SIAM J. CONTROL OPTIM.
Vol. 38, No. 1, pp. 1–21

THE GENERIC NUMBER OF INVARIANT ZEROS OF A STRUCTURED LINEAR SYSTEM*

JACOB VAN DER WOUDE†

Abstract. In this paper we consider a structured linear system represented by means of a directed graph. We present a graph theoretic method to compute the generic number of invariant zeros of the corresponding Rosenbrock matrix. The method is based on a fundamental decomposition of the directed graph representing the structured system. The generic number of invariant zeros is important in generic versions of problems involving controllability and pole placement.

Key words. graph theory, linking, separating subset, linear system, invariant zeros

AMS subject classifications. 05C50, 15A22, 93A99, 93C05, 93C41, 94C15

PII. S0363012996310119

1. Introduction.

This paper deals with a method to compute the generic number of invariant zeros of a structured system. This number of zeros has already been studied in various papers (see, for instance, [10], [11]) and also in the book [7]. However, in none of the cited references was the most general case treated or was a complete solution given. In this paper we aim for the most general case and we present a graph theoretic method to compute the generic number of invariant zeros of a general structured linear system. Fundamental in our approach is a decomposition of the directed graph representing the structured system into a number of structured subsystems.

The material in this paper on separators and (separator-based) decompositions is largely based on results in [5] on M-decompositions. In [5] M-decompositions are developed in order to hierarchically decompose a (large) set of (algebraic) equations into a number of smaller sets of equations such that the solvability of the original set of equations can be investigated by checking the solvability of the smaller sets of equations individually and by using their hierarchical ordering.

In this paper we use only a simplified version of the M-decomposition and we apply it on linear structured systems described by differential equations. We show that after applying the simplified M-like decomposition, we obtain in principle three essentially different linear structured subsystems that each have their own specific properties. These properties can be used to compute the generic number of invariant zeros of each of the subsystems individually. By combining the obtained results for the subsystems we get the generic number of invariant zeros of the original system. In [16] the subsystems with their specific properties are used to investigate the generic solvability of the disturbance decoupling problem with pole placement for structured systems.

For the sake of completeness and the readers' interest, we have included some illustrative results on separators and separator-based decompositions. Also we present a simple and straightforward algorithm for the decomposition used in this paper.

*Received by the editors October 7, 1996; accepted for publication (in revised form) November 4, 1998; published electronically November 10, 1999.
http://www.siam.org/journals/sicon/38-1/31011.html
†Faculty Information Technology and Systems, Delft University of Technology, Mekelweg 4, 2628 CD Delft, The Netherlands (j.w.vanderwoude@math.tudelft.nl).

We want to also point out that we can apply a general version of the M-decomposition on linear structured systems described by differential equations. In principle, this will yield a minimal inconsistent part, a number of consistent parts, and a maximal inconsistent part (see [5]). Two of the three subsystems mentioned above can be obtained directly from the minimal part and the maximal inconsistent part, respectively. The third subsystem is obtained by aggregating all the components of the consistent parts.

The main goal of our paper is the presentation of the three different subsystems with their specific properties and their use for computing the generic number of invariant zeros of a structured linear system. Since the subsystems can already be obtained by means of our simplified version of the M-decomposition, we have refrained from using a more general version of this decomposition. For more details on a general version of the M-decomposition, we refer to [5].

The outline of this paper is as follows. In section 2 we give the mathematical formulation of the problem of the paper. In section 3 we explain how a structured system can be represented by means of a directed graph. In section 4 we present some simplifying assumptions that in the context of this paper can be made without harming the generality and we present the decomposition of a graph on which this paper is based. In section 5 we present results that are crucial for this paper. These results are combined to yield our main result: a graph theoretic method for computing the generic number of invariant zeros of a general structured system. In section 6 we illustrate the method by means of an elementary example. In section 7 we present some conclusions and remarks. We give (sketches of) proofs of some of the results of this paper in the appendix.

2. Problem formulation.

2.1. Systems and number of zeros. In this paper we study the following linear time-invariant system:

$$\Sigma: \quad \left\{ \begin{array}{rcl} \dot{x}(t) &=& Ax(t) + Bu(t), \\ y(t) &=& Cx(t) + Du(t), \end{array} \right. \tag{1}$$

with state $x(t) \in \mathbb{R}^n$, input $u(t) \in \mathbb{R}^m$, and output $y(t) \in \mathbb{R}^p$. The matrices A, B, C, and D are real and have dimensions $n \times n$, $n \times m$, $p \times n$, and $p \times m$, respectively. With I_n the $n \times n$ identity matrix, we write the *system pencil* of system (1) as (see [8])

$$P(s) = \begin{pmatrix} A - sI_n & B \\ C & D \end{pmatrix}.$$

From the next subsection on we assume that system (1) is *structured* in the sense that only the zero/nonzero structure of the matrices A, B, C, and D is known. However, here we still assume that system (1) is numerically specified. Then, regarding $P(s)$ as a rational matrix, we call its rank the *normal-rank* and we denote this normal-rank by n-rank $P(s)$. For a particular complex value $\hat{s}$, we denote the *rank* of the constant (complex) matrix $P(\hat{s})$ by rank $P(\hat{s})$. It is well known that if n-rank $P(s) = q$, then rank $P(\hat{s}) = q$ for almost all values $\hat{s} \in \mathbb{C}$ and $\max_{\hat{s} \in \mathbb{C}}$ rank $P(\hat{s}) = q$ (see [13]). Given that n-rank $P(s) = q$, there may exist numbers $\hat{s} \in \mathbb{C}$ such that rank $P(\hat{s}) < q$. We call such numbers the *invariant zeros* of the system (see [1]). Formally, if n-rank $P(s) = q$, we have the following definition.

DEFINITION 2.1. *The number $s_0 \in \mathbb{C}$ is an invariant zero of system* (1) *if rank* $P(s_0) < q$.

For certain applications, for instance, applications involving pole placement, the *number* of invariant zeros of system (1) is important (see [8]), where the number of zeros is counted with multiplicity. See also [16] for an application of the results of this paper.

In this paper we study methods to compute *the number of invariant zeros*. Our goal is to compute the number of invariant zeros for *structured systems*. For numerically specified (nonstructured) systems, we have various methods available to compute the number of invariant zeros. In the following we review some cases that will illustrate the approach followed in this paper:

Special Cases.

1. We begin by considering the case when $p = m$. Then $P(s)$ is square with dimensions $(n + p) \times (n + p)$. If in addition n-rank $P(s) = n + p$, then $P(s)$ is invertible as a rational matrix. The number of invariant zeros of system (1) then equals the degree of the determinant of $P(s)$.

2. Next we consider the case when $p < m$. If in addition n-rank $P(s) = n + p$, then $P(s)$ has a right inverse, seen as a rational matrix. Now if s_0 is an invariant zero of (1), then rank $P(s_0) < n + p$. Hence, any $(n + p) \times (n + p)$ submatrix of $P(s_0)$ has zero determinant. Therefore, s_0 is a zero of any $(n + p)$th order minor of $P(s)$, implying that s_0 is a zero of the greatest common divisor of all these minors. Conversely, if s_0 is a zero of the greatest common divisor of all $(n+p)$th order minors of $P(s)$, then any $(n+p)\times(n+p)$ submatrix of $P(s)$ has zero determinant for $s = s_0$, implying that rank $P(s_0) < n + p$ and that s_0 is an invariant zero of (1). Therefore, if $p < m$ and n-rank $P(s) = n + p$, then the number of invariant zeros of system (1) is equal to the degree of the greatest common divisor of all $(n + p)$th order minors of $P(s)$.

3. Similarly as above, we can treat the case when $m < p$ with n-rank $P(s) = n + m$.

4. We now consider the case that by row and column permutations the system pencil $P(s)$ is transformed into $\tilde{P}(s)$ given by

$$\tilde{P}(s) = \begin{pmatrix} \tilde{P}_1(s) & \tilde{Q}_{12}(s) \\ 0 & \tilde{P}_2(s) \end{pmatrix},$$

with

- $\tilde{P}_i(s)$ *system pencils like* $P(s)$ *of dimensions* $a_i \times b_i$ *for* $i = 1, 2$,
- *n-rank* $\tilde{P}_1(s) = a_1$,
- *n-rank* $\tilde{P}_2(s) = b_2$.

Hence, seen as rational matrices, we have that $\tilde{P}_1(s)$ has full row rank and $\tilde{P}_2(s)$ has full column rank. Then using some basic calculus for rational matrices we can show that n-rank $P(s)$ = n-rank $\tilde{P}(s) = q$ with $q = a_1 + b_2$. The $a_1 \times b_2$ polynomial matrix $\tilde{Q}_{12}(s)$ is not relevant in this context. By simple reasoning it is easy to prove that any invariant zero of $\tilde{P}(s)$ is an invariant zero of $\tilde{P}_1(s)$ and/or $\tilde{P}_2(s)$, and also the converse is true. Any invariant zero of $\tilde{P}_1(s)$ and/or $\tilde{P}_2(s)$ is also an invariant zero of $\tilde{P}(s)$. Moreover, counting the zeros with multiplicities it can be shown that the number of invariant zeros of $\tilde{P}(s)$ is equal to the sum of the number of invariant zeros of $\tilde{P}_1(s)$ and the number of invariant zeros of $\tilde{P}_2(s)$. Therefore, to compute the number of invariant zeros of $\tilde{P}(s)$, we can compute the number of invariant zeros of the smaller matrices $\tilde{P}_1(s)$ and $\tilde{P}_2(s)$ and add them together. Since row and column

permutations are of no influence, we then also have obtained the number of invariant zeros of $P(s)$.

2.2. Structured systems and generic number of zeros. In the previous subsection we considered systems that are numerically specified. However, in this paper we mainly treat systems (1) that are not numerically specified. Indeed, in the rest of this paper we assume that we know only the *zero/nonzero* structure of the matrices A, B, C, and D. This means that of each entry in these matrices we know only whether its value is fixed to zero, in which case we call it a *fixed zero*, or that it has an unknown real value, in which case we call the entry a *free parameter*. We say that a system (1) is *structured* if only the zero/nonzero structure of the matrices A, B, C, and D is given.

In a structured system (1) with l nonzero entries in the matrices A, B, C, and D, we can parametrize these entries by a scalar real parameters λ_i, $i = 1, 2, \ldots, l$, together forming a parameter vector $\lambda = (\lambda_1, \lambda_2, \ldots, \lambda_l)^\top \in \mathbb{R}^l$, where $^\top$ means transpose.

We denote the matrices obtained by replacing the nonzeros in A, B, C, and D by the corresponding parameters λ_i, $i = 1, 2, \ldots, l$, by A_λ, B_λ, C_λ, and D_λ, respectively. Following this notation we also write

$$P_\lambda(s) = \begin{pmatrix} A_\lambda - sI_n & B_\lambda \\ C_\lambda & D_\lambda \end{pmatrix}.$$

For each value of $\lambda \in \mathbb{R}^l$, we have a numerically specified system with system pencil $P_\lambda(s)$ of which we can compute the n-rank and the number of invariant zeros. It turns out that the n-rank of $P_\lambda(s)$ will have the same value for almost all parameter values $\lambda \in \mathbb{R}^l$ (see [7], [14]). This is the so-called *generic n-rank* of $P(s)$ and will be denoted by g-n-rank $P(s)$.

If the g-n-rank of $P(s)$ is equal to q, it can be shown that the degree of each qth order minor of $P(s)$ individually will have the same value for almost all parameter values (see [7]). The same applies to the degree of the greatest common divisor of the qth order minors of $P(s)$. This means that for almost all parameter values the number of invariant zeros of $P(s)$ is the same. Therefore, we can refer to this number as the *generic number of invariant zeros* of $P(s)$.

Above, each of the expressions "for almost all" is to be understood as "for all except for those in some proper algebraic variety in $\mathbb{R}^l$" (see [12]).

2.3. Approach followed in this paper. In this paper we develop a method to determine the generic number of invariant zeros of a structured system (1). Basically, the method consists of two steps.

In the first step the system pencil of a structured system

$$P(s) = \begin{pmatrix} A - sI_n & B \\ C & D \end{pmatrix}$$

is transformed using row and column permutations into the next form

$$P(s) = \begin{pmatrix} P_1(s) & Q_{12}(s) & Q_{13}(s) \\ 0 & P_2(s) & Q_{23}(s) \\ 0 & 0 & P_3(s) \end{pmatrix},$$

in which $P_i(s)$, $i = 1, 2, 3$, are system pencils like $P(s)$, such that

- $P_1(s)$ *generically has full row rank (even after deleting an arbitrary column; see in the later sections),*
- $P_2(s)$ *is generically invertible, and*
- $P_3(s)$ *generically has full column rank (even after deleting an arbitrary row; see in the later sections).*

In the second step, as in subsection 2.1, we determine the generic number of invariant zeros of the structured system (1) by adding the generic number of invariant zeros of the structured subsystems characterized by the system pencils $P_1(s)$, $P_2(s)$, and $P_3(s)$.

The generic number of invariant zeros of the structured subsystem characterized by $P_2(s)$ is simply the generic degree of the determinant of $P_2(s)$. The generic number of invariant zeros of the structured subsystem characterized by $P_1(s)$ equals the generic degree of the greatest common divisor of all maximum order minors of $P_1(s)$. Similarly, the generic number of invariant zeros of the structured subsystem is characterized by $P_3(s)$.

Using graph theory we are able to determine row and column permutations such that $P(s)$ is transformed into the above triangular form. In addition to obtaining the above triangular form, the permutations are such that $P_1(s)$ has special properties that enable us to determine the generic degree of the greatest common divisor of all maximum order minors of $P_1(s)$ in a straightforward way. This also holds for the generic degree of the greatest common divisor of all maximum order minors of $P_3(s)$.

3. Graph representation of structured systems.

3.1. Graphs, paths, linkings, and cycles families. We introduce the graph $G = (V, E)$ of a structured system of the form (1). The vertex set V of the graph is given by $U \cup X \cup Y$ with $U = \{u_1, \dots, u_m\}$ the set of input vertices, $X = \{x_1, \dots, x_n\}$ the set of state vertices, $Y = \{y_1, \dots, y_p\}$ the set of output vertices. Hence, V consists of $n + m + p$ vertices. The indices n, m, and p come from system (1) and denote the dimension of the state space, the input space, and the output space, respectively. Denoting (v, v') for a directed edge from the vertex $v \in V$ to the vertex $v' \in V$, the edge set E of the graph is described by $E_A \cup E_B \cup E_C \cup E_D$ with $E_A = \{(x_j, x_i) | A_{ij} \neq 0\}$, $E_B = \{(u_j, x_i) | B_{ij} \neq 0\}$, $E_C = \{(x_j, y_i) | C_{ij} \neq 0\}$, $E_D = \{(u_j, y_i) | D_{ij} \neq 0\}$. In the latter, for instance, $A_{ij} \neq 0$ means that the (i, j)th entry of the matrix A is a free parameter (a nonzero). An example of a graph is given in section 6.

Let W, W' be two nonempty subsets of the vertex set V of the graph G. We say that there exists a *path* from W to W', if there is an integer t and there are vertices $w_0, w_1, \dots, w_t \in V$ such that $w_0 \in W$, $w_t \in W'$, and $(w_{i-1}, w_i) \in E$ for $i = 1, 2, \dots t$. We call the vertex w_0 the *begin vertex* of the path and w_t the *end vertex*. We say that the path consists of the vertices $w_0, w_1, \dots, w_t$, where it may happen that some of the vertices occur more than once. We also say that each of the vertices in $w_0, w_1, \dots, w_t$ is contained in the path. We call the path *simple* if every vertex on the path occurs only once. Occasionally, we denote a path by the sequence of directed edges it consists of, i.e., by $(w_0, w_1), (w_1, w_2), \dots, (w_{t-1}, w_t)$. If $(w_0, w_1), \dots, (w_{i-1}, w_i), (w_i, w_{i+1}), \dots, (w_{j-1}, w_j), (w_j, w_{j+1}), \dots, (w_{t-1}, w_t)$ is a path from the vertex w_0 to the vertex w_t, first along the vertex w_i and then along the vertex w_j, the part from the vertex w_i to the vertex w_j is called a *subpath* of the original path.

We say that two paths from W to W' are *disjoint* if they consist of disjoint sets of vertices. We call l paths from W to W' disjoint if they are mutually disjoint, i.e., each of the two of them are disjoint. We call a set of l *disjoint* and *simple* paths from

W to W' a *linking from W to W' of size l*. Since there are only a finite number of linkings, there obviously exist linkings consisting of a maximum number of disjoint paths. We call such linkings *maximum (size) linkings*.

We call a *simple* path a *U-rooted path* if the path has its begin vertex in U. We call a number of mutually disjoint U-rooted paths a *U-rooted path family*.

We call a *closed* and *simple* path a *cycle*, i.e., a cycle is a path of the form $(w_0, w_1), (w_1, w_2), \ldots, (w_{t-1}, w_0)$, consisting of distinct vertices $w_0, w_1, \ldots, w_{t-1}$. We call two cycles *disjoint* if they consist of disjoint sets of vertices. We say that l cycles are disjoint if they are mutually disjoint. We call such a set of l disjoint cycles a *cycle family of size l*.

We say that the union of a linking, a U-rooted path family and a cycle family is *disjoint* if they mutually have no vertices in common.

3.2. Separators, essential vertices, and subgraphs. Given the graph $G = (V, E)$ we call a set of vertices $S \subseteq V$ a *separator* between the sets U and Y if every path from U to Y contains at least one vertex in S. The number of vertices in a separator is called the *size* of the separator. Since there are only a finite number of separators between U and Y, there clearly exist such separators having the smallest size. We call such separators *minimum (size) separators* between U and Y. We denote the family of all minimum separators between U and Y by $\mathcal{S}$.

The following theorem is now well known (see [5]).

THEOREM 3.1 (Menger). *The size of a maximum linking from U to Y is equal to the size of a minimum separator between U and Y.*

Next we assume that the maximum size of a linking from U to Y is equal to k and we concentrate on linkings of size k from U to Y. We define the set of so-called *essential vertices* as follows.

DEFINITION 3.2. *$V_{ess} = \{v \in V |$ v is included in every linking of size k from U to Y $\}$.*

We have the following property which is proved in the report version of this paper [15].

PROPOSITION 3.3. *$V_{ess} = \cup\{S | S \in \mathcal{S}\}$.*

Thus, V_{ess} consists of all vertices that are present in minimum size separators between U and Y. The following observations are immediate.

COROLLARY 3.4.

- *$S \subseteq V_{ess}$ for all $S \in \mathcal{S}$,*
- *V_{ess} is a separator between U and Y.*

We now introduce subgraphs of the original graph G. However, before doing this we observe that input vertices (vertices in U) can be seen as vertices from which only edges are coming out and to which no edges are going into. Similarly, output vertices (vertices in Y) can be seen as vertices from which no edges are coming out and to which only edges are going into.

Introducing a class of subgraphs we occasionally want to change the role of vertices. For instance, we may want to think of a state vertex $x \in X$ as an input vertex. To do so, we simply ignore the edges that are going into x and concentrate on the edges that are coming out of x. Similarly, if we want to think of x as an output vertex, we simply ignore the edges that are coming out of x and concentrate on the edges that are going into x. In the same manner we can think of an output vertex $y \in Y$ as an input vertex by ignoring the edges that are going into y, yielding a vertex y that is isolated (to which no edges are going into or from which no edges are coming out). Similarly, an input vertex $u \in U$ can be thought of as an output vertex by ignoring

the edges that are coming out of u, yielding a vertex u that is isolated. We discuss these isolated vertices in some more detail later.

Now we let W and $\tilde{W}$ be two nonempty subsets of the vertex set V and we concentrate on all so-called *IO-paths* in G from W to $\tilde{W}$. Here, an IO-path from W to $\tilde{W}$ is a path from W to $\tilde{W}$ where we think of the vertices in W as input vertices (only outgoing edges) and of the vertices in $\tilde{W}$ as output vertices (only incoming edges). Hence, we have the following definition.

DEFINITION 3.5. *An IO-path in G from W to $\tilde{W}$ is a path from W to $\tilde{W}$ that only has one vertex in W (the begin vertex) and only one vertex in $\tilde{W}$ (the end vertex).*

Given the sets W and $\tilde{W}$ we collect the vertices on all IO-paths from W to $\tilde{W}$ in the set V' and we collect all the edges in E between vertices in V' in the set E'. The combination of V' and E' then defines the subgraph G', i.e., $G' = (V', E')$. To indicate that the subgraph G' is obtained by concentrating on all IO-paths in G from W to $\tilde{W}$ we write $G' = H(G; W, \tilde{W})$. We note that we do not require that the sets W and $\tilde{W}$ have an empty intersection. If $v \in W \cap \tilde{W}$, then the vertex v is at the same time seen as an input vertex and an output vertex. As we explained above, all the edges that are coming out of v and all the edges that are going into v, respectively, are then ignored in G'. Hence, the vertex v in G' is an isolated vertex.

Given two pairs of subsets $(W_1, \tilde{W}_1)$ and $(W_2, \tilde{W}_2)$ in V, we can compare (partially order) the associated subgraphs $G_i = H(G; W_i, \tilde{W}_i)$, $i = 1, 2$, in the following way.

DEFINITION 3.6. *Given two pairs of subsets $(W_1, \tilde{W}_1)$ and $(W_2, \tilde{W}_2)$ in V, defining two subgraphs $G_i = H(G; W_i, \tilde{W}_i)$, $i = 1, 2$, of the graph G. Then $G_1 \preceq G_2$ if any IO-path from W_2 to $\tilde{W}_2$ contains a subpath that is an IO-path from W_1 to $\tilde{W}_1$.*

Put differently, $G_1 \preceq G_2$ if and only if any IO-path from W_1 to $\tilde{W}_1$ can be extended (in both directions if necessary) to an IO-path from W_2 to $\tilde{W}_2$. With this ordering of subgraphs we have the following important result. For a proof we refer to [15].

THEOREM 3.7. *There exist two uniquely determined minimum separators between U and Y, denoted as S^* and S_*, such that*

- *$H(G; U, S_*) \preceq H(G; U, S)$ for all $S \in \mathcal{S}$,*
- *$H(G; S^*, Y) \preceq H(G; S, Y)$ for all $S \in \mathcal{S}$.*

To give some intuition for the minimum separators S^* and S_*, we consider a simple path P from $u \in U$ to $y \in Y$ and a minimum separator $S \in \mathcal{S}$. Clearly, the path must contain a vertex in S, say, s. Next we think of the first part of the path from u to s as an IO-path in $H(G; U, S)$. According to the above theorem there is a begin part of this IO-path that can be seen as an IO-path in $H(G; U, S_*)$, say, from u to $s_* \in S_*$. As S_* is the "smallest" with respect to the ordering $\preceq$ introduced above, the previous means that when following any path from U to Y the first vertex in a minimum separator must be a vertex in S_*. Similarly, the last vertex on any path P from U to Y in a minimum separator must be a vertex in S^*.

As $S_* \in \mathcal{S}$ and $V_{ess} = \cup\{S | S \in \mathcal{S}\}$ it follows that $S_* \subseteq V_{ess}$. It also follows that any path from U to V_{ess} has to end in some $S \in \mathcal{S}$ and has to pass through S_*. Thus, any IO-path from U to V_{ess} is in fact an IO-path from U to S_*. The latter implies the following corollary.

COROLLARY 3.8.

- $H(G; U, V_{ess}) = H(G; U, S_*)$,
- $H(G; V_{ess}, Y) = H(G; S^*, Y)$.

Hence, we can compute S_* by considering the set of output vertices of the subgraph $H(G; U, V_{ess})$, where vertices of V_{ess} in U are also considered as output vertices

for the moment. Similarly, we can compute S^* by considering the set of input vertices of the subgraph $H(G; V_{ess}, Y)$. The subset V_{ess} can be computed beforehand by applying algorithms based on the computation of the maximum flow through an associated network.

4. Simplifying assumptions and graph decomposition.

4.1. Simplifying assumptions and basic result. We consider a structured system of the type (1) with the graph $G = (V, E)$ with $V = U \cup X \cup Y$. In the context of this paper we can make the following assumptions on the graph G without harming the generality. We recall that U, X, and Y denote the set of input, state, and output vertices, respectively.

ASSUMPTION 4.1.

- *Every vertex in U is the begin vertex of an edge that ends in a vertex in $X \cup Y$,*
- *Every vertex in Y is the end vertex of an edge that begins in a vertex in $U \cup X$,*
- *Every vertex in X is contained in a path (not necessarily simple) from U to Y.*

We can motivate the assumptions as follows. Following the definition of the graph we see that there is no edge from input vertex u_j to a state or output vertex if and only if the jth column of the compound matrix $\binom{B}{D}$ consists of fixed zeros only. In case of such a zero column the jth input has no effect at all on the system and therefore can be ignored without loss of generality.

Similarly, we see that there is no edge that ends in the output vertex y_i and that starts in a state or input vertex if and only if the ith row of the compound matrix $\begin{pmatrix} C & D \end{pmatrix}$ consists of only zeros. In case of such a zero row the ith output has no information at all on the system and therefore can be ignored without loss of generality. This explains the two first assumptions.

Now we come to the third assumption. The next proposition is well known (see [2], [4], [9]).

PROPOSITION 4.2. *Every state vertex $x \in X$ is the end vertex of some U-rooted path if and only if the pair (A, B) is not reducible, where we say that the pair (A, B) is* reducible *if there is a permutation matrix Q such that*

$$QAQ^{-1} = \begin{pmatrix} A_{11} & 0 \\ A_{21} & A_{22} \end{pmatrix}, \qquad QB = \begin{pmatrix} 0 \\ B_2 \end{pmatrix},$$

with A_{11}, A_{22}, and B_2 are matrices of dimensions $n_1 \times n_1$, $n_2 \times n_2$, and $n_2 \times m$, respectively, with $n_1 + n_2 = n$ and $n_1 > 0$.

Proposition 4.2 states that if there is a state vertex $x \in X$ that is not the end vertex of a U-rooted path, there is a permutation matrix Q such that the system pencil $P(s)$ transforms into

$$\begin{pmatrix} A_{11} - sI_{n_1} & 0 & 0 \\ A_{21} & A_{22} - sI_{n_2} & B_2 \\ C_1 & C_2 & D \end{pmatrix},$$

where $CQ^{-1} = (C_1 \;\; C_2)$. From the latter we immediately see that every eigenvalue of A_{11} is an invariant zero of the original $P(s)$. This means that the generic number of invariant zeros of $P(s)$ is equal to n_1 plus the generic number of invariant zeros of

$$\begin{pmatrix} A_{22} - sI_{n_2} & B_2 \\ C_2 & D \end{pmatrix}.$$

Therefore, as far as the generic number of invariant zeros of $P(s)$ is concerned we may assume without loss of generality that we have split off the "reducible part" and that we may concentrate on the "nonreducible remainder." Inspired by Proposition 4.2 and the above decomposition this means that we just may concentrate on those structured subsystems of which all state vertices are end vertices of U-rooted paths.

Similarly, if there are state vertices that are not begin vertices of paths ending in Y, we can reorder the state vertices in the graph and split off the vertices that are not begin vertices of paths ending in Y. Similarly as before with n_1, only the *number* of those vertices is relevant for the generic number of invariant zeros. Once this number is known the associated vertices can be ignored. We therefore may concentrate on those structured subsystems of which all state vertices are begin vertices of paths ending in Y.

Thus, we can say that without loss of generality we may focus on graphs of which the state vertices are end vertices of paths beginning in U as well as begin vertices of paths ending in Y. Put differently, we may focus on those graphs of which the state vertices are contained in a path from U to Y, which is our third assumption.

To conclude this subsection we recall a well known result (see [13] for a similar result for systems without D matrix). The results relates the g-n-rank of the system pencil with the existence of a maximum size linking from U to Y. See also [15] for a proof.

THEOREM 4.3. *Let $P(s)$ be the system pencil of the structured system* (1)*. Then the n-rank of $P(s)$ is generically equal to n* plus *the* maximum *size of a linking from U to Y.*

4.2. Separator-based decomposition. In this subsection we present a decomposition of the graph $G = (V, E)$ based on a minimum separator between U and Y. We assume that the maximum size of a linking from U to Y is equal to k. Further, we let S be a minimum separator between U and Y. According to the theorem of Menger we know that the set S has to consist of k vertices. Finally, we assume Assumption 4.1 to be valid.

First we consider the subgraph $H(G; U, S)$ of the graph G. As we want to think of the vertices in S as new output vertices, the vertices in $U \cap S$ have to play a double role, i.e., at the same time they have be both input and output vertices. These vertices then have an unclear system theoretic interpretation and have to be treated specially. We do this by first considering only the IO-paths from U to S that have positive length. In this way we first ignore the vertices in $U \cap S$.

We collect all the vertices on the positive length IO-paths from U to S in the set V_α. To distinguish the various vertices in V_α we adopt the following general notation: if K, L are sets then $K\backslash L$ denotes the set of elements in K that are not in L, i.e., $K\backslash L = \{v \in K | v \notin L\}$. With this notation we now define the next new input, output, and state vertex sets

$$U_\alpha = V_\alpha \cap U, \;\; Y_\alpha = V_\alpha \cap S, \;\; X_\alpha = V_\alpha \backslash (U_\alpha \cup Y_\alpha).$$

Due to Assumption 4.1 every vertex in U is the begin vertex of a path from U to Y. Hence, we have that every vertex in U is the begin vertex of an IO-path from U to S. Moreover, since we exclude IO-paths of zero length, it follows that $U_\alpha = U\backslash S$. Furthermore, since S is a separator of minimum size, every vertex of S is the end vertex of at least one IO-path from U to S. Concentrating on positive length IO-paths only we obtain that $Y_\alpha = S\backslash U = S \cap (X \cup Y)$. Therefore

$$U_\alpha = U\backslash S, \;\; Y_\alpha = S\backslash U = S \cap (X \cup Y), \;\; X_\alpha = V_\alpha \backslash (U_\alpha \cup Y_\alpha).$$

By construction $X_\alpha \cap U_\alpha = \emptyset$ and $X_\alpha \cap Y_\alpha = \emptyset$. From the above it follows directly that also $U_\alpha \cap Y_\alpha = \emptyset$. Hence, the collection $\{U_\alpha, Y_\alpha, X_\alpha\}$ forms a partitioning of V_α, meaning that the union of the subsets in the collection is V_α and that the intersection of any two subsets in the collection is empty.

In the above we thought of the vertices in S as new output vertices. However, we can also think of these vertices as new input vertices. Then the vertices in $S \cap Y$ have to be treated in a special way. In view of the above we collect all vertices on positive length IO-paths from S to Y into the set V_β. Then we can define new output, input, and state vertex sets as follows:

$$Y_\beta = V_\beta \cap Y, \;\; U_\beta = V_\beta \cap S, \;\; X_\beta = V_\beta \backslash (U_\beta \cup Y_\beta).$$

Due to Assumption 4.1 every vertex in Y is the end vertex of a path from U to Y. Similarly as above it follows that

$$Y_\beta = Y \backslash S, \;\; U_\beta = S \backslash Y = S \cap (X \cup U), \;\; X_\beta = V_\beta \backslash (U_\beta \cup Y_\beta).$$

The collection $\{U_\beta, Y_\beta, X_\beta\}$ forms a partitioning of V_β.

We now consider both V_α and V_β simultaneously. Hence, we have the graph $G = (V, E)$ with $V = U \cup X \cup Y$ and we consider the subgraphs $G_\alpha = (V_\alpha, E_\alpha)$ and $G_\beta = (V_\beta, E_\beta)$, where V_α and V_β are defined as above and the edge sets E_α and E_β are obtained by restricting the edge set E to V_α and V_β, respectively.

It easily follows that $U \subseteq U_\alpha \cup U_\beta$ and $Y \subseteq Y_\alpha \cup Y_\beta$. We next observe that the intersection of V_α and V_β must be contained in S, i.e., $V_\alpha \cap V_\beta \subseteq S$. Indeed, suppose that $v \in V_\alpha \cap V_\beta$, but that $v \notin S$. Because $v \in V_\alpha$, the vertex v is contained in an IO-path from U to S. Since $v \notin S$ the part from U to v of such an IO-path does not contain any vertex in S. Similarly, since $v \in V_\beta$, there is a part of an IO-path from S to Y forming a path from v to Y that does not contain any vertex in S. Combining these two subpaths results in a path from U to Y along v that does not contain any vertex in S. However, this is in contradiction with the fact that S is a separator between U and Y and, consequently, must contain at least one vertex of any path from U to Y. Thus, we have proved that

$$V_\alpha \cap V_\beta \subseteq S.$$

Intersecting both sides with $V_\alpha \cap V_\beta$ we obtain that $V_\alpha \cap V_\beta = S \cap (V_\alpha \cap V_\beta) = (S \cap V_\alpha) \cap (S \cap V_\beta) = Y_\alpha \cap U_\beta$. This implies that $V_\alpha \cap V_\beta = (S \backslash U) \cap (S \backslash Y) = S \backslash (U \cup Y) = S \cap X$.

To come to a complete partitioning of the vertex set V we define

$$X_\delta = V \backslash (V_\alpha \cup V_\beta).$$

The notation suggests that $X_\delta \subseteq X$. As $U \subseteq U_\alpha \cup U_\beta$ and $Y \subseteq Y_\alpha \cup Y_\beta$ it is obvious that this is indeed true. From above it now immediately follows that the next collection of subsets forms a partitioning of V

$$\{X_\alpha, X_\beta, X_\delta, U_\alpha, Y_\beta, S \cap X, S \cap U, S \cap Y\}.$$

Indeed, by construction $\{V_\alpha \cup V_\beta, X_\delta\}$ is a partitioning of V. Observe further that $Y_\alpha = S \backslash U = (S \cap X) \cup (S \cap Y)$ and $U_\beta = S \backslash Y = (S \cap X) \cup (S \cap U)$, where clearly $\{S \cap X, S \cap Y\}$ is a partitioning of Y_α and $\{S \cap X, S \cap U\}$ is a partitioning of U_β. This implies that $\{X_\alpha, U_\alpha, S \cap X, S \cap Y\}$ is a partitioning of V_α, $\{X_\beta, Y_\beta, S \cap X, S \cap U\}$ is

a partitioning of V_β, and because $V_\alpha \cap V_\beta = S \cap X$ that $\{X_\alpha, X_\beta, U_\alpha, Y_\beta, S \cap X, S \cap U, S \cap Y\}$ is a partitioning of $V_\alpha \cup V_\beta$.

Moreover, it is easy to see that the collection $\{X_\alpha, X_\beta, X_\delta, U_\alpha, U_\beta\}$ is a partitioning of $X \cup U$ and $\{X_\alpha, X_\beta, X_\delta, Y_\alpha, Y_\beta\}$ is a partitioning of $X \cup Y$.

4.3. Properties of separator-based decomposition. Given the above partitioning of V into the subsets $\{X_\alpha,\ X_\beta,\ X_\delta,\ U_\alpha,\ Y_\beta,\ S \cap X,\ S \cap U,\ S \cap Y\}$ we can prove the following result of which a proof can be found in the appendix.

PROPOSITION 4.4. *Let S be a minimum separator between U and Y and let $V = U \cup X \cup Y$ be partitioned as above. Then there are no edges*

- *from $X_\alpha \cup U_\alpha$ to $X_\beta \cup Y_\beta$,*
- *from $X_\alpha \cup U_\alpha$ to X_δ,*
- *from X_δ to $X_\beta \cup Y_\beta$.*

The partitioning of the vertex sets $X \cup U$ and $X \cup Y$ induces a partitioning of the system matrix $P(s)$. To see this, we first consider the matrix $P(s)$ for $s = 0$,

$$\begin{pmatrix} A & B \\ C & D \end{pmatrix},$$

and think of it to act on the vector $\left(\begin{smallmatrix} x \\ u \end{smallmatrix}\right)$ to yield the vector $\left(\begin{smallmatrix} \dot{x} \\ y \end{smallmatrix}\right)$. Next, we reorder the components of these vectors in accordance to the partitionings $\{X_\alpha, U_\alpha, X_\delta, X_\beta, U_\beta\}$ and $\{X_\alpha, Y_\alpha, X_\delta, X_\beta, Y_\beta\}$, respectively, to obtain

$$\begin{pmatrix} x_\alpha \\ u_\alpha \\ x_\delta \\ x_\beta \\ u_\beta \end{pmatrix} \text{ and } \begin{pmatrix} \dot{x}_\alpha \\ y_\alpha \\ \dot{x}_\delta \\ \dot{x}_\beta \\ y_\beta \end{pmatrix}, \text{respectively.}$$

For instance, the vector x_α contains all components of the vector x in X_α. Similarly for the other subvectors. Since there is no edge from $X_\alpha \cup U_\alpha$ to $X_\delta \cup X_\beta \cup Y_\beta$ and no edge from $X_\alpha \cup U_\alpha \cup X_\delta$ to $X_\beta \cup Y_\beta$, there are row and column permutations that transform the matrix $\left(\begin{smallmatrix} A & B \\ C & D \end{smallmatrix}\right)$ into the upper block triangular matrix

$$\begin{pmatrix} A_\alpha & B_\alpha & * & * & * \\ C_\alpha & D_\alpha & * & * & * \\ 0 & 0 & A_\delta & * & * \\ 0 & 0 & 0 & A_\beta & B_\beta \\ 0 & 0 & 0 & C_\beta & D_\beta \end{pmatrix},$$

where all matrices have suitable dimensions and with the $*$'s denoting matrices that are not relevant in the present context.

For the matrix $P(s)$, the above triangular matrix means that using a suitable row and column permutation $P(s)$ can be transformed into

$$\begin{pmatrix} A_\alpha - sI_{n_\alpha} & B_\alpha & * & * & * \\ C_\alpha & D_\alpha & * & * & * \\ 0 & 0 & A_\delta - sI_{n_\delta} & * & * \\ 0 & 0 & 0 & A_\beta - sI_{n_\beta} & B_\beta \\ 0 & 0 & 0 & C_\beta & D_\beta \end{pmatrix},$$

where all matrices are as before and n_i denotes the number of vertices in X_i, $i = \alpha, \delta, \beta$.

Given the above block triangular form of the matrix $P(s)$ we can prove the following result. For a proof, we refer to the appendix.

PROPOSITION 4.5. *Let the system pencil $P(s)$ be transformed into the upper block triangular form as given above. Then the following hold:*

- *The subsystem pencil*

$$P_\alpha(s) = \begin{pmatrix} A_\alpha - sI_{n_\alpha} & B_\alpha \\ C_\alpha & D_\alpha \end{pmatrix}$$

  generically has full row rank.
- *The subsystem pencil*

$$\begin{pmatrix} A_\delta - sI_{n_\delta} & * & * \\ 0 & A_\beta - sI_{n_\beta} & B_\beta \\ 0 & C_\beta & D_\beta \end{pmatrix}$$

  generically has full column rank.
- *If in the above $S = S_*$, the subsystem pencil $P_\alpha(s)$ generically has full row rank, even after deleting an arbitrary column.*

5. Main results. In this section we present the main results of this paper.

5.1. Fundamental results for special cases. In this subsection we first consider a square structured system of type (1) that is generically invertible. The following theorem indicates how the number of invariant zeros can be computed.

THEOREM 5.1. *Assume that $m = p$, and let $P(s)$ be generically invertible. Then the degree of the determinant of $P(s)$ is generically equal to $n+p$* minus *the* minimum *number of edges in a maximum size (size p) linking from U to Y.*

A proof of Theorem 5.1 is given in the appendix where it is related to one of the results in [14].

Next we assume that $m > p$ and that $P(s)$ generically has full row rank, even after the deletion of an arbitrary column. Then we have the following important result.

THEOREM 5.2. *Let $P(s)$ generically have full row n-rank, even after the deletion of an arbitrary column. Then the greatest common divisor of all the $(n+p)$th order minors of $P(s)$ generically is a monomial in s with a degree equal to $n+p$* minus *the* maximum *number of edges in the disjoint union of a*

- *linking of size p from U to Y,*
- *a U-rooted path family,*
- *a cycle family in X.*

A sketch of a proof of Theorem 5.2 will be given in the appendix. A full proof can be found in [15]. Theorem 5.2 is an extension of one of the main results in [3]. Theorems 5.1 and 5.2 will be illustrated in the next section.

5.2. Partitioning of a structured system. In this subsection we consider a structured system of the type (1) and we assume that Assumption 4.1 is satisfied. The next result states how in general a structured system can be seen as the combination of three smaller subsystems each having properties that are useful for computing the generic number of invariant zeros.

THEOREM 5.3. *There exist row and column permutations such that the system pencil $P(s)$ of a structured system* (1) *with matrices A, B, C, and D can be trans-*

formed into the block triangular form,

$$\begin{pmatrix} A_1 - sI_{n_1} & B_1 & * & * & * & * \\ C_1 & D_1 & * & * & * & * \\ 0 & 0 & A_2 - sI_{n_2} & B_2 & * & * \\ 0 & 0 & C_2 & D_2 & * & * \\ 0 & 0 & 0 & 0 & A_3 - sI_{n_3} & B_3 \\ 0 & 0 & 0 & 0 & C_3 & D_3 \end{pmatrix},$$

with A_i, B_i, C_i, and D_i structured matrices of dimensions $n_i \times n_i$, $n_i \times m_i$, $p_i \times n_i$, and $p_i \times m_i$, respectively, $i = 1, 2, 3$, such that, if present

- $\left(\begin{smallmatrix} A_1 - sI_{n_1} & B_1 \\ C_1 & D_1 \end{smallmatrix}\right)$ *generically has full row n-rank, even after the deletion of an arbitrary column;*
- $\left(\begin{smallmatrix} A_2 - sI_{n_2} & B_2 \\ C_2 & D_2 \end{smallmatrix}\right)$ *generically is invertible (as a rational matrix);*
- $\left(\begin{smallmatrix} A_3 - sI_{n_3} & B_3 \\ C_3 & D_3 \end{smallmatrix}\right)$ *generically has full column n-rank, even after the deletion of an arbitrary row.*

In the above block triangular form the $*$'s denote matrices (matrix pencils) that are not relevant in the context of this paper. A sketch of a proof of Theorem 5.3 is given in the appendix. It is based on Propositions 4.4 and 4.5, combined with the minimum separators S_* and S^*.

5.3. Algorithm. In this subsection we present an algorithm to compute the generic number of invariant zeros of a structured system. The algorithm consists of a first part in which the graph of the system is split into smaller subgraphs for which the generic number of invariant zeros can be computed by Theorems 5.1 and 5.2. This is done in the second part of the algorithm. The results are then taken together and yield the generic number of invariant zeros of the original system.

ALGORITHM 5.4.

- *Make a directed graph of the structured system as explained in subsection* 3.1.
- *Check if the assumptions in subsection* 4.1 *are valid. If not, do preliminary computations as explained in subsection* 4.1 *that result in a smaller structured system for which the assumptions in subsection* 4.1 *are valid.*
- *Using, for instance, Menger's theorem determine the set of essential vertices V_{ess} and the sets S_* and S^* using the subgraphs $H(G; U, V_{ess})$ and $H(G; V_{ess}, Y)$, respectively.*
- *Construct the subgraphs G_1 and G_3 by considering all IO-paths with positive length from U to S_* and from S^* to Y, respectively. Let the subgraph G_2 be determined by considering the vertices in S_* as input vertices, the vertices in S^* as output vertices, and all the remaining vertices of V that are not present in G_1 or G_3.*
- *Using Theorems* 5.1, 5.2, *and the dual of Theorem* 5.2 *compute the generic number of invariant zeros of the subsystems corresponding to the obtained partitioning.*
- *Add the obtained generic numbers to get the generic number of invariant zeros of the original structured system.*

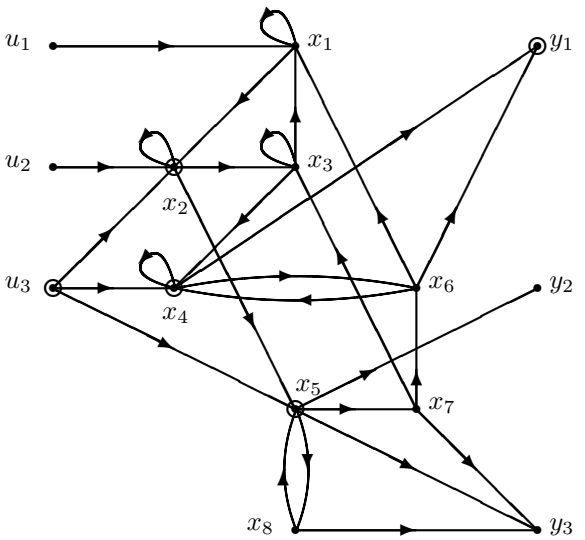


FIG. 6.1. *The graph G with the vertices in V_{ess} already encircled.*

6. Example. We consider the structured system of type (1) as depicted in Figure 6.1. We can parametrize the matrices of the structured system by a parameter $\lambda \in \mathbb{R}^{27}$ as follows.

$$A_\lambda = \begin{pmatrix} \lambda_1 & 0 & \lambda_3 & 0 & 0 & \lambda_4 & 0 & 0 \\ \lambda_5 & \lambda_7 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & \lambda_{10} & \lambda_9 & 0 & 0 & 0 & \lambda_{11} & 0 \\ 0 & 0 & \lambda_{12} & \lambda_{15} & 0 & \lambda_{14} & 0 & 0 \\ 0 & \lambda_{16} & 0 & 0 & 0 & 0 & 0 & \lambda_{18} \\ 0 & 0 & 0 & \lambda_{19} & 0 & 0 & \lambda_{20} & 0 \\ 0 & 0 & 0 & 0 & \lambda_{23} & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & \lambda_{24} & 0 & 0 & 0 \end{pmatrix}, B_\lambda = \begin{pmatrix} \lambda_2 & 0 & 0 \\ 0 & \lambda_6 & \lambda_8 \\ 0 & 0 & 0 \\ 0 & 0 & \lambda_{13} \\ 0 & 0 & \lambda_{17} \\ 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix},$$

$$C_\lambda = \begin{pmatrix} 0 & 0 & 0 & \lambda_{22} & 0 & \lambda_{21} & 0 & 0 \\ 0 & 0 & 0 & 0 & \lambda_{25} & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & \lambda_{28} & 0 & \lambda_{26} & \lambda_{27} \end{pmatrix}, D_\lambda = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}.$$

By Figure 6.1 it can be easily verified that Assumption 4.1 is valid. Hence, we are in the position to directly apply our main result and we do not need to perform some preliminary computations. From Figure 6.1 we can conclude that the maximum size of a linking from U to Y is 2. To see this, we note that the following two paths from U to Y are disjoint: (u_2, x_2),(x_2, x_5),(x_5, y_2) and (u_3, x_4),(x_4, y_1). Furthermore, we note that paths from u_1 to Y and paths from u_2 to Y always contain x_2, and, therefore, never can be disjoint. As there are only three input vertices it follows that the maximum size of a linking from U to Y has to be 2.

Further, from the above two paths we know already that the vertices u_1, x_1, x_3, x_6, x_7, x_8, and y_3 are not contained in V_{ess}. In fact, it is easy to see that V_{ess} is given by $\{u_3, x_2, x_4, x_5, y_1\}$. In Figure 6.1 we have already encircled the vertices in V_{ess}.

Next we concentrate on the subgraphs $H(G; U, V_{ess}) = H(G; U, S_*)$ and $H(G; V_{ess}\ , Y) = H(G; S^*, Y)$; see Corollary 3.8. From Figure 6.1 it is easy to see

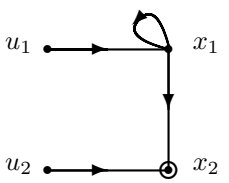


FIG. 6.2a G_1.

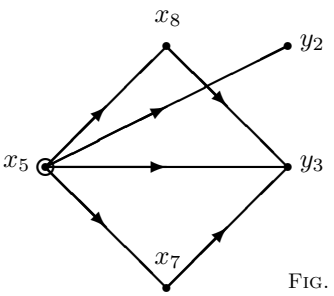


FIG. 6.2b G_3.

that $S_* = \{u_3, x_2\}$ and $S^* = \{x_5, y_1\}$, respectively. In Figures 6.2a and 6.2b we have depicted the subgraphs G_1 and G_3, respectively. We have obtained the subgraph G_1 by concentrating on all IO-paths from U to V_{ess} (or S_*) that have a positive length or, alternatively, by just deleting from $H(G; U, V_{ess})$ all isolated vertices. The subgraph G_3 is obtained in a similar way from $H(G; V_{ess}, Y)$. Finally, we collect all vertices on IO-paths in G from S_* to S^* in the subgraph G_2. We have depicted G_2 in Figure 6.3.

As a consequence of the above we can conclude that with a suitable row and column permutation the system matrix of system (1) can be transformed into the next form

$$\left(\begin{array}{ccc|ccccc|ccc}
-s+\lambda_1 & \lambda_2 & 0 & \lambda_3 & 0 & \lambda_4 & 0 & 0 & 0 & 0 & 0 \\
\lambda_5 & 0 & \lambda_6 & 0 & 0 & 0 & -s+\lambda_7 & \lambda_8 & 0 & 0 & 0 \\
\hline
0 & 0 & 0 & -s+\lambda_9 & 0 & 0 & \lambda_{10} & 0 & \lambda_{11} & 0 & 0 \\
0 & 0 & 0 & \lambda_{12} & -s+\lambda_{15} & \lambda_{14} & 0 & \lambda_{13} & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & \lambda_{19} & -s & 0 & 0 & \lambda_{20} & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & \lambda_{16} & \lambda_{17} & 0 & \lambda_{18} & -s \\
0 & 0 & 0 & 0 & \lambda_{22} & \lambda_{21} & 0 & 0 & 0 & 0 & 0 \\
\hline
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & -s & 0 & \lambda_{23} \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & -s & \lambda_{24} \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \lambda_{25} \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \lambda_{26} & \lambda_{27} & \lambda_{28}
\end{array}\right).$$

In this block triangular form we can distinguish the three subsystems corresponding to the three subgraphs G_1, G_2, and G_3, respectively. These subgraphs G_1, G_2, and G_3 correspond to the following system pencils:

$$P_1(s) = \begin{pmatrix} A_1 - sI_{n_1} & B_1 \\ C_1 & D_1 \end{pmatrix} = \begin{pmatrix} -s+\lambda_1 & \lambda_2 & 0 \\ \lambda_5 & 0 & \lambda_6 \end{pmatrix}$$

with $n_1 = 1, m_1 = 2, p_1 = 1$,

$$P_2(s) = \begin{pmatrix} A_2 - sI_{n_2} & B_2 \\ C_2 & D_2 \end{pmatrix} = \begin{pmatrix} -s+\lambda_9 & 0 & 0 & \lambda_{10} & 0 \\ \lambda_{12} & -s+\lambda_{15} & \lambda_{14} & 0 & \lambda_{13} \\ 0 & \lambda_{19} & -s & 0 & 0 \\ 0 & 0 & 0 & \lambda_{16} & \lambda_{17} \\ 0 & \lambda_{22} & \lambda_{21} & 0 & 0 \end{pmatrix}$$

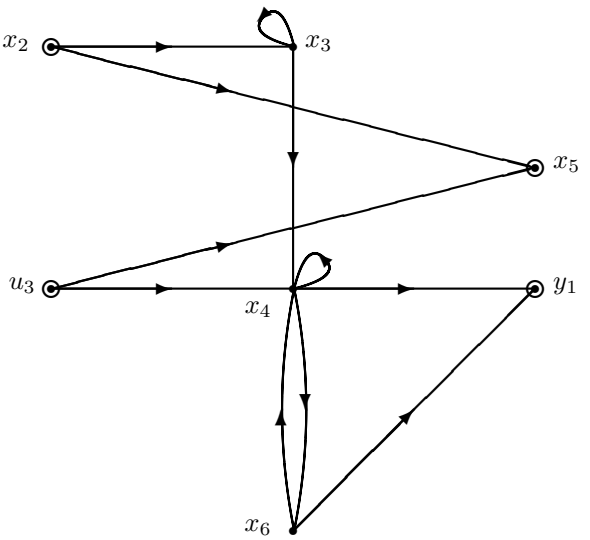


FIG. 6.3. G_2

with $n_2 = 3, m_2 = 2, p_2 = 2$, and

$$P_3(s) = \begin{pmatrix} A_3 - sI_{n_3} & B_3 \\ C_3 & D_3 \end{pmatrix} = \begin{pmatrix} -s & 0 & \lambda_{23} \\ 0 & -s & \lambda_{24} \\ 0 & 0 & \lambda_{25} \\ \lambda_{26} & \lambda_{27} & \lambda_{28} \end{pmatrix}$$

with $n_3 = 2, m_3 = 1, p_3 = 2$, respectively.

Applying Theorem 5.1 on graph G_2 it follows that the generic number of invariant zeros is equal to 2, namely, $n_2 + p_2$ minus the minimum number of edges in a size p_2 linking from U_2 $(= \{x_2, u_3\})$ to Y_2 $(= \{x_5, y_1\})$ in G_2. Since $n_2 = 3$, $p_2 = 2$ and any size 2 linking from U_2 to Y_2 in G_2 contains at least 3 edges, the generic number of invariant zeros of the subsystem corresponding to the graph G_2 equals $3 + 2 - 3 = 2$.

For the graph G_1, the application of Theorem 5.2 yields that the generic number of invariant zeros is equal to 0, being $n_1 + p_1$ minus the maximum number of edges in the disjoint union of a linking of size p_1 from U_1 $(= \{u_1, u_2\})$ to Y_1 $(= \{x_2\})$, a U_1-rooted path family, and a cycle family in X_1 $(= \{x_1\})$. Here $n_1 = 1$, $p_1 = 1$, and it is easy to see that the number of edges in a disjoint union just described is at most 2. For instance, the disjoint union of the size 1 linking from U_1 to Y_1 given by the edge (u_2, x_2) and the U_1-rooted path family given by the edge (u_1, x_1) contains 2 edges. The same applies to all other such unions. Using Theorem 5.2 it, therefore, follows that the generic number of invariant zeros of the subsystem corresponding to the graph G_1 equals $1 + 1 - 2 = 0$.

For the graph G_3, the application of Theorem 5.2 in transposed form yields that the generic number of invariant zeros is equal to $3 - 2 = 1$. This is left to the reader.

Combining the above we can now conclude that the generic number of invariant zeros of the original structured system (1) is $0 + 2 + 1 = 3$.

7. Conclusions. In this paper we studied structured systems and we presented a graph theoretical method to compute the generic number of invariant zeros of such

systems. The method uses elementary and efficient network algorithms. We think that the results in the paper are of an elementary and basic nature, and we believe that they are important in the many areas of systems theory that involve pole placement issues when only the zero/nonzero structure is known.

Appendix. In this appendix we give the proofs of some of the results of this paper.

Proof of Proposition 4.4. To prove the first part of the proposition, we take $v \in X_\alpha \cup U_\alpha$ and $w \in X_\beta \cup Y_\beta$ and we assume that there is an edge from v to w, i.e., $(v, w) \in E$. From the definition of X_α and U_α it is clear that $v \notin S$. Similarly, it follows from the definitions of X_β and Y_β that $w \notin S$. Since $v \in V_\alpha$ and $v \notin S$, there is an IO-path from U to v that does not contain any vertex in S. Similarly, because $w \in V_\beta$ and $w \notin S$ there is an IO-path from w to Y that does not contain any vertex in S. Connecting these two paths by the edge between v and w, i.e., by (v, w), we obtain a path from U to Y that does not contain any vertex in S. However, this is in contradiction with the fact that S is a separator between U and Y. Thus, we must conclude that there does not exists an edge from v to w, implying that there can not be an edge from $X_\alpha \cup U_\alpha$ to $X_\beta \cup Y_\beta$.

To prove the second part of the proposition, we again take a vertex $v \in X_\alpha \cup U_\alpha$. But we now consider a vertex $w \in X_\delta$ and we assume that there is an edge from v to w, i.e., $(v, w) \in E$. From the proof of the first part of the proposition we know that there exists a path, say, P_1, from U to v that does not contain any vertex in S. Since we assume that every state vertex is contained in a path from U to Y, there is a simple path from U to Y that contains the vertex $w \in X_\delta$. We denote this path by P and we denote the part of P from w to Y by P_2. Now there are two possibilities.

- P_2 does not contain any vertex in S,
- P_2 does contain a vertex in S.

In the first case the combination of the path P_1, the edge (v, w), and the path P_2 yields a path from U to Y that does not contain any vertex in S. This is in contradiction with the fact that S is a separator between U and Y. In the second case denote the first vertex on P_2 starting from w that is contained in S by w' and denote the part of P_2 from w to w' by P_2'. Recall that $w \in X_\delta$. Now consider the combination of the path P_1, the edge (v, w) and the path P_2'. This combination establishes an IO-path from U to S. Indeed, only the end vertex w' is in S. Since w is contained in the combined path made up of the path P_1, the edge (v, w) and the path P_2' it follows by construction that $w \in V_\alpha$. This is not possible since V_α and X_δ have empty intersection.

Since both cases are not possible, we must conclude that there does not exist an edge from v to w, implying that there can not be an edge from $X_\alpha \cup U_\alpha$ to X_δ.

The third part of the proposition can be proved analogously to the second part.

This concludes the proof of Proposition 4.4. □

Proof of Proposition 4.5. We recall that the number of vertices in a minimum separator S is k. We denote k_U, k_X, and k_Y for the number of vertices in the sets $S \cap U$, $S \cap X$ and $S \cap Y$, respectively, so that $k_U + k_X + k_Y = k$. The number of vertices in V_α is given by the sum of the number of vertices in X_α, $U_\alpha = U \backslash S$, and $Y_\alpha = S \backslash U$. The number of vertices in V_α is therefore given by the expression $n_\alpha + (m - k_U) + (k - k_U)$.

We observe that a linking of size k from U to Y in G induces a linking of size $(k - k_U)$ from U_α to Y_α. Indeed, restrict the paths of the linking to the subpaths that begin in U and end in S, and of these subpaths take the $(k - k_U)$ subpaths that have

a positive length. According to Theorem 4.3 it follows now that the generic rank of the system pencil

$$\begin{pmatrix} A_\alpha - sI_{n_\alpha} & B_\alpha \\ C_\alpha & D_\alpha \end{pmatrix}$$

is equal to $n_\alpha + (k - k_U)$, implying that the pencil generically has full row rank. This completes the proof of the first statement of Proposition 4.5.

In the same way we can prove that the pencil

$$\begin{pmatrix} A_\beta - sI_{n_\beta} & B_\beta \\ C_\beta & D_\beta \end{pmatrix}$$

generically has full column rank equal to $n_\beta + (k - k_Y)$. As $(A_\delta - sI_{n_\delta})$ (generically) has full rank the second statement of Proposition 4.5 easily follows.

If $S = S_*$, we have that $H(G; U, S) = H(G; U, V_{ess})$; see Corollary 3.8. Concentrating on $H(G; U, V_{ess})$ we have that in this subgraph there is a linking of size k from U to V_{ess}. We recall that $G_\alpha = (V_\alpha, E_\alpha)$ can be obtained from $H(G; U, V_{ess})$ by just ignoring the isolated vertices in $U \cap V_{ess}$. As by definition vertices in X_α and U_α are therefore not contained in V_{ess}, they are not essential and each of them can be deleted without decreasing the maximum size of a linking from U to V_{ess}. Also in G_α the deletion of a single vertex from U_α or X_α does not decrease the maximum size of a linking from U_α to Y_α. We note that instead of the deletion of vertices and incident edges we can also delete the outgoing edges from these vertices alone.

Deletion of vertex u_i in U_α corresponds to the deletion of the column $n_\alpha + i$ from $P_\alpha(s)$. Deletion of the outgoing edges of vertex x_i corresponds to the deletion of the column i from $P_\alpha(s)$. As in both cases a linking of size $(k - k_U)$ from U_α to Y_α still is possible, it follows that the generic rank of $P_\alpha(s)$ is $n_\alpha + (k - k_U)$ even after the deletion of an arbitrary column from $P_\alpha(s)$. This completes the proof of the third statement of Proposition 4.5. □

Proof of Theorem 5.1. In the case of the matrix $D = 0$, i.e, D consists of fixed zeros only, Theorem 5.1 is formulated and proved in [14] as Theorem 6.2. For general D a proof of Theorem 5.1 can be given by a straightforward modification of the proof in [14]. This is left for the reader. To conclude this proof we note that in Theorem 5.1 the number of edges is counted, while in Theorem 6.2 in [14] state vertices are counted. This explains the difference in formulation of Theorem 5.1 here and Theorem 6.2 in [14]. This concludes the proof of Theorem 5.1. □

Sketch of proof of Theorem 5.2. A full proof can be found in [15]. Here we have divided the proof into a number of steps.

1. Assuming that g-n-rank $P(s) = n + p$, we first indicate that the rank of $P(s)$ for $s = 0$ generically is equal $n + p$ if and only if there is a disjoint union of
   - a linking of size p from U to Y,
   - a U-rooted path family,
   - a cycle family in X

that contains $n + p$ (the largest possible number) edges.

2. Next we prove that under the conditions of Theorem 5.2 the greatest common divisor of all the $(n + p)$th order minors of $P(s)$ generically is a monomial in s (a single power in s). See also [5], Lemma 14.5, or Proposition A.1 in [17].

3. Then we may follow the line of thought in [3] and we define

$$\mu = \min \left\{ s | \exists i_1, i_2, \ldots, i_s : \begin{pmatrix} A & B \\ C & D \end{pmatrix}^{(i_1, i_2, \ldots, i_s)}_{(i_1, i_2, \ldots, i_s)} \text{ generically has full row rank} \right\},$$

where we assume that $1 \leq i_1 < i_2 < \cdots < i_s \leq n$, and where we have written

$$\begin{pmatrix} A & B \\ C & D \end{pmatrix}^{(i_1,i_2,\ldots,i_s)}_{(i_1,i_2,\ldots,i_s)}$$

for the matrix obtained from $\left(\begin{smallmatrix} A & B \\ C & D \end{smallmatrix}\right)$ by deleting all rows and all columns with index in $\{i_1, i_2, \ldots, i_s\}$. We note that $\left(\begin{smallmatrix} A & B \\ C & D \end{smallmatrix}\right)^{(i_1,i_2,\ldots,i_s)}_{(i_1,i_2,\ldots,i_s)}$ itself defines a system with an $n-s$ dimensional state. The input vector remains m dimensional and the output vector p dimensional.

In the spirit of [3] we can prove that every $(n+p)$th order minor of $P(s)$ can be divided by s^μ. Furthermore, we can indicate that at least one $(n+p)$th order minor can be found by specifying the values of the nonzeros in $P(s)$ that contains a term of the form as^μ with $a \neq 0$. The existence of such a minor can be proved in the same way as a similar result is proved in [3].

Finally, combining Steps 2 and 3 we obtain that

- the greatest common divisor of all the $(n+p)$th order minors of $P(s)$ is generically a monomial in s;
- every $(n+p)$th order minor of $P(s)$ can be divided by s^μ, while there is an $(n+p)$th order minor of $P(s)$ that can not be divided by $s^{\mu+1}$.

Hence, the greatest common divisor of all the $(n+p)$th order minors of $P(s)$ is generically of the form as^μ with $a \neq 0$. Combining the definition of μ with Step 1, we obtain that μ is the minimal number of vertices that have to be deleted from the graph G such that in the remaining graph there exists a disjoint union of a linking of size p from U to Y, a U-rooted path family, and a cycle family in X that contains $(n-\mu)+p$ edges.

Instead of trying to take μ minimal we can try to take $(n-\mu)+p$ maximal. Suppose that the maximal number of edges in a disjoint union as above is ρ. Then $\rho = (n-\mu)+p$ or $\mu = (n+p)-\rho$, and we can conclude that the greatest common divisor of all the $(n+p)$th order minors of $P(s)$ is generically a monomial with a degree equal to $(n+p)$ minus the maximal number of edges in a disjoint union of

- a linking of size p from U to Y,
- a U-rooted path family,
- a cycle family in X,

which completes the proof of Theorem 5.2. □

Proof of Theorem 5.3. Theorem 5.3 is in fact a consequence of the combination of Propositions 4.4 and 4.5 and the sets S_* and S^*. Below a sketch of the proof of Theorem 5.3 is given.

First the decomposition of subsection 4.2 is considered for $S = S_*$. This decomposition is based on $H(G; U, V_{ess})$ and yields a subgraph $G_1 = (V_1, E_1)$ of G, where V_1 denotes the set of all vertices contained in IO-paths from U to S_* that have a positive length. Note that we use the index 1 instead of the index α as in subsection 4.2. The vertex set can be partitioned as $V_1 = U_1 \cup X_1 \cup Y_1$, where

$$U_1 = U \backslash S_*, \;\; Y_1 = S_* \backslash U, \;\; X_1 = V_1 \backslash (U_1 \cup Y_1).$$

In the spirit of sections 4.2 and 4.3 define $\bar{V}$ as the set of all vertices contained in IO-paths from S_* to Y that have a positive length. (In terms of subsection 4.2 this means that $\bar{V} = V_\beta$.) Furthermore, we define

$$\tilde{U} = S_* \backslash Y, \; \tilde{Y} = Y \backslash S_*, \; \tilde{X} = (\bar{V} \backslash (\tilde{U} \cup \tilde{Y})) \cup (V \backslash (V_1 \cup \bar{V})), \; \tilde{V} = \tilde{U} \cup \tilde{X} \cup \tilde{Y}.$$

Then according to the results of subsections 4.2 and 4.3 there are row and column permutations such that the system pencil $P(s)$ can be transformed into

$$\begin{pmatrix} P_1(s) & * \\ 0 & \tilde{P}(s) \end{pmatrix},$$

with $P_1(s)$ a subsystem pencil of the form

$$\begin{pmatrix} A_1 - sI_{n_1} & B_1 \\ C_1 & D_1 \end{pmatrix},$$

generically having full row rank, even after deleting an arbitrary column, and with $\tilde{P}(s)$ a subsystem pencil of the form

$$\begin{pmatrix} \tilde{A} - sI_{\tilde{n}} & \tilde{B} \\ \tilde{C} & \tilde{D} \end{pmatrix},$$

generically having full column rank.

In terms of the decomposition of subsection 4.2 with $S = S_*$, we have that $P_1(s)$ in fact equals

$$\begin{pmatrix} A_\alpha - sI_{n_\alpha} & B_\alpha \\ C_\alpha & D_\alpha \end{pmatrix}$$

and that $\tilde{P}(s)$ equals

$$\begin{pmatrix} A_\delta - sI_{n_\delta} & * & * \\ 0 & A_\beta - sI_{n_\beta} & B_\beta \\ 0 & C_\beta & D_\beta \end{pmatrix}.$$

Next we use the set S^* to decompose the subsystem pencil $\tilde{P}(s)$. For instance, it can be shown that $S^* \subseteq \tilde{V} \cup S_*$ and that $S^* \backslash (Y \cap S_*)$ is the "largest" minimum separator between $\tilde{U}$ and $\tilde{Y}$ in $\tilde{G}$ (see [15]). This implies that we can apply the decomposition of section 4.2 in a dual way on $\tilde{P}(s)$. We then obtain row and column permutations such that $\tilde{P}(s)$ is transformed into

$$\begin{pmatrix} A_2 - sI_{n_2} & B_2 & * & * \\ C_2 & D_2 & * & * \\ 0 & 0 & A_3 - sI_{n_3} & B_3 \\ 0 & 0 & C_3 & D_3 \end{pmatrix},$$

with $P_2(s)$ a subsystem pencil of the form

$$\begin{pmatrix} A_2 - sI_{n_2} & B_2 \\ C_2 & D_2 \end{pmatrix},$$

generically having full row rank and with $P_3(s)$ a subsystem pencil of the form

$$P_3(s) = \begin{pmatrix} A_3 - sI_{n_3} & B_3 \\ C_3 & D_3 \end{pmatrix},$$

generically having full column rank, even after deleting an arbitrary row. As $\tilde{P}(s)$ generically has full column rank, it immediately follows that $P_2(s)$ is square and generically invertible.

Combining the previous two decompositions we have shown that there are row and column permutations such that the system pencil $P(s)$ transforms into the block triangular form described in Theorem 5.3. □

REFERENCES

[1] H. Aling and J. M. Schumacher, *A nine-fold canonical decomposition for linear systems*, Internat. J. Control, 39 (1984), pp. 779–805.

[2] K. Glover and L. M. Silverman, *Characterization of structural controllability*, IEEE Trans. Automat. Control, AC-21 (1976), pp. 534–537.

[3] S. Hosoe, *Determination of generic dimensions of controllable subspaces and its application*, IEEE Trans. Automat. Control, AC-25 (1980), pp. 1192–1196.

[4] C. T. Lin, *Structural controllability*, IEEE Trans. Automat. Control, AC-19 (1974), pp. 201–208.

[5] K. Murota, *Systems Analysis by Graphs and Matroids—Structural Solvability and Controllability*, Algorithms Combin. 3, Springer-Verlag, New York, 1987.

[6] Y. Ohta and S. Kodama, *Structural invertibility of transfer functions*, IEEE Trans. Automat. Control, AC-30 (1985), pp. 818–819.

[7] K. J. Reinschke, *Multivariable Control, A Graph-Theoretic Approach*, Springer-Verlag, New York, 1988.

[8] H. H. Rosenbrock, *State-Space and Multivariable Theory*, Wiley, New York, 1970.

[9] R. W. Shields and J. B. Pearson, *Structural controllability of multiinput linear systems*, IEEE Trans. Automat. Control, AC-21 (1976), pp. 203–212.

[10] W. Söte, *Eine graphische Methode zur Ermittlung der Nullstellen in Mehrgrössensystemen*, Reglungstechnik, 28 (1980), pp. 346–348.

[11] F. Svaricek, *Graphentheoretische Ermittlung der Anzahl von strukturellen und streng strukturellen invarianten Nullstellen*, Automatisierungstechnik, 34 (1986), pp. 488–497.

[12] W. M. Wonham, *Linear Multivariable Control: A Geometric Approach*, 3rd ed., Springer-Verlag, New York, 1985.

[13] J. W. van der Woude, *A graph-theoretic characterization for the rank of the transfer matrix of a structured system*, Math. Control Signals Systems, 4 (1991), pp. 33–40.

[14] J. W. van der Woude, *On the structure at infinity of a structured system*, Linear Algebra Appl., 148 (1991), pp. 145–169.

[15] J. W. van der Woude, *The generic number of invariant zeros of a structured linear system*, Report 96-96, Faculty of Technical Mathematics and Informatics, available at ftp.twi.tudelft.nl in the directory /pub/publications/tech-reports/1996 as DUT-TWI-96-96.ps.gz

[16] J. W. van der Woude, *Graph theoretic conditions for structural disturbance decoupling with pole placement*, TU-E E2, in CD-ROM 4th European Control Conference ECC '97, Brussels.

[17] J. W. van der Woude and K. Murota, *Disturbance decoupling with pole placement for structured systems: A graph theoretic approach*, SIAM J. Matrix Anal. Appl., 162 (1995), pp. 922–942.

SIAM J. CONTROL OPTIM.
Vol. 38, No. 1, pp. 22–49

DESIGN OF HOMOGENEOUS TIME-VARYING STABILIZING CONTROL LAWS FOR DRIFTLESS CONTROLLABLE SYSTEMS VIA OSCILLATORY APPROXIMATION OF LIE BRACKETS IN CLOSED LOOP*

PASCAL MORIN†, JEAN-BAPTISTE POMET†, AND CLAUDE SAMSON†

Abstract. A constructive method for time-varying stabilization of smooth driftless controllable systems is developed. It provides time-varying homogeneous feedback laws that are continuous and smooth away from the origin. These feedbacks make the closed-loop system globally exponentially asymptotically stable if the control system is homogeneous with respect to a family of dilations and, using local homogeneous approximation of control systems, locally exponentially asymptotically stable otherwise.

The method uses some known algorithms that construct oscillatory control inputs to approximate motion in the direction of iterated Lie brackets that we adapt to the closed-loop context.

Key words. nonlinear control, stabilization, time-varying stabilization, controllability, Lie brackets

AMS subject classifications. 93D15, 34C29, 93B52

PII. S0363012997315427

1. Introduction.

1.1. Related work and contribution. Stabilization by continuous time-varying feedback laws of nonlinear systems that cannot be stabilized by time-invariant continuous feedback laws has been an ongoing subject of research in the past few years.

The fact that for many controllable systems no continuous stabilizing feedback exists was first pointed out by Sussmann [23]. A simple necessary condition was given by Brockett [1], since known as "Brockett's condition." It allows us to identify a wide class of controllable systems for which no continuous stabilizing feedback exists; these include most controllable driftless systems. More recently, Coron gave a stronger necessary condition [2].

A possible way of stabilizing systems for which these necessary conditions are violated is to use discontinuous (time-invariant) control laws. This has been explored in the literature, but the present work does not go in this direction at all.

The possibility of stabilizing nonlinear controllable systems via *continuous time-varying feedback control laws* was first noticed in the very detailed study of stabilization of one-dimensional systems by Sontag and Sussmann [21]. More recently, smooth stabilizing control laws for some nonholonomic mechanical systems were given by Samson [18]; this was the starting point of a systematic study of time-varying stabilization. Coron [3] proved that all controllable driftless systems may be stabilized by continuous (and even smooth) time-varying feedback and that "most" controllable systems (even with drift) can also be stabilized by continuous time-varying feedback [4]. Pomet deals with a less general class of controllable driftless systems [15].

*Received by the editors January 22, 1997; accepted for publication (in revised form) August 10, 1998; published electronically November 10, 1999.
http://www.siam.org/journals/sicon/38-1/31542.html
†INRIA Sophia-Antipolis, B. P. 93, 06902 Sophia Antipolis cedex, France (pmorin@sophia.inria.fr, pomet@sophia.inria.fr, samson@sophia.inria.fr).

From here on, only driftless systems are considered in this paper. After the general existence result given in [3], studies on the subject have focused on methods to *construct* continuous time-varying stabilizing feedback laws and on obtaining feedback laws that provide sufficiently fast convergence.

As far as the constructiveness aspect is concerned, for simplicity let us divide the construction methods into two kinds. The first kind of method applies to rather large classes of controllable driftless systems, such as the work of Coron [3] (general controllable driftless systems; the paper is not oriented toward construction of the control, but a method can be extracted from the proofs), Pomet [15] (controllable driftless systems for which the control Lie algebra is generated by a specific set of vector fields), or of M'Closkey and Murray [12] (same conditions as in [15]). These studies all share the following feature: they use the solution of a linear PDE, or the expression of the flow of a vector field, to construct the control law. This solution, or this flow, has to be calculated beforehand, either analytically or numerically, and this introduces, especially when no analytical solution is available, a degree of complication which may not be necessary. The second kind of method found in the literature provides explicit expressions. Its drawback is that it applies only to specific subclasses of driftless systems, such as models of mobile robots or systems in the "chain form" or "power-form," like the work of Samson [18], Teel, Murray, and Walsh [27], and Sépulchre, Campion, and Vertz [20], among others.

Alternatively, a need to improve the speed of convergence came out of the slow convergence associated with the *smooth* control laws that were first proposed. This concern motivated several studies, starting with the work by M'Closkey and Murray [11], yielding continuous control laws which are not smooth, or even Lipschitz everywhere, but are homogeneous with respect to some dilation, and thus exponentially stabilizing, not in the standard sense but with respect to some homogeneous norm (this notion was introduced by Kawski [7]). See, for instance, further work by the authors of this paper [16, 14] or by M'Closkey and Murray [12], who have also proposed recently a procedure that transforms a given smooth stabilizing control law into a homogeneous one [13]. Except for this last reference, which requires that a smooth stabilizing control law has been designed beforehand, the construction of homogeneous exponentially stabilizing control laws in the literature is restricted to specific subclasses of driftless systems.

The design method described in the present paper has the advantage of being totally explicit, in the sense that it requires only ordinary differentiation and linear algebraic operations, while it applies to general controllable systems and provides exponential stability. This method gives homogeneous feedbacks, which ensure global stability if the control vector fields are homogeneous and local stability otherwise. The fact that it relates controllability with the construction of a stabilizing control law in a more direct way than previous designs also makes it conceptually appealing, all the more so as it may be viewed as converting the open-loop control techniques reported by Liu and Sussmann in [25] and Liu [9] into closed-loop techniques.

However, the generality of the method also has a price. When applied to particular systems for which explicit solutions have long been available, the present method often yields solutions which are significantly more complicated. This comes partly from the complexity of the approximation algorithm proposed in [25, 9], which we use. This is also a consequence of the modifications that we have made to adapt this algorithm to our feedback control objective.

1.2. Outline of the method. Nonlinear controllability results were first derived for driftless systems; see, for instance, the work by Lobry [10], where it is shown that such systems are controllable if and only if any direction in the state space can be obtained as a linear combination of iterated Lie brackets of the control vector fields, at least in real-analytic cases. It was also shown very early on by Haynes and Hermes [5] that, under this same condition, *any* curve in the state-space can be approached by open-loop solutions of the controlled system. (Note that this property is not shared by all controllable systems, but rather is specific to driftless systems.) In these studies the key element is that, in addition to the directions of motion corresponding to the control vector fields, motion along other directions corresponding to iterated Lie brackets is also possible by quickly switching motions along the original control vector fields. Take, for example, a system with two controls

$$\dot{x} \;=\; u_1\, b_1(x) \;+\; u_2\, b_2(x) \tag{1.1}$$

with state x in $\mathbb{R}^5$, and assume that at each point x the vectors

$$b_1(x)\,,\; b_2(x)\,,\; [b_1,b_2](x)\,,\; [b_1,[b_1,b_2]](x)\,,\; [b_2,[b_1,b_2]](x) \tag{1.2}$$

are linearly independent, and thus span $\mathbb{R}^5$. The idea in [5] is the following: first, it is clear that *any* (e.g., differentiable) parameterized curve $t \mapsto \gamma(t)$ is a possible solution of the "extended" system with five controls:

$$\begin{aligned}\dot{x} = v_1\, b_1(x) \;+\; v_2\, b_2(x) \;+\; v_3\, [b_1,b_2](x)\\ +\; v_4\, [b_1,[b_1,b_2]](x) \;+\; v_5\, [b_2,[b_1,b_2]](x)\end{aligned} \tag{1.3}$$

(simply decompose $\dot{\gamma}(t)$ on the basis (1.2) to obtain the controls). Then it is proved in [5] that there exists a sequence of (oscillatory) controls $u_1(\varepsilon, t, v_1, v_2, v_3, v_4, v_5)$ and $u_2(\varepsilon, t, v_1, v_2, v_3,\; v_4, v_5)$ such that the system (1.1) "converges to" the system (1.3) when $\varepsilon \to 0$ in the sense that the solutions of (1.1) with these controls u_k converge uniformly on finite time intervals to the solutions of (1.3). The proof in [5] does not give a process to build these sequences of approximating sequence of oscillatory control, and although the case of a simple bracket (approximating $[b_1, b_2]$ by switching between b_1 and b_2) is elementary and well known, the case above of brackets of order 3 is more complex. The more recent work by Liu [9] and Sussmann and Liu [25] gives an explicit construction of the approximating sequence. The process of building this sequence is amazingly intricate compared to the simplicity of the existence proof in [5]. Of course, the controls u_k are not defined for $\varepsilon = 0$, and both their frequency and their amplitude tend to infinity when ε goes to zero.

Being aware of these results, and faced with the problem of proving that any controllable driftless system may be stabilized by means of a periodic feedback, the most natural idea is probably the following, which we illustrate for (1.1) (5 states, 2 controls):

(a) Stabilize the extended system (1.3) by a control law $v_i(x)$. This is very easy, and $\dot{x}$ may even be assigned to be any desired function, for instance, $-x$.

(b) Use the approximation results and build the controls $u_k(\varepsilon, t, v_1(x), v_2(x), v_3(x),\; v_4(x),\; v_5(x))$, according to the process given in [25, 9] so that when ε tends to zero, the system (1.1) controlled with these controls "tends to" the extended system (1.3) controlled with the controls $v_i(x)$.

(c) Since the limit system is asymptotically stable (for instance, $\dot{x} = -x$), and asymptotic stability is somehow robust, the constructed control laws are, it is to be hoped, stabilizing for ε nonzero but small enough. For instance, one may take $\|x\|^2$ as a Lyapunov function for the limit system, its time-derivative along the limit system is $-2\|x\|^2$, and it is tempting to believe that its time-derivative along the original system controlled by $u_k(\varepsilon, t, v_1(x), v_2(x), v_3(x), v_4(x), v_5(x))$ is no larger than $-\|x\|^2$ for ε small enough.

Unfortunately, these arguments, which would have been somewhat simpler than those in [3], are not rigorous as they stand. The meaning of "tends to" in point (b) is very imprecise. In [5], and in [25, 9], only uniform convergence of the trajectories on finite-time intervals are considered. This is not adequate for asymptotic stabilization. The Lyapunov function-based argument in point (c) does not work because, in general, when ε tends to zero, the time derivative of a given function along the system (1.1) in feedback with the controls u_k from point (b) does not tend to the time-derivative of this function along the "limit" system (1.3). In addition, the fact that feedback controls are considered instead of open-loop controls complicates the proofs because the controls depend on the state and therefore may have a very high derivative with respect to time not only through the high frequencies and amplitudes built into the approximation process but also through their dependence on the state, whose speed is proportional to these high amplitudes.

However, we show in the present paper that the above sketch is basically correct, provided that homogeneous controls associated with a homogeneous Lyapunov function are used and that the construction of the approximating sequence is modified to take into account the closed-loop nature of the controls. An argument of the type of point (c) is possible based on a notion of approximation that is not in terms of uniform convergence of trajectories, but in terms of the differential operator defined by derivation along the system.

This paper is organized as follows. After a brief recall of technical material in section 2, we state in section 3 the control objective, make homogeneity assumptions, and explain how they will yield local results for general controllable systems. The design method is developed in section 4 through four steps: choice of the "useful" Lie brackets, construction of the stabilizing controls for the extended system (system (1.3) in the above example), construction of the "state dependent" amplitudes for the feedback law, and construction of the oscillatory controls by the method exposed in [9]; the material from these steps is then gathered to give the control law, and the stabilization result is stated. We present in this section all that is needed for the construction of the control law, but the proofs of some properties needed at each steps, and of the theorem, are given separately in section 7. Section 6 is devoted to a convergence result needed in the proof of the stability theorem; it is a translation in terms of differential operators (instead of trajectories) of the averaging results presented in [25, 9, 26] and in [8]. An illustrative example is given in section 5.

2. Background on homogeneous vector fields. For any $\lambda > 0$, the "dilation operator" δ_λ associated with a "weight vector" $r = (r_1, \ldots, r_n)$ $(r_i > 0)$ is defined on $\mathbb{R}^n$ by

$$\delta_\lambda(x_1, \ldots, x_n) = (\lambda^{r_1} x_1, \ldots, \lambda^{r_n} x_n).$$

A function $f \in C^o(\mathbb{R}^n; \mathbb{R})$ is said to be homogeneous of degree τ with respect to the family of dilations (δ_λ) if

$$\forall \lambda > 0, \quad f(\delta_\lambda(x)) = \lambda^\tau f(x)\,.$$

A *homogeneous norm* is any proper continuous positive function that is homogeneous of degree 1.

A continuous vector field X on $\mathbb{R}^n$ is said to be homogeneous of degree σ with respect to the family of dilations (δ_λ) if one of the following equivalent properties is satisfied:

(1) For any $i = 1, \ldots, n$, its ith component, i.e., the function $x \mapsto X_i(x)$, is homogeneous of degree $r_i + \sigma$.
(2) For any function h homogeneous of degree $\tau > 0$ with respect to the same dilation, the function $L_X h$ (its Lie derivative along X) is homogeneous of degree $\sigma + \tau$.
(3) For all positive constant λ, the vector field $((\delta_\lambda)_* X)$, conjugate of X by the diffeomorphism δ_λ —away from the origin— satisfies $((\delta_\lambda)_* X)(x) = \lambda^{-\sigma} X(x)$ for $x \neq 0$.

The previous definitions of homogeneity can be extended to time-varying functions and vector fields by considering an "extended dilation":

$$\delta_\lambda(x_1, \ldots, x_n, t) = (\lambda^{r_1} x_1, \ldots, \lambda^{r_n} x_n, t).$$

Finally, let $f \in C^0(\mathbb{R}^n \times \mathbb{R}; \mathbb{R}^n)$, with $f(x, .)$ T-periodic, defining a homogeneous vector field of degree zero with respect to a family of dilations (δ_λ). Then, the two following properties are equivalent (see [7] for the autonomous case):

(i) the origin $x = 0$ of the system $\dot{x} = f(x, t)$ is locally asymptotically stable.
(ii) $x = 0$ is globally *ρ-exponentially asymptotically stable*, i.e., for any homogeneous norm ρ, there exist $K, \gamma > 0$ such that, for any solution $x(.)$ of the system,

$$\rho(x(t)) \leq K \rho(x(0)) e^{-\gamma t}.$$

In what follows, when using the expression *exponentially asymptotically stable*, we will refer to the ρ-exponential asymptotic stability defined above.

3. Problem statement. Consider a smooth driftless controllable system

$$\dot{x} = \sum_{i=1}^{m} u_i f_i(x). \tag{3.1}$$

In general, there does not exist a dilation with respect to which the control vector fields are homogeneous. However, controllability implies that after some adequate change of coordinates, there exist a dilation and a controllable *homogeneous approximation* [6, 7]—with respect to this dilation—of the system (3.1) around the origin. Different methods exist to find such a change of coordinates and dilation. For instance, a constructive method (i.e., requiring only algebraic computations and derivations) is given in [22]. Using this method, one obtains a driftless control system with control vector fields homogeneous of degree -1. Moreover, any homogeneous feedback law that asymptotically stabilizes this system also locally asymptotically stabilizes the original system.

The present work constructs a homogeneous feedback that ensures global exponential stabilization for homogeneous systems. Applied to the homogeneous approximation of a general system (3.1), it provides *local* exponential stabilization of (3.1).

Throughout this paper, we always consider a system

$$\dot{x} = \sum_{i=1}^{m} u_i b_i(x), \tag{3.2}$$

where the b_i's are smooth vector fields and the system of coordinates is such that there exist some integers $(r_1, \dots, r_n)$ such that

(1) each vector field b_i is homogeneous of degree -1 with respect to the family of dilations δ_λ with weights $(r_1, \dots, r_n)$;
(2) the rank at the origin of the Lie algebra generated by the b_i's is n:

$$\text{Rank}(\text{Lie}\{b_1, \dots, b_m\}(0)\ ) \ = \ n. \tag{3.3}$$

The integer valued weights $r_1, \dots, r_n$ are now fixed, and we denote

$$P \ = \ \text{Max}\,\{r_i;\ i = 1, \dots, n\}. \tag{3.4}$$

Our objective is to design feedback laws $u = (u_1, \dots, u_m) \in C^0(\mathbb{R} \times \mathbb{R}^n; \mathbb{R}^m)$ such that the origin $x = 0$ of the closed-loop system (3.2) is exponentially asymptotically stable.

Remark 3.1. We require only full rank control Lie algebra at the origin, but controllability follows, because homogeneity allows us to deduce the same rank condition everywhere.

Remark 3.2. We assume that the degrees are all equal to -1. These are the degrees given by the construction of a homogeneous approximation in [22]. If a system is naturally homogeneous, but the degrees are not all equal (if they are equal, a simple scaling makes them all equal to -1), it might be better to use this natural homogeneity than to construct a different homogeneous approximation that will have all the degrees equal to -1. The present method can be adapted to the case when the degrees of homogeneity are not all equal; this requires only a modification of the first step (see Remark 4.3).

4. Controller design. The control design consists of four steps described below.

Step 1 *(selection of Lie brackets).*

In this step, we select some vector fields $\tilde{b}_j$ $(j = 1, \dots, N)$, obtained as Lie brackets of the control vector fields $b_1, \dots, b_m$. The $\tilde{b}_j$ are chosen recursively as follows. For any $p = 1, \dots, P$ (with P defined by (3.4)),

(1) compute all brackets of length p made from the control vector fields b_i $(i = 1, \dots, m)$;
(2) select among the vector fields so obtained a maximal number of vector fields independent[1] over $\mathbb{R}$. These vector fields are the $\tilde{b}_j$ $(m_{p-1} + 1 \le j \le m_p)$. (We set $m_0 = 0$ so that all the integers m_p $(p = 0, \dots, P)$ are defined, with $N = m_P$.)

It follows from this construction that with each vector field $\tilde{b}_j$ we can associate a Lie bracket of some b_i's, i.e.,

$$\tilde{b}_j \ = \ \mathcal{C}_j(b_{\tau_j^1}, \dots, b_{\tau_j^{\ell(j)}})\,, \tag{4.1}$$

with

- $\mathcal{C}_j$ a formal bracket and $b_{\tau_j^1}, \dots, b_{\tau_j^{\ell(j)}}$ the elements that are bracketed (listed in the order they appear in the bracket);
- $\ell(j)$ the number of vector fields that are bracketed in (4.1), i.e.,

$$\ell(j) = p \quad \Leftrightarrow \quad m_{p-1} + 1 \le j \le m_p.$$

[1] Recall that some vector fields $X_1, \cdots, X_r$ are said to be *linearly independent over* $\mathbb{R}$ if and only if for any $(\lambda_1, \dots, \lambda_r)$ in $\mathbb{R}^r$, the vector field $\lambda_1 X_1 + \cdots + \lambda_r X_r$ is identically zero on $\mathbb{R}^n$ only if $\lambda_1 = \cdots = \lambda_r = 0$.

For instance, if we choose a vector field $\tilde{b}_6 = [[b_2, b_1], [b_1, [b_1, b_2]]]$, then we encode this as (4.1) with $\ell(6) = 5$, $\tau_6^2 = \tau_6^3 = \tau_6^4 = 1$, $\tau_6^1 = \tau_6^5 = 2$, and the symbol $\mathcal{C}_6$ defined by $\mathcal{C}_6(z_1, z_2, z_3, z_4, z_5) = [[z_1, z_2], [z_3, [z_4, z_5]]]$. This notation is sloppy but avoids using formal Lie brackets and the evaluation operator (see [24]) from a free Lie algebra to vector fields, which would make the exposition uselessly heavy. Of course, the decomposition (4.1) is not unique in general. From now on, we consider that one decomposition has been chosen and that the $\mathcal{C}_j$'s and τ_j^k's have been defined accordingly.

Remark 4.1. (1) In Step 1 above, we do not need to compute all brackets of length p. More precisely, let $\mathcal{F}$ denote the free Lie algebra generated by some indeterminates $s_1, \ldots, s_m$. Then, one can select a basis $\mathcal{B}$ of this Lie algebra (for instance a P. Hall basis, as used by Sussmann and Liu [25, 26] and Liu [9]). If $\mathcal{B}_p$ denotes the elements of $\mathcal{B}$ of order p, then it is clearly sufficient to consider Lie brackets of the b_i obtained by *evaluating* (in the sense of [24]) the elements of $\mathcal{B}_p$ at $s_i = b_i$ $(i = 1, \ldots, m)$. One usually takes this into account when checking controllability.

(2) Since the vector fields b_i $(i = 1, \ldots, m)$ are homogeneous of degree -1, each bracket of length p of these vector fields is homogeneous of degree $-p$. Moreover, the weights of the dilation being integers, any smooth vector field homogeneous of integer degree is, in fact, polynomial. Using a (finite) basis of the polynomials homogeneous of degree k ($k \in \{0, \ldots, P-1\}$), selecting Lie brackets of a given length consists only of computing a basis of a finite dimensional vector space.

(3) We do not need to consider brackets of order larger than P because they are identically zero; indeed, all components of these vector fields are homogeneous of negative degree and, therefore, they would tend to infinity at the origin if they were not identically zero.

Example. Let us illustrate this step on the following academic example:

$$\begin{aligned} \dot{x}_1 &= u_1, \\ \dot{x}_2 &= x_3^2(u_1 + u_2), \\ \dot{x}_3 &= u_3, \end{aligned}$$

which is of the form (3.2) with $m = 3$ and

$$b_1 = \frac{\partial}{\partial x_1} + x_3^2 \frac{\partial}{\partial x_2}, \quad b_2 = x_3^2 \frac{\partial}{\partial x_2}, \quad b_3 = \frac{\partial}{\partial x_3}.$$

The control vector fields are homogeneous of degree -1 with respect to the dilation with weights $r_1 = 1$, $r_2 = 3$, and $r_3 = 1$.

For the brackets of length 1, i.e., the control vector fields, b_1 and b_3 are independent at the origin while b_2 is zero at the origin but independent from b_1 and b_3 away from $x_3 = 0$. Hence $m_1 = 3$, and one can take $\tilde{b}_1 = b_1 = \mathcal{C}_1(b_1)$, $\tilde{b}_2 = b_2 = \mathcal{C}_2(b_2)$, and $\tilde{b}_3 = b_3 = \mathcal{C}_3(b_3)$.

At length 2 all the brackets vanish at the origin, but they are not identically zero: $[b_2, b_3] = -2x_3 \frac{\partial}{\partial x_2}$, and $[b_3, b_1] = -[b_2, b_3]$. Since $[b_1, b_2] = 0$, we have $m_2 = 4$. We define, for instance, $\tilde{b}_4 = [b_2, b_3] = \mathcal{C}_4(b_2, b_3)$.

Finally, since $[b_3, [b_2, b_3]] = -2\frac{\partial}{\partial x_2}$, $m_3 = 5$ with, for instance, $\tilde{b}_5 = [b_3, [b_2, b_3]] = \mathcal{C}_5(b_3, b_2, b_3)$. Note that here, due to the origin being a singular point for the distributions spanned by the control vector fields and by the brackets of order at most 2, N is strictly larger than n.

With this general construction, we have the following proposition.

PROPOSITION 4.2. *For any family* $(\tilde{b}_j)_{j=1,\dots,N}$ *defined as above, we have the following:*

(a) *Let* $j_1, \dots, j_n$ *be such that* $\mathrm{Span}\{\tilde{b}_{j_1}(0), \dots, \tilde{b}_{j_n}(0)\} = \mathbb{R}^n$. *Then*

$$\forall x \in \mathbb{R}^n, \quad \mathrm{Span}\{\tilde{b}_{j_1}(x), \dots, \tilde{b}_{j_n}(x)\} = \mathbb{R}^n .$$

(b) *Any vector field* b *that can be written as a Lie bracket of order* p *of some* b_i*'s is a linear combination of the* $\tilde{b}_j$*'s with* $\ell(j) = p$, *i.e.,*

$$b = \sum_{j=m_{p-1}+1}^{m_p} \lambda_j \tilde{b}_j = \sum_{\ell(j)=p} \lambda_j \tilde{b}_j$$

for some real numbers $\lambda_j \in \mathbb{R}$.

(c) *The vector fields* $\{\tilde{b}_j\}_{j=1,\dots,N}$ *are linearly independent over* $\mathbb{R}$.

(The proof is in section 7.1.)

Remark 4.3. If the degrees of the vector fields b_i are not all equal, the above construction has to be modified. More precisely, in the recursive construction of the family $(\tilde{b}_j)_{j=1,\dots,N}$, we have to consider an induction on the degree of homogeneity instead of an induction on the length of the Lie brackets. (Note that this is just a generalization of the above construction, since for vector fields of the same degree -1 the set of Lie brackets of length p is the same as the set of Lie brackets of degree $-p$.) This means that at each step, we have to compute the set of Lie brackets of a certain degree and select from among them a finite number of vector fields that form a basis of this set.

Step 2 *(stabilization of the extended system).*

Let a be a smooth vector field, homogeneous of degree zero with respect to the family of dilations (δ_λ), and such that the origin $x = 0$ of the system $\dot{x} = a(x)$ is asymptotically stable. One may take, for instance, $a(x) = -x$. In view of Proposition 4.2(a), the $n \times n$ matrix whose columns are $\tilde{b}_{j_1}(x), \dots, \tilde{b}_{j_n}(x)$ is invertible for all x. Define the functions $\tilde{u}_j$ $(j = 1, \dots, N)$ by

$$\bullet \begin{pmatrix} \tilde{u}_{j_1}(x) \\ \vdots \\ \tilde{u}_{j_n}(x) \end{pmatrix} = \left(\tilde{b}_{j_1}(x), \dots, \tilde{b}_{j_n}(x)\right)^{-1} a(x), \tag{4.2}$$

$$\bullet\ \tilde{u}_j = 0 \quad \forall j \notin \{j_1, \dots, j_n\}.$$

These functions are obviously such that

$$a = \sum_{j=1}^{N} \tilde{u}_j \, \tilde{b}_j, \tag{4.3}$$

and furthermore, we may state this proposition.

PROPOSITION 4.4. *For any* $j = 1, \dots, N$, *the above-constructed function* $\tilde{u}_j$ *is in* $C^\infty(\mathbb{R}^n - \{0\}; \mathbb{R}) \cap C^0(\mathbb{R}^n; \mathbb{R})$ *and is homogeneous of degree* $\ell(j)$.

Proof. Continuity and smoothness away from the origin are inherited from the vector fields $\tilde{b}_j$ and the vector field a. Each $\tilde{u}_{j_k}$ is homogeneous of degree $\ell(j_k)$ because the lth component of the vector field a is homogeneous of degree r_l and the element (k, l) of the matrix $(\tilde{b}_{j_1}(x), \dots, \tilde{b}_{j_n}(x))^{-1}$ is homogeneous of degree $\ell(j_k) - r_l$. This

last statement is true because the element (k,l) of the matrix $(\tilde{b}_{j_1}(x),\dots,\tilde{b}_{j_n}(x))$ is homogeneous of degree $r_l-\ell(j_k)$ for the vector field; $\tilde{b}_{j_k}$ is an iterated Lie bracket of $\ell(j_k)$ homogeneous vector fields of degree -1 and hence is homogeneous of degree $-\ell(j_k)$. □

Step 3 *(construction of the state-dependent amplitudes).*

This step consists of finding some functions $v_j^k \in C^\infty(\mathbb{R}^n-\{0\};\mathbb{R})\cap C^0(\mathbb{R}^n;\mathbb{R})$ $(j=1,\dots,N,\ k=1,\dots\ell(j))$ homogeneous of degree one and such that

$$\sum_{j=1}^{N}\tilde{u}_j\,\mathcal{C}_j(b_{\tau_j^1},\dots,b_{\tau_j^{\ell(j)}})=\sum_{j=1}^{N}\mathcal{C}_j(b_{\tau_j^1}v_j^1,\dots,b_{\tau_j^{\ell(j)}}v_j^{\ell(j)}). \tag{4.4}$$

Recall that the $\mathcal{C}_j$'s, defined in Step 1, are the brackets associated with the $\tilde{b}_j$'s, i.e.,

$$\tilde{b}_j=\mathcal{C}_j(b_{\tau_j^1},\dots,b_{\tau_j^{\ell(j)}})\,. \tag{4.5}$$

The construction of the functions v_j^k is based on the following lemma.

LEMMA 4.5. *Let $\mathcal{C}(b_{i_1},\dots,b_{i_p})$ $(i_k\in\{1,\dots,m\})$ be any Lie bracket of some vector fields b_{i_k} $(i_k\in\{1,\dots,m\})$, and $v_k\in C^\infty(\mathbb{R}^n-\{0\};\mathbb{R})\cap C^0(\mathbb{R}^n;\mathbb{R})$ $(k=1,\dots,p)$ some functions homogeneous of degree* 1*. Then,*

(i) $\mathcal{C}(b_{i_1}v_1,\dots,b_{i_p}v_p)=v_1\dots v_p\,\mathcal{C}(b_{i_1},\dots,b_{i_p})-\sum_{j=1}^{m_{p-1}}h_j\ \tilde{b}_j$;

(ii) *for any $j=1,\dots,m_{p-1}$, $h_j\in C^\infty(\mathbb{R}^n-\{0\};\mathbb{R})\cap C^0(\mathbb{R}^n;\mathbb{R})$ is homogeneous of degree $\ell(j)$.*

The proof of this lemma, left to the reader, follows from Proposition 4.2(b) by a direct induction on the length p of the bracket $\mathcal{C}(b_{i_1}v_1,\dots,b_{i_p}v_p)$. It is a generalization of the fact that for two functions v_1 and v_2, and vector fields b_{i_1} and b_{i_2}, $[v_1b_{i_1},v_2b_{i_2}]=v_1v_2[b_{i_1},b_{i_2}]-v_2(L_{b_{i_2}}v_1)b_{i_1}+v_1(L_{b_{i_1}}v_2)b_{i_2}$.

Note that the functions h_j in Lemma 4.5 can be explicitly computed by expressing brackets of order not larger than $p-1$ as linear combinations of $\tilde{b}_1,\dots,\tilde{b}_{m_{p-1}}$.

Based on Lemma 4.5, the functions v_j^k can be constructed recursively as follows.

Step $p=P$: For any $j\in\{m_{P-1}+1,\dots,m_P\}$, we define

$$v_j^P=\frac{\tilde{u}_j}{\rho^{P-1}}\ \text{ and }\ v_j^k=\rho\ (k=1,\dots,P-1)\,, \tag{4.6}$$

with ρ any homogeneous norm in $C^\infty(\mathbb{R}^n-\{0\};\mathbb{R})\cap C^0(\mathbb{R}^n;\mathbb{R})$ (for instance, one may take $\rho(x)=(\sum|x_i|^{\frac{q}{r_i}})^{\frac{1}{q}}$ with $q=2\prod_{i=1}^n r_i$).

In view of (4.5), (4.6), and Lemma 4.5, we have

$$\sum_{j=m_{P-1}+1}^{m_P}\mathcal{C}_j(b_{\tau_j^1}v_j^1,\dots,b_{\tau_j^P}v_j^P)\ =\ \sum_{j=m_{P-1}+1}^{m_P}\tilde{u}_j\tilde{b}_j-\sum_{j=1}^{m_{P-1}}h_j^P\tilde{b}_j \tag{4.7}$$

with h_j^P $(j=1,\dots,m_{P-1})$ obtained by expanding the brackets in the left-hand side of (4.7) with respect to the variables v_j^k and their derivatives.

Step $1\le p<P$: Assume that the functions v_j^k $(j=m_p+1,\dots,m_P,\ k=1,\dots,\ell(j))$ and h_j^k $(j=m_p+1,\dots,m_P,\ k=p+1,\dots,P)$ have been computed in Steps P to $p+1$ and satisfy the induction assumption

$$\sum_{j=m_p+1}^{N}\mathcal{C}_j(b_{\tau_j^1}v_j^1,\dots,b_{\tau_j^{\ell(j)}}v_j^{\ell(j)})\ =\ \sum_{j=m_p+1}^{N}\tilde{u}_j\tilde{b}_j-\sum_{j=1}^{m_p}h_j^{p+1}\tilde{b}_j\,. \tag{4.8}$$

We define, for any $j \in \{m_{p-1}+1, \dots, m_p\}$,

$$v_j^p = \frac{1}{\rho^{p-1}}(\tilde{u}_j + h_j^{p+1}) \text{ and } v_j^k = \rho \ (k = 1, \dots, p-1)\,. \tag{4.9}$$

In view of (4.5), (4.9), and Lemma 4.5, we have

$$\sum_{j=m_{p-1}+1}^{m_p} \mathcal{C}_j(b_{\tau_j^1} v_j^1, \dots, b_{\tau_j^p} v_j^p) = \sum_{j=m_{p-1}+1}^{m_p} (\tilde{u}_j + h_j^{p+1})\tilde{b}_j + \sum_{j=1}^{m_{p-1}} (h_j^{p+1} - h_j^p)\tilde{b}_j \tag{4.10}$$

for an adequate choice of the h_j^p $(j = m_{p-1}+1, \dots, m_p)$ obtained again by expanding the brackets in the left-hand side of (4.7) with respect to the variables v_j^k and their derivatives. In view of (4.8) and (4.10), we have

$$\sum_{j=m_{p-1}+1}^{N} \mathcal{C}_j(b_{\tau_j^1} v_j^1, \dots, b_{\tau_j^{\ell(j)}} v_j^{\ell(j)}) = \sum_{j=m_{p-1}+1}^{N} \tilde{u}_j \tilde{b}_j - \sum_{j=1}^{m_{p-1}} h_j^p \tilde{b}_j \tag{4.11}$$

so that the induction assumption (4.8) on Steps P to $p+1$ is also true for Steps P to p.

The computation of the functions v_j^k and h_j^k ends after Step $p = 1$ has been performed. Let us remark that in the last step ($p = 1$), there is no function h_j^p to compute. With this construction, we have the next proposition.

PROPOSITION 4.6. *Consider the functions v_j^k defined above. Then*

(a) *each v_j^k $(j = 1, \dots, N,\ k = 1, \dots, \ell(j))$ belongs to $C^\infty(\mathbb{R}^n - \{0\}; \mathbb{R}) \cap C^0(\mathbb{R}^n; \mathbb{R})$ and is homogeneous of degree 1;*

(b) *(4.4) is satisfied.*

Proof. Point (b) is a direct consequence of (4.11) with $p = 1$. Point (a) is an easy consequence of Proposition 4.4, equations (4.6) and (4.9), and Lemma 4.5. ∎

Step 4 (oscillatory approximation of Lie brackets).

The last step of our construction relies on the work of Liu [9] and Sussmann and Liu [25, 26]. More precisely, consider a control system

$$\dot{x} = \sum_{\alpha=1}^{A} u_\alpha X_\alpha(x) \tag{4.12}$$

with $X_1, \dots, X_A$ some smooth vector fields on a smooth n-dimensional manifold, a "Lie bracket extended" system

$$\dot{x} = \sum_{\beta=1}^{B} w_\beta X_\beta(x) \qquad (B \geq A), \tag{4.13}$$

where the A first vector fields are the same as in (4.12), and the other vector fields are Lie brackets of $X_1, \dots, X_A$. In [9], an algorithm is given that builds, for any set of integrable functions of time w_β $(\beta = 1, \dots, B)$, some "highly oscillatory" functions of time u_α^ε such that the trajectories of (4.12), with $u_\alpha = u_\alpha^\varepsilon$, approximate those of (4.13).

We do not describe this algorithm here, though we use the notation

$$u_\alpha^\varepsilon = \mathcal{F}(\alpha, \varepsilon, (w_\beta)_{1 \leq \beta \leq B}), \tag{4.14}$$

where $\mathcal{F}$ is a function described algorithmically in [9]. It depends only on which Lie brackets have to be performed to obtain the vector fields $X_{A+1}, \ldots, X_B$ from the vector fields $X_1, \ldots, X_A$. It is of the form

$$u_\alpha^\varepsilon(t) \;=\; \eta_{\alpha,0}(t) + \varepsilon^{-\frac{1}{2}} \sum_{\omega\in\Omega(2,\alpha)} \eta_{\omega,\alpha}(t) e^{i\omega t/\varepsilon} + \sum_{n=3}^{N} \varepsilon^{\frac{n-1}{n}} \sum_{\omega\in\Omega(n,\alpha)} \eta_\omega(t) e^{i\omega t/\varepsilon}, \tag{4.15}$$

with N the length of the higher order bracket X_β in (4.13), $\eta_{\alpha,0}, \eta_{\omega,\alpha}$, and η_ω some functions, and $\Omega_{2,\alpha}$, $\Omega_{n,\alpha}$ some finite subsets of $\mathbb{R}$, that are all built precisely in [9]. In particular, the construction of the "approximating inputs" u_α^ε given in [9] implies the following.

THEOREM 4.7 (see [9]). *For any T $(0 < T < +\infty)$ and any family w_β $(\beta = 1, \ldots, B)$ of integrable functions on $[0,T]$, the functions u_α^ε $(\alpha = 1, \ldots, A)$ given by (4.14), where $\mathcal{F}$ symbolizes the algorithm described in [9], are integrable and are such that the trajectories of (4.12)–(4.15) converge to the trajectories of (4.13) in the following sense: For any $p \in \mathbb{R}^n$, if the system (4.13) with $x(0) = p$ has a unique solution x^∞ defined on $[0,T]$ and if x_ε is a maximal solution of system (4.12)–(4.15) with $x(0) = p$, then x_ε is defined on $[0,T]$ for ε small enough and converges uniformly to x^∞ on $[0,T]$ as $\varepsilon \to 0$.*

Remark 4.8. (1) The functions u_α^ε in (4.15) are real-valued because each $\Omega_{n,\alpha}$ $(n = 2, \ldots, N)$ is symmetric $(\omega \in \Omega_{n,\alpha} \Rightarrow -\omega \in \Omega_{n,\alpha})$, $\eta_{-\omega} = \overline{\eta_\omega}$, and $\eta_{-\omega,\alpha} = \overline{\eta_{\omega,\alpha}}$.

(2) If the functions w_β in (4.13) are constant, the functions $\eta_{\alpha,0}, \eta_{\omega,\alpha}$, and η_ω are also constant.

Consider now the following two systems:

$$\dot{x} = \sum_{j=1}^{N} \sum_{s=1}^{\ell(j)} u_{j,s} b_{\tau_j^s} v_j^s\,, \tag{4.16}$$

$$\dot{x} = \sum_{j=1}^{N} \mathcal{C}_j(b_{\tau_j^1} v_j^1, \ldots, b_{\tau_j^{\ell(j)}} v_j^{\ell(j)})\,. \tag{4.17}$$

Systems (4.16) and (4.17) are of the same form as (4.12) and (4.13), respectively, with the vector fields X_α being the $b_{\tau_j^s} v_j^s$'s (with α a double index (j,s)), the vector fields X_β being these plus the brackets in (4.17), i.e., $\mathcal{C}_j(X_{j,1}, \ldots, X_{j,\ell(j)})$, $1 \le j \le N$, and each w_β in (4.13) being constant: 0 in front of the X_β's that are also X_α's and 1 in front of the added brackets. Note that since each original vector field from (3.2) appears many times in the brackets selected in Step 1, we consider here as independent control vector fields in (4.12) some vector fields that are in fact "multiples" of each other: for instance if the vector field b_1 appears more than one time, we have $\tau_j^s = \tau_{j'}^{s'} = 1$ for some $(j,x) \neq (j',s')$, and $v_j^s b_1$ and $v_{j'}^{s'} b_1$ are distinct control vector fields X_α in (4.12).

Following Liu's algorithm, we construct some functions

$$u_{j,s}^\varepsilon \;=\; \mathcal{F}\left((j,s), \varepsilon, (0, \ldots, 0, 1, \ldots, 1)\right),$$

where $\mathcal{F}$ is the notation introduced in (4.14), such that the trajectories of (4.16)–(4.18) (which exist on any time interval because the system is degree zero homogeneous) converge uniformly on any time interval $[0,T]$ to those of (4.17), as ε tends to zero.

Recall (see (4.15)) that they are of the form

$$u^\varepsilon_{j,s}(t) = \eta_{j,s,0} + \varepsilon^{-\frac{1}{2}} \sum_{\omega\in\Omega(2,j,s)} \eta_{\omega,j,s} e^{i\omega t/\varepsilon} + \sum_{n=3}^{P} \varepsilon^{\frac{n-1}{n}} \sum_{\omega\in\Omega(n,j,s)} \eta_\omega e^{i\omega t/\varepsilon}\,. \tag{4.18}$$

Note that the functions η in (4.18) are constant in view of Remark 4.8 above. We rewrite system (4.16)–(4.18) as

$$\dot{x} = \sum_{i=1}^{m} \left(\sum_{(j,s):\tau^s_j=i} u^\varepsilon_{j,s}(t) v^s_j(x) \right) b_i(x)\,. \tag{4.19}$$

Our final control laws are defined by

$$u^\varepsilon_i(x,t) = \sum_{(j,s):\tau^s_j=i} u^\varepsilon_{j,s}(t) v^s_j(x)\,. \tag{4.20}$$

As stated in the following theorem, they ensure asymptotic stability of system (3.2) for "sufficiently large" frequencies.

THEOREM 4.9. *Let the controls u^ε_i be these described above. Then, the vector field in the right-hand side of the time-varying closed-loop system*

$$\dot{x} = \sum_{i=1}^{m} u^\varepsilon_i(x,t)\, b_i(x) \tag{4.21}$$

is homogeneous of degree zero and, for $\varepsilon > 0$ sufficiently small, the origin is exponentially uniformly asymptotically stable. (See the proof in section 7.3.)

Remark 4.10. Our construction a priori implies uniform convergence of the trajectories of (4.21) to those of (4.17), the origin of which is asymptotically stable from (4.3) to (4.5). However, this is not enough to infer asymptotic stability of (4.21). In the proofs, and in section 6, we introduce a stronger kind of convergence (DO-convergence), sufficient to infer asymptotic stability of (4.21). However, we quote uniform convergence here (instead of the DO-convergence, which we really need) because we base our construction on [9]. It makes the present construction clearer. (To construct the controls, one needs only to follow the algorithm in [9]; the kind of convergence does not matter.) Also, using the convergence result from [9] (Theorem 4.7) provides a shortcut in the proof on DO-convergence. This may make the paper less self-contained, but it avoids reproducing some difficult calculations made in [9].

5. An illustrative example. We now illustrate the control design method shown in section 4. Let us consider the following system in $\mathbb{R}^4$:

$$\dot{x} = b_1 u_1 + b_2 u_2\,, \tag{5.1}$$

with $b_1 = \frac{\partial}{\partial x_1} + x_3 \frac{\partial}{\partial x_2} + x_4 \frac{\partial}{\partial x_3}$ and $b_2 = \frac{\partial}{\partial x_4}$, which can be used to model the kinematic equations of a car-like mobile robot. One easily verifies that the vector fields b_1 and b_2 are homogeneous of degree -1 with respect to the family of dilations of weight $r = (1,3,2,1)$, and that this system is controllable. We follow this example with the four steps of our control design procedure.

Step 1. Since $[b_1,b_2] = -\frac{\partial}{\partial x_3}$, $[b_1,[b_1,b_2]] = \frac{\partial}{\partial x_2}$, and $[b_2,[b_2,b_1]] = 0$, the family $(\tilde{b}_j)$ is directly given by

$$\begin{aligned}(\tilde{b}_j) = (\tilde{b}_1,\tilde{b}_2,\tilde{b}_3,\tilde{b}_4) &= (b_1, b_2, [b_1,b_2], [b_1,[b_1,b_2]]) \\ &= (\mathcal{C}_1(b_{\tau^1_1}), \mathcal{C}_2(b_{\tau^1_2}), \mathcal{C}_3(b_{\tau^1_3}, b_{\tau^2_3}), \mathcal{C}_4(b_{\tau^1_4}, b_{\tau^2_4}, b_{\tau^3_4})).\end{aligned} \tag{5.2}$$

This implies that $\tau_1^1 = 1, \tau_2^1 = 2, \tau_3^1 = 1, \tau_3^2 = 2, \tau_4^1 = \tau_4^2 = 1, \tau_4^3 = 2$, and that $m_1 = 2$, $m_2 = 3$, and $m_3 = N = 4$.

Step 2. Let us, for instance, define the vector field a by $a(x) = -x$. (The origin $x = 0$ of $\dot{x} = a(x)$ is obviously asymptotically stable.) Then the integers j_k are simply defined by $j_k = k$ $(k = 1, \ldots, 4)$. By a direct computation, one obtains the following expression for the functions $\tilde{u}_j$:

$$\begin{aligned}(\tilde{u}_1, \tilde{u}_2, \tilde{u}_3, \tilde{u}_4)^T(x) &= (\tilde{b}_1, \tilde{b}_2, \tilde{b}_3, \tilde{b}_4)^{-1}(x)\, a(x) \\ &= (-x_1, -x_4, -x_1x_4 + x_3, x_1x_3 - x_2)^T\,.\end{aligned} \tag{5.3}$$

Step 3. From Step 1, the brackets $\mathcal{C}_k$ are defined by

$$\mathcal{C}_1(x_1) = x_1, \mathcal{C}_2(x_2) = x_2, \mathcal{C}_3(x_1, x_2) = [x_1, x_2], \mathcal{C}_4(x_1, x_1, x_2) = [x_1, [x_1, x_2]]\,.$$

We now follow the procedure exposed in section 4.

Step $p = P = 3$: The functions v_4^1, v_4^2, and v_4^3 are given, in view of (4.6), by

$$v_4^1 = v_4^2 = \rho, \quad v_4^3 = \frac{\tilde{u}_4}{\rho^2}\,, \tag{5.4}$$

with $\rho \in C^\infty(\mathbb{R}^4 - \{0\}; \mathbb{R}) \cap C^0(\mathbb{R}^4; \mathbb{R})$ a homogeneous norm. (For instance, one may take $\rho(x) = (x_1^{12} + x_2^4 + x_3^6 + x_4^{12})^{\frac{1}{12}}$.)

We also compute the functions h_j^P involved in (4.7). A tedious but simple calculation gives

$$\begin{aligned} h_1^3 &= v_4^2 v_4^3 L_{[b_1,b_2]} v_4^1 + v_4^2 L_{b_1} v_4^3 L_{b_2} v_4^1 + v_4^1 L_{b_1}(v_4^3 L_{b_2} v_4^2) - v_4^3 L_{b_1} v_4^1 L_{b_2} v_4^2\,, \\ h_2^3 &= -v_4^1 L_{b_1}(v_4^2 L_{b_1} v_4^3)\,, \\ h_3^3 &= -v_4^1 L_{b_1} v_4^2 v_4^3 - v_4^1 v_4^2 L_{b_1} v_4^3\,. \end{aligned} \tag{5.5}$$

Step $p = 2$: The functions v_3^1 and v_3^2 are given, in view of (4.9), by

$$v_3^1 = \rho, \quad v_3^2 = \frac{(\tilde{u}_3 + h_3^3)}{\rho}\,. \tag{5.6}$$

The functions h_1^2 and h_2^2 defined by (4.10) can be computed using (5.6):

$$\begin{aligned} h_1^2 &= h_1^3 + v_3^2 L_{b_2} v_3^1\,, \\ h_2^2 &= h_2^3 - v_3^1 L_{b_1} v_3^2\,. \end{aligned} \tag{5.7}$$

Step $p = 1$: Finally, the functions v_1^1 and v_2^1 are defined, from (4.9) again, by

$$v_1^1 = \tilde{u}_1 + h_1^2, \quad v_2^1 = \tilde{u}_2 + h_2^2\,. \tag{5.8}$$

Step 4. First, we need to find functions $u_{j,s}\,(j = 1, \ldots, 4\,, s = 1, \ldots, \ell(j))$ such that the trajectories of the system

$$\dot{x} = \sum_{j=1}^{4} \sum_{s=1}^{\ell(j)} u_{j,s} b_{\tau_j^s} v_j^s \tag{5.9}$$

converge uniformly to those of the system

$$\dot{x} = \sum_{j=1}^{4} \mathcal{C}_j(b_{\tau_j^1} v_j^1, \ldots, b_{\tau_j^{\ell(j)}} v_j^{\ell(j)})\,. \tag{5.10}$$

We remark that, in view of (5.2), (5.4), and (5.6), the vector fields $b_{\tau_3^1} v_3^1, b_{\tau_4^1} v_4^1$, and $b_{\tau_4^2} v_4^2$ are in fact identical. As a consequence, there are only 5—not 7 (the number of terms in the sum (5.9))—different vector fields in (5.9) or (5.10). Therefore, the system (5.9) can be rewritten as

$$\dot{x} = \sum_{i=1}^{5} u_i X_i \tag{5.11}$$

with $X_1 = b_1 v_1^1, X_2 = b_2 v_2^1, X_3 = b_1 v_3^1 = b_1 v_4^1 = b_1 v_4^2, X_4 = b_2 v_3^2$, and $X_5 = b_2 v_4^3$, and u_1, u_2, u_3, u_4, u_5 standing, respectively, for $u_{1,1}$, $u_{2,1}$, $u_{3,1} + u_{4,1} + u_{4,2}$, $u_{3,2}$, and $u_{4,3}$, and the system (5.10) can then be rewritten as

$$\dot{x} = X_1 + X_2 + [X_3, X_4] + [X_3, [X_3, X_5]]\,. \tag{5.12}$$

We choose some candidate functions u_i, for the approximation of trajectories of (5.12) by solutions of (5.11), of the following form:

$$\left\{ \begin{array}{rcl} u_1(t) &=& \eta_{1,0}\,, \\ u_2(t) &=& \eta_{2,0}\,, \\ u_3(t) &=& \varepsilon^{-\frac{1}{2}} \eta_{\omega_{1,1}} \cos \omega_{1,1} t/\varepsilon + \varepsilon^{-\frac{2}{3}} \left(\eta_{\omega_{2,1}} \cos \omega_{2,1} t/\varepsilon + \eta_{\omega_{2,2}} \cos \omega_{2,2} t/\varepsilon \right), \\ u_4(t) &=& \varepsilon^{-\frac{1}{2}} \eta_{\omega_{1,2}} \sin \omega_{1,2} t/\varepsilon, \\ u_5(t) &=& \varepsilon^{-\frac{2}{3}} \eta_{\omega_{2,3}} \cos \omega_{2,3} t/\varepsilon \end{array} \right. \tag{5.13}$$

with $\omega_{k,j}$ defined for instance by

$$\begin{array}{rcl} \Omega_1 &=& \{\omega_{1,1}, \omega_{1,2}\} = \{\frac{7}{2}, -\frac{7}{2}\} \\ \text{and} \quad \Omega_2 &=& \{\omega_{2,1}, \omega_{2,2}, \omega_{2,3}\} = \{2, 3, -5\}. \end{array}$$

Note in particular that each set Ω_k is "minimally cancelling" (MC) in the sense of [9, 25, 26]. Using [9, Theorem 5.1] (see also [9, section 8]), where a very similar example is treated), one can show that the trajectories of system (5.11)–(5.13) converge to those of the system

$$\dot{x} = \eta_{1,0} X_1 + \eta_{2,0} X_2 - \frac{\eta_{\omega_{1,1}} \eta_{\omega_{1,2}}}{2\omega_{1,1}} [X_3, X_4] - \frac{\eta_{\omega_{2,1}} \eta_{\omega_{2,2}} \eta_{\omega_{2,3}}}{4\omega_{2,1}\omega_{2,2}} [X_3, [X_3, X_5]]\,. \tag{5.14}$$

In order to identify system (5.12) with system (5.14), one can, for instance, define

$$\eta_{1,0} = \eta_{2,0} = \eta_{\omega_{1,1}} = \eta_{\omega_{2,1}} = \eta_{\omega_{2,2}} = 1$$

and

$$\eta_{\omega_{1,2}} = -2\omega_{1,1},\ \eta_{\omega_{2,3}} = -4\omega_{2,1}\omega_{2,2}\,.$$

Expressing the right-hand term of (5.11) as a function of the control vector fields b_1 and b_2, we finally obtain the expression of our stabilizing feedbacks:

$$\left\{ \begin{array}{rcl} u_1^\varepsilon(x,t) &=& u_1(t/\varepsilon) v_1^1(x) + u_3(t/\varepsilon) v_3^1(x), \\ u_2^\varepsilon(x,t) &=& u_2(t/\varepsilon) v_2^1(x) + u_4(t/\varepsilon) v_3^2(x) + u_5(t/\varepsilon) v_4^3(x) \end{array} \right. \tag{5.15}$$

with the u_i's defined by (5.13) and the v_j^s's defined by (5.4), (5.6), and (5.8).

Although the above expression of the control laws appears quite simple, it is, in fact, quite involved due to the terms contained in the v_j^s's, and in particular due to the functions h_j^p defined by (5.5) and (5.7). This is a negative aspect of our construction: solving the equation (4.4) in the v_j^s's leads to heavy computations.

6. Convergence of highly oscillatory vector fields as differential operators. As explained in the introduction (section 1.2), the convergence results which are implicitly contained in [5], and explicitly in [25, 9] or [8], in terms of uniform convergence of solutions on finite time intervals, are not sufficient here. In this section, we state separately the convergence result that is used to prove Theorem 4.9. The word convergence is perhaps a bit farfetched since there is no notion of limit in the topological sense; the convergence is more of an algebraic nature: we simply decompose the operator as the sum of a nonoscillating term (the "limit") and a term which is a differential operator—of order higher than 1—whose coefficients are, when ε goes to zero and x remains in a compact set, $\mathcal{O}(\varepsilon^\gamma)$, with $\gamma > 0$. However, this result will prove to be sufficient for our needs. It is also sufficient to recover the uniform convergence stated in [5, 8, 25, 9]. In what follows, $\mathcal{T}$ denotes any time interval (possibly infinite).

DEFINITION 6.1. *Let $F^\varepsilon\,(\varepsilon \in (0,\varepsilon_0]\,\varepsilon_0 > 0)$ and F^0 be vector fields on $\mathbb{R}^{1+n}$, defined by $F^\varepsilon(t,x) = \frac{\partial}{\partial t} + f(\varepsilon,t,x)$ and $F^0(t,x) = \frac{\partial}{\partial t} + f^0(t,x)$ with $f \in \mathcal{C}^0((0,\varepsilon_0]\times\mathcal{T}\times\mathbb{R}^n;\mathbb{R}^n)\cap\mathcal{C}^\infty((0,\varepsilon_0]\times\mathcal{T}\times(\mathbb{R}^n-\{0\});\mathbb{R}^n)$, and $f^0 \in \mathcal{C}^0(\mathcal{T}\times\mathbb{R}^n;\mathbb{R}^n)\cap\mathcal{C}^\infty(\mathcal{T}\times(\mathbb{R}^n-\{0\});\mathbb{R}^n)$.*

We say that F^ε converges as a differential operator of order one on functions of t and x, in brief "DO-converges," to F^0, as $\varepsilon \longrightarrow 0$, if

$$F^\varepsilon = F^0 + \varepsilon^{\gamma_1}\left(F^\varepsilon D_1^\varepsilon - D_1^\varepsilon\frac{\partial}{\partial t}\right) + \varepsilon^{\gamma_2} D_2^\varepsilon\,. \tag{6.1}$$

The above equality is understood as an equality of differential operators. γ_1 and γ_2 are strictly positive reals, and D_1^ε and D_2^ε are differential operators whose coefficients are continuous, smooth outside the origin, and locally uniformly bounded when $\varepsilon \to 0$, i.e., there exists $\varepsilon_0 > 0$ such that for all compact subset K of $\mathbb{R}^n$, each component of these differential operators is bounded for $(\varepsilon,t,x) \in (0,\varepsilon_0]\times\mathcal{T}\times K$.

This kind of convergence carries with it two important properties.

PROPOSITION 6.2. *Suppose that a vector field F^ε DO-converges, as $\varepsilon \longrightarrow 0$, to a vector field F^0 on a time interval $\mathcal{T}$. Then we have the following.*

1. *The trajectories of $\dot{x} = f(\varepsilon,t,x)$ converge uniformly to those of $\dot{x} = f^0(t,x)$ on finite time intervals. More precisely, let $[0,T] \subset \mathcal{T}$, and let x^0 be the (unique) solution of*

$$\begin{aligned}\dot{x} &= f^0(t,x),\\ x(0) &= x_0\,.\end{aligned} \tag{6.2}$$

Then, for ε small enough, the unique solution x^ε of

$$\begin{aligned}\dot{x} &= f(\varepsilon,t,x),\\ x(0) &= x_0\end{aligned} \tag{6.3}$$

is defined on $[0,T]$, and $x^\varepsilon(t)$ converges to $x^0(t)$ uniformly on $[0,T]$.

2. *If $\mathcal{T} = [0,+\infty)$, and if all vector fields in (6.1) are homogeneous of degree zero and f^0 is autonomous, then, if the origin of*

$$\dot{x} = f^0(x) \tag{6.4}$$

is asymptotically stable, the origin of

$$\dot{x} = f(\varepsilon,t,x) \tag{6.5}$$

is (exponentially) asymptotically stable too for $\varepsilon > 0$ sufficiently small.

Proof. We prove (1). First, we rewrite (6.1) as

$$F^\varepsilon(I - \varepsilon^{\gamma_1} D_1^\varepsilon) = F^0 - \varepsilon^{\gamma_1} D_1^\varepsilon \frac{\partial}{\partial t} \; + \; \varepsilon^{\gamma_2} D_2^\varepsilon \, .$$

This is an equality between differential operators. We apply each side to the coordinate functions x_i. $D_1^\varepsilon x_i$ and $D_2^\varepsilon x_i$ are simply the coefficients in front of $\frac{\partial}{\partial x_i}$ in the expression of the differential operator D_1^ε or D_2^ε. This implies (coordinate by coordinate) that the differential equation (6.3) can be rewritten

$$\frac{\mathrm{d}}{\mathrm{d}t}\left(x \, - \, \varepsilon^{\gamma_1} d_1(\varepsilon, t, x)\right) \;\; = \;\; f^0(t,x) \; + \; \varepsilon^{\gamma_2} d_2(\varepsilon, t, x), \tag{6.6}$$

where $d_i(\varepsilon, t, x)$ $(i \in \{1, 2\})$ is the vector whose jth component is the coefficient of $\frac{\partial}{\partial x_j}$ in D_i^ε. This implies that the difference between $x^0(t) - x^\varepsilon(t)$ satisfies

$$\begin{aligned} \|x^\varepsilon(t) - x^0(t)\| \leq \varepsilon^{\gamma_1} \|d_1(\varepsilon, t, x^0(t))\| \; + \; \varepsilon^{\gamma_1} \|d_1(\varepsilon, t, x_0)\| \\ + \; \int_0^t \|f^0(\tau, x^\varepsilon(\tau)) - f^0(\tau, x^0(\tau))\| \mathrm{d}\tau \; + \; \varepsilon^{\gamma_2} \int_0^t \|d_2(\varepsilon, \tau, x^\varepsilon(\tau))\| \mathrm{d}\tau. \end{aligned}$$

The standard Gronwall lemma then yields, $\forall\, \varepsilon \in (0, \varepsilon_0]$ and $\forall\, t \in [0, T]$ such that x^ε remains in the interior of a certain compact neighborhood K of the trajectory x^0, the estimate $\|x^\varepsilon(t) - x^0(t)\| \leq (2\varepsilon^{\gamma_1} + T\varepsilon^{\gamma_2}) M e^{\lambda t}$, where λ is a Lipschitz constant (with respect to x) of F on $[0, T] \times K$ and M is an upperbound on $(0, \varepsilon_0] \times [0, T] \times K$ for both $\|d_1\|$ and $\|d_2\|$. This proves (1).

Let us prove (2). Since the right-hand side of (6.4) is homogeneous of degree zero, there exists [17] a *homogeneous* and autonomous Lyapunov function V, positive definite, whose derivative along (6.4) is given by

$$\dot{V}_{(6.4)} \; = \; F^0 V \; = \; -W. \tag{6.7}$$

Here $X\,V$, for X a vector field, denotes the Lie derivative of V along X with W homogeneous positive definite of the same degree as V, i.e.,

$$W(x) \; \geq \; c\, V(x). \tag{6.8}$$

Let us now compute the derivative of V along system (6.5). From (6.1) and (6.7),

$$\dot{V}_{(6.5)} \; = \; F^\varepsilon V \;\; = \;\; -\, W \; + \; \varepsilon^{\gamma_1} F^\varepsilon D_1^\varepsilon V \; - \; \varepsilon^{\gamma_1} D_1^\varepsilon \frac{\partial V}{\partial t} \; + \; \varepsilon^{\gamma_2} D_2^\varepsilon V,$$

which can be rewritten, since V is autonomous, as

$$F^\varepsilon V_\varepsilon \;\; = \;\; -\, W \; + \; \varepsilon^{\gamma_2} D_2^\varepsilon V, \tag{6.9}$$

with

$$V_\varepsilon \;\; = \;\; V \; - \; \varepsilon^{\gamma_1} D_1^\varepsilon V. \tag{6.10}$$

Since, by assumption, the operators D_1^ε and D_2^ε are homogeneous of degree zero, and locally uniformly bounded with respect to $\varepsilon > 0$, one has, since V is positive definite,

$$|D_1^\varepsilon V| \; \leq \; k\, V, \quad |D_2^\varepsilon V| \; \leq \; k\, V$$

$\forall\ \varepsilon > 0$. Hence, for ε sufficiently small, V_ε is arbitrarily close to V and hence positive definite, and also

$$\dot{V}_\varepsilon \;=\; F^\varepsilon V_\varepsilon \;\leq\; -\frac{c}{2} V. \tag{6.11}$$

Therefore, for ε small enough, V_ε is a strict Lyapunov function for system (6.5). This ends the proof of 2 via Lyapunov's first method. □

Before stating our convergence result, we recall two definitions introduced in [25, 9].

DEFINITION 6.3 (see [25, 9]). *Let Ω be a finite subset of $\mathbb{R}$ and $|\Omega|$ denote the number of elements of Ω. The set Ω is said to be MC if and only if*

(i) $\sum_{\omega\in\Omega} \omega = 0$;

(ii) *this is the only zero sum with at most $|\Omega|$ terms taken in Ω with possible repetitions:*

$$\left.\begin{array}{c} \displaystyle\sum_{\omega\in\Omega} \lambda_\omega \omega = 0 \\ (\lambda_\omega)_{\omega\in\Omega} \in \mathbb{Z}^{|\Omega|} \\ \displaystyle\sum_{\omega\in\Omega} |\lambda_\omega| \leq |\Omega| \end{array}\right\} \Longrightarrow \left\{\begin{array}{lll} (\lambda_\omega)_{\omega\in\Omega} & = & (0,\dots,0), \\ & or & (1,\dots,1), \\ & or & (-1,\dots,-1). \end{array}\right. \tag{6.12}$$

For example, a set $\{\omega_1, \omega_2\}$ is MC if and only if $\omega_2 = -\omega_1$ with $\omega_1 \neq 0$, a set $\{\omega_1, \omega_2, \omega_3\}$ is MC if and only if $\omega_3 = -\omega_1 - \omega_2$ with $\omega_1 \neq 0$, $\omega_2 \neq 0$, $\omega_1 + \omega_2 \neq 0$, $\omega_1 - \omega_2 \neq 0$, $\omega_1 + 2\omega_2 \neq 0$, $2\omega_1 + \omega_2 \neq 0$, $\omega_1 - 2\omega_2 \neq 0$, $2\omega_1 - \omega_2 \neq 0 \dots$.

DEFINITION 6.4 (see [25, 9]). *Let $(\Omega_\alpha)_{\alpha\in I}$ be a finite family of finite subsets Ω_α of $\mathbb{R}$. The family $(\Omega_\alpha)_{\alpha\in I}$ is said to be "independent with respect to p" if and only if*

$$\left.\begin{array}{l} \bullet\ \displaystyle\sum_{\alpha\in I}\sum_{\omega\in\Omega_\alpha} \lambda_\omega \omega = 0 \\ \bullet\ (\lambda_\omega)_{\omega\in\Omega_\alpha, \alpha\in I} \in \mathbb{Z}^{\Sigma|\Omega_\alpha|} \\ \bullet\ \displaystyle\sum_{\alpha\in I}\sum_{\omega\in\Omega} |\lambda_\omega| \leq p \end{array}\right\} \Longrightarrow \sum_{\omega\in\Omega_\alpha} \lambda_\omega \omega = 0 \quad \forall \alpha \in I. \tag{6.13}$$

For example, the sets $(\{\omega_1, \omega_2, \omega_3\}, \{\omega_4, \omega_5\})$ are both MC and independent with respect to 2 if and only if $\omega_3 = -\omega_1 - \omega_2$ and $\omega_5 = -\omega_4$ with $\omega_1 \neq 0$, $\omega_2 \neq 0$, $\omega_1 + \omega_2 \neq 0$, $\omega_1 - \omega_2 \neq 0$, $\omega_1 + 2\omega_2 \neq 0$, $2\omega_1 + \omega_2 \neq 0$, $\omega_1 - 2\omega_2 \neq 0$, $2\omega_1 - \omega_2 \neq 0$, $\omega_4 \neq 0$ (this is MC), and $\omega_1 + \omega_4 \neq 0$, $\omega_1 - \omega_4 \neq 0$, $\omega_2 + \omega_4 \neq 0$, $\omega_2 - \omega_4 \neq 0$, $\omega_1 + \omega_2 + \omega_4 \neq 0$, $\omega_1 + \omega_2 - \omega_4 \neq 0$ (this is independence).

We are now ready to state our convergence result.

THEOREM 6.5. *Let N be a positive integer and consider, for $j = 1, \dots, N$,*

- *some vector fields $X_j^s \in C^\infty(\mathbb{R}^n - \{0\}; \mathbb{R}^n) \cap C^0(\mathbb{R}^n; \mathbb{R}^n)$ $(s = 1, \dots, \ell(j))$,*
- *some smooth complex valued functions of time η_j^s $(s = 1, \dots, \ell(j))$ such that, for some M,*

$$\left|\eta_j^s(t)\right| \leq M \text{ and } \left|\dot{\eta}_j^s(t)\right| \leq M \quad \forall t \in \mathcal{T}, \tag{6.14}$$

- *some sets $\Omega_j = \{\omega_j^1, \dots, \omega_j^{\ell(j)}\}$ of real numbers such that $\omega_j^s = 0$ if $\ell(j) = 1$, Ω_j is MC if $\ell(j) \geq 2$, and the family (Ω_j) $(\ell(j) \geq 2)$ is independent with respect to $P \triangleq \mathrm{Max}_j \ell(j)$.*

Then the vector field

$$F^\varepsilon \;=\; \frac{\partial}{\partial t} + \sum_{j=1}^{N}\sum_{s=1}^{\ell(j)} \alpha^s_{j,\varepsilon} X^s_j\,, \tag{6.15}$$

with

$$\alpha^s_{j,\varepsilon}(t) = 2\varepsilon^{-\frac{\ell(j)-1}{\ell(j)}} \Re\left(\eta^s_j(t)\, e^{i\omega^s_j t/\varepsilon}\right), \tag{6.16}$$

DO-converges, as $\varepsilon \to 0$, to the vector field

$$F^0 = \frac{\partial}{\partial t} + \sum_{j=1}^{N} \frac{2}{\ell(j)} \Re\left(\frac{\eta^1_j \cdots \eta^{\ell(j)}_j}{i^{\ell(j)-1}}\right) B_j \tag{6.17}$$

$$\text{with}\quad B_j \;=\; \sum_{\sigma\in\mathfrak{S}(\ell(j))} \frac{[X^{\sigma(1)}_j, [X^{\sigma(2)}_j, [\ldots, X^{\sigma(\ell(j))}_j]\ldots]]}{\omega^{\sigma(1)}_j(\omega^{\sigma(1)}_j + \omega^{\sigma(2)}_j)\ldots(\omega^{\sigma(1)}_j + \ldots + \omega^{\sigma(\ell(j)-1)}_j)}\,.$$

Furthermore, if all the vector fields X^s_j are homogeneous of degree zero, then all the differential operators in (6.1) *are homogeneous of degree zero also.*

Remark 6.6. This result is very much related to the theory of "normal forms" for time-varying differential equations, as shown, for instance, in [19, Chapter 6]. Let us recall (see [19] for details) that a vector field $\frac{\partial}{\partial t} + \varepsilon f^0(t,x)\frac{\partial}{\partial x}$ is said to be in normal form if and only if $[\frac{\partial}{\partial t}, f^0\frac{\partial}{\partial x}] = 0$, i.e., if f^0 does not depend on t. For a system

$$(\Sigma_\varepsilon) \qquad \dot{x} \;=\; f(\varepsilon, t, x),$$

finding a normal form means finding a change of coordinates $x \mapsto y = x + \alpha(\varepsilon, x)$ that transforms (Σ_ε) into

$$(\Sigma_0) \qquad \dot{y} \;=\; \varepsilon f_0(y).$$

In general, deciding whether a normal form exists for a system, and then possibly finding this normal form, is a difficult problem and there are no systematic tools available.

Let us, however, rephrase Theorem 6.5 in the terms of [19]. By a time-scaling $t \mapsto \varepsilon t$, the system $\dot{x} = f(\varepsilon, x, t)$, where f is defined by $F^\varepsilon = \frac{\partial}{\partial t} + f\frac{\partial}{\partial x}$, with F^ε given by (6.15), is rewritten as

$$(\Sigma'_\varepsilon) \qquad \dot{x} \;=\; \varepsilon f_1(t,x) \,+\, \varepsilon^{1/2} f_2(t,x) \,+\, \cdots \,+\, \varepsilon^{1/P} f_P(t,x).$$

In the context of "normal forms," Theorem 6.5 states that (Σ_0), with f^0 defined by $F^0 = \frac{\partial}{\partial t} + f^0\frac{\partial}{\partial x}$ and F^0 given by (6.17), is a normal form for (Σ'), up to terms of higher order in ε.

7. Proofs.

7.1. Proof of Proposition 4.2 (section 4). Point (b) is strictly a consequence of the construction. Point (c) follows from the fact that if a linear combination of all the vector fields $\tilde{b}_j$ with constant real coefficients is identically zero, then homogeneity implies that each linear combination where only the terms corresponding to the brackets of same length must also be zero, and since by construction all the brackets

$\tilde{b}_j$ of same length are linearly independent over $\mathbb{R}$, this implies that all the coefficients are zero.

Let us prove point (a). Recall that any Lie bracket of length $p > P$ made with the vector fields b_i is identically zero (see Remark 4.1). From this fact, the controllability assumption (3.3), and the construction itself, there clearly exist integers $j_1, \ldots, j_n \in \{1, \ldots, N\}$ such that $\{\tilde{b}_{j_1}(0), \ldots, \tilde{b}_{j_n}(0)\}$ is a basis of $\mathbb{R}^n$. Hence $\{\tilde{b}_{j_1}(x), \ldots, \tilde{b}_{j_n}(x)\}$ is a basis of $\mathbb{R}^n$ for x in some neighborhood W of the origin. Let us show that this is true for any x in $\mathbb{R}^n$. Let x be outside W. There exist $\lambda > 0$ such that $\bar{x} = \delta_\lambda(x)$ is in W and hence $\{\tilde{b}_{j_1}(\bar{x}), \ldots, \tilde{b}_{j_n}(\bar{x})\}$ is a basis of $\mathbb{R}^n$. This implies, since δ_λ is a local diffeomorphism from a neighborhood of x to a neighborhood of $\bar{x}$, that

$$\left\{ \left((\delta_\lambda^{-1})_* \tilde{b}_{j_1} \right)(x), \ldots, \left((\delta_\lambda^{-1})_* \tilde{b}_{j_n} \right)(x) \right\}$$

is also a basis of $\mathbb{R}^n$. Now, from the homogeneity, $(\delta_\lambda^{-1})_* \tilde{b}_{j_k} = \lambda^{-\ell(j_k)} \tilde{b}_{j_k}$. This proves point (a). □

7.2. Proof of Theorem 6.5. In [5, 8, 25, 9], the main ingredient of the proof was iterated integrations by parts. Here we mimic these integrations by parts but at the level of products of differential operators instead of integrals along the solutions. The closed-loop vector field F^ε can be rewritten as

$$\begin{aligned} F^\varepsilon = & \frac{\partial}{\partial t} + \sum_{\substack{1 \le j \le N \\ \ell(j)=1}} 2\eta_j^1 X_j^1 \\ & + \sum_{\substack{1 \le j \le N \\ \ell(j) \ge 2}} \sum_{s=1}^{\ell(j)} \varepsilon^{-\frac{\ell(j)-1}{\ell(j)}} \left(\eta_j^s e^{i\omega_j^s t/\varepsilon} + \overline{\eta_j^s} e^{-i\omega_j^s t/\varepsilon} \right) X_j^s. \end{aligned} \tag{7.1}$$

Let us make some conventions and definitions, used only in the present proof. We define the sets of indices

$$J = \{j \in \{1, \ldots, N\}, \ell(j) \ge 2\} = \{m_1 + 1, \ldots, N\}, \tag{7.2}$$

$$J_l = \{j \in \{1, \ldots, N\}, \ell(j) = l\} = \{m_{l-1} + 1, \ldots, m_l\}, \tag{7.3}$$

$$K_j = \{-\ell(j), -\ell(j) - 1, \ldots, -1, 1, 2, \ldots \ell(j)\} \tag{7.4}$$

and the sets of pairs of indices

$$I = \{(j, s), j \in J, s \in K_j\} = \bigcup_{j \in J} \{j\} \times K_j, \tag{7.5}$$

$$I_l = \{(j, s) \in I, \ell(j) = l\} = \bigcup_{j \in J_l} \{j\} \times K_j. \tag{7.6}$$

We call F_1 the vector field

$$F_1 = \sum_{\substack{1 \le j \le N \\ \ell(j)=1}} 2\eta_j^1 X_j^1 = \sum_{(j,s) \in I_1} 2\eta_j^s X_j^s. \tag{7.7}$$

Clearly, if we define, for $s < 0$, the real numbers ω_j^s, the complex numbers η_j^s, and the vector fields X_j^s by

$$\left.\begin{aligned} \omega_j^{-s} &= -\omega_j^s \\ \eta_j^{-s} &= \overline{\eta_j^s} \\ X_j^{-s} &= X_j^s \end{aligned}\right\} \text{ for } j \in J\,,\ s \in K_j\,,\ s > 0, \tag{7.8}$$

the vector field F^ε from (7.1) may be rewritten as

$$F^\varepsilon = \frac{\partial}{\partial t} \;+\; F_1 \;+\; \sum_{(j,s)\in I} \varepsilon^{-\frac{\ell(j)-1}{\ell(j)}}\, \eta_j^s e^{i\omega_j^s t/\varepsilon} X_j^s \tag{7.9}$$

$$= \frac{\partial}{\partial t} \;+\; F_1 \;+\; \varepsilon^{-\frac{1}{2}}\, F_2^\varepsilon \;+\; \varepsilon^{-\frac{2}{3}}\, F_3^\varepsilon \;+\; \cdots \;+\; \varepsilon^{-\frac{P-1}{P}}\, F_P^\varepsilon, \tag{7.10}$$

where

$$F_l^\varepsilon \;=\; \sum_{(j,s)\in I_l} \eta_j^s e^{i\omega_j^s t/\varepsilon} X_j^s. \tag{7.11}$$

Note that the interest of (7.10) is that the negative powers of ε are written apart, and the vector fields F_j^ε have the "boundedness" property that their coefficients are continuous functions of x and t, smooth outside $x = 0$, indexed by $\varepsilon > 0$, and locally uniformly bounded with respect to $\varepsilon > 0$. (It is not the case of F^ε itself because of the negative powers of ε.) In the remainder of the proof, we shall always write the negative powers of ε apart so that all the differential operators written as capital letters never contain coefficients that are unbounded when ε goes to zero.

We now define a certain number of differential operators $F^\varepsilon_{p_1,p_2,\ldots,p_d}$ of order d for d between 1 and P, and for all d-tuple $(p_1, p_2, \ldots, p_d)$ of integers such that

$$\left\{\begin{array}{rcl} 1 \le p_k \le P \;\; \text{for } 1 \le k \le d, \\ \frac{1}{p_1} + \cdots + \frac{1}{p_{d-1}} \le 1, \\ (p_1, p_2) \;\neq\; (2,2), \\ (p_1, p_2, p_3) \;\neq\; (3,3,3), \\ \vdots \\ (p_1, \ldots, p_{d-1}) \;\neq\; (d-1, \ldots, d-1). \end{array}\right. \tag{7.12}$$

We define $F^\varepsilon_{p_1,p_2,\ldots,p_d}$ to be equal to

$$\sum_{((j_1,s_1),\ldots,(j_d,s_d))\,\in\, I^d(p_1,\ldots,p_d)} \frac{\eta_{j_1}^{s_1}\eta_{j_2}^{s_2}\cdots\eta_{j_d}^{s_d} e^{i(\omega_{j_1}^{s_1}+\cdots+\omega_{j_d}^{s_d})\frac{t}{\varepsilon}}\, X_{j_d}^{s_d} X_{j_{d-1}}^{s_{d-1}} \ldots X_{j_1}^{s_1}}{i^{(d-1)}\omega_{j_1}^{s_1}(\omega_{j_1}^{s_1}+\omega_{j_2}^{s_2})\cdots(\omega_{j_1}^{s_1}+\cdots+\omega_{j_{d-1}}^{s_{d-1}})}, \tag{7.13}$$

where $I^d(p_1, \ldots, p_d)$ is the set of d-tuples of indices $((j_1, s_1), \ldots, (j_d, s_d))$ such that $\ell(j_k) = p_k$, and which are neither a collection of $\frac{d}{2}$ pairs of the form (j, s), $(j, -s)$ nor such that, for some (even) k, $2 \le k \le d$, $((j_1, s_1), \ldots, (j_k, s_k))$ would be a collection of $\frac{k}{2}$ pairs of the form (j, s), $(j, -s)$. More precisely, $I^d(p_1, \ldots, p_d)$ may be defined recursively by $I^1(p) = I_1$ and

$$((j_1, s_1), \ldots, (j_d, s_d)) \;\in\; I^d(p_1, \ldots, p_d)$$

$$\Leftrightarrow \left\{\begin{array}{l} \bullet\; (j_k, s_k) \;\in\; I_{p_k} \;\forall\; k, \\ \bullet\; ((j_1, s_1), \ldots, (j_{d-1}, s_{d-1})) \;\in\; I^{d-1}(p_1, \ldots, p_{d-1}), \\ \bullet\; \text{there exists no permutation } \tau \in \mathfrak{S}(d) \\ \quad \text{such that } \;(j_{\tau(k)}, s_{\tau(k)}) \;=\; (j_k, -s_k). \end{array}\right. \tag{7.14}$$

With the above definition of the sets of indices $I^d(p_1,\dots,p_d)$, the denominators in (7.13) cannot be zero because of the following lemma.

LEMMA 7.1. *Let* $((j_1,s_1),\dots,(j_d,s_d)) \in I^d$ *(see the definition of* I *in* (7.5)*) be such that* $\omega_{j_1}^{s_1}+\dots+\omega_{j_d}^{s_d} = 0$. *Then*

- *either* $(\ell(j_1),\dots,\ell(j_d)) = (d,\dots,d)$ *and there exists a permutation* $\sigma\in\mathfrak{S}(d)$ *such that the p-tuple* $((j_1,s_1),\dots,(j_d,s_d))$ *is exactly equal to* $((j,\sigma(1)),\dots,(j,\sigma(p)))$ *or* $((j,-\sigma(1)),\dots,(j,-\sigma(p)))$ *(with* $j_1=\dots=j_d=j$*),*
- *or* $\frac{1}{\ell(j_1)}+\dots+\frac{1}{\ell(j_d)} > 1$,
- *or there exists a permutation* $\tau\in\mathfrak{S}(d)$ *such that* $(j_{\tau(k)},s_{\tau(k)}) = (j_k,-s_k)\ \forall\ k$.

Proof. The equality $\omega_{j_1}^{s_1}+\dots+\omega_{j_d}^{s_d}=0$ may be rewritten as

$$\sum_{\substack{j\in\{1,\dots,N\}\\ \ell(j)\geq 2}}\ \sum_{s=1}^{\ell(j)}\lambda_j^s\omega_j^s \ = \ 0, \tag{7.15}$$

where the integer λ_j^s is equal to the number of times that (j,s) appears in $((j_1,s_1),\dots,(j_d,s_d))$ minus the number of times $(j,-s)$ appears. Of course, (7.15) may be rewritten as

$$\sum_{j\in J}\sum_{\omega\in\Omega_j}\lambda_\omega\omega = 0$$

with $\lambda_{\omega_j^s}=\lambda_j^s$. Note that

$$\sum_{\omega}|\lambda_w| \ = \ \sum_{j,s}|\lambda_j^s| \ \leq \ d \ \leq \ P.$$

Hence, from the assumption that the sequences of frequencies are mutually "independent with respect to P" and are all MC (see (6.12)–(6.13)), each $(\lambda_j^1,\dots,\lambda_j^{\ell(j)})$ is equal to either $(0,\dots,0)$, $(1,\dots,1)$, or $(-1,\dots,-1)$. If it is different from $(0,\dots,0)$ for at least one j, then all the couples $(j,1),\dots,(j,\ell(j))$ or all the couples $(j,-1),\dots,(j,-\ell(j))$ appear in $((j_1,s_1),\dots,(j_d,s_d))$. If $d=\ell(j)$ for this j, i.e., if $((j_1,s_1),\dots,(j_d,s_d))$ is a reordering of $((j,1),\dots,(j,\ell(j)))$, or of $((j,-1),\dots,(j,-\ell(j)))$, then we are in the first case of the lemma; if $d>\ell(j)$, then there is at least another couple (j',s') in $((j_1,s_1),\dots,(j_d,s_d))$ and hence the sum $\frac{1}{\ell(j_1)}+\dots+\frac{1}{\ell(j_d)}$ can be no less than $1+\frac{1}{\ell(j')}$ and we are in the second case of the lemma. Let us now examine the case where all the $(\lambda_j^1,\dots,\lambda_j^{\ell(s)})$'s are equal to $(0,\dots,0)$. This means that for all j,s, the couple (j,s) and the couple $(j,-s)$ appear the same number of times in $((j_1,s_1),\dots,(j_d,s_d))$. This allows one to build the permutation having the property required in the third point of the lemma: it is the one that exchanges 1 with the first k_1 such that $(j_{k_1},s_{k_1})=(j_1,-s_1)$, 2 (3 if $k_1=2$) with the first $k_2\neq k_1$ such that $(j_{k_2},s_{k_2})=(j_2,-s_2)$, and so on. □

We shall now prove the following two facts.

Fact 1. For all q, $1\leq q\leq P$, there exist $\gamma_{1,q}$ and $\gamma_{2,q}$ strictly positive such that

$$F^\varepsilon = \frac{\partial}{\partial t} \ + \ F_1 \ + \ \sum_{p=2}^{q}(-1)^{p-1}F^\varepsilon_{\underbrace{p,p,\dots,p}_{p\text{ times}}} + \varepsilon^{\gamma_{1,q}}\left(F^\varepsilon D^\varepsilon_{1,q} \ - \ D^\varepsilon_{1,q}\frac{\partial}{\partial t}\right) \ + \ \varepsilon^{\gamma_{2,q}}\,D^\varepsilon_{2,q} \tag{7.16}$$

$$+ \sum_{\substack{(p_1,\dots,p_q)\in\{2,\dots,P\}^q,\\ \frac{1}{p_1}+\dots+\frac{1}{p_q}\le 1,\\ (p_1,\dots,p_q)\neq(q,\dots,q)}} (-1)^{q-1}\varepsilon^{-\left(1-\frac{1}{p_1}-\dots-\frac{1}{p_q}\right)} F^{\varepsilon}_{p_1,\dots,p_q}$$

Fact 2. For all p, $1 \le p \le P$, there exist $\gamma'_{1,p}$ and $\gamma'_{2,p}$ strictly positive such that

$$F^{\varepsilon}_{\underbrace{p,p,\dots,p}_{p \text{ times}}} = 2\,\frac{(-1)^{p-1}}{p} \sum_{j\in J_p} \Re\left(\frac{\eta_j^1\cdots\eta_j^p}{i^{(p-1)}}\right) B_j + \varepsilon^{\gamma'_{1,p}}\left(F^{\varepsilon} D'^{\varepsilon}_{1,p} - D'^{\varepsilon}_{1,p}\frac{\partial}{\partial t}\right) + \varepsilon^{\gamma'_{2,p}} D'^{\varepsilon}_{2,p} \tag{7.17}$$

with

$$B_j = \sum_{\sigma\in\mathfrak{S}(\ell(j))} \frac{[X_j^{\sigma(1)},[X_j^{\sigma(2)},[\dots,X_j^{\sigma(\ell(j))}]\dots]]}{\omega_j^{\sigma(1)}(\omega_j^{\sigma(1)}+\omega_j^{\sigma(2)})\cdots(\omega_j^{\sigma(1)}+\cdots+\omega_j^{\sigma(\ell(j)-1)})}. \tag{7.18}$$

These two facts imply Theorem 6.5. Indeed, for $q = P$, the last sum in (7.16) is empty since $\frac{1}{p_1}+\cdots+\frac{1}{p_P} \le 1$ with all the integers p_j no larger that P implies $(p_1,\dots,p_P) = (P,\dots,P)$. Hence for $q = P$, (7.16) reads

$$F^{\varepsilon} = \frac{\partial}{\partial t} + F_1 + \sum_{p=2}^{P}(-1)^{p-1} F^{\varepsilon}_{\underbrace{p,p,\dots,p}_{p \text{ times}}} + \varepsilon^{\gamma_{1,P}}\left(F^{\varepsilon} D^{\varepsilon}_{1,P} - D^{\varepsilon}_{1}\frac{\partial}{\partial t}\right) + \varepsilon^{\gamma_{2,P}} D^{\varepsilon}_{2,P}. \tag{7.19}$$

Substituting in the above the expression of $F^{\varepsilon}_{p,\dots,p}$ given by (7.17), one clearly gets (6.1) with the appropriate differential operators D^{ε}_1 and D^{ε}_2 and the appropriate positive real numbers γ_1 and γ_2.

Proof of Fact 1. We prove (7.16) by induction on q, from $q = 1$ to $q = P$.

For $q = 1$, the sum on the first line of (7.16) is empty, one may take $D^{\varepsilon}_{1,1}, D^{\varepsilon}_1$, and $D^{\varepsilon}_{2,1}$ to be zero, and (7.16) is simply (7.10).

Let us now suppose that (7.16) holds for a certain $q \ge 1$ and let us prove it for $q+1$. This is done through a manipulation on differential operators that more or less mimics an integration by parts. Since we shall use it elsewhere, let us explain it on a "general" differential operator Y before applying it.

Consider a differential operator of order d on functions of t and x that does not contain derivations with respect to t:

$$Y = \sum_{\text{multi-indices } I \text{ of length } d} \eta_I(t) a_I(t,x) \frac{\partial^{|I|}}{\partial x_I}. \tag{7.20}$$

Define $Y^{[-1]}$ and $Y^{[1]}$ to be

$$Y^{[-1]} = \sum_{\text{multi-indices } I \text{ of length } d} \eta_I(t)\left(\int_*^t a_I(\tau,x)\mathrm{d}\tau\right)\frac{\partial^{|I|}}{\partial x_I}, \tag{7.21}$$

$$Y^{[1]} = \sum_{\text{multi-indices } I \text{ of length } d} \frac{\mathrm{d}\eta_I}{\mathrm{d}t}(t)\left(\int_*^t a_I(\tau,x)\mathrm{d}\tau\right)\frac{\partial^{|I|}}{\partial x_I}. \tag{7.22}$$

Note that these are defined up to a function of x (through the initial time in the integrals) and that $Y^{[1]}$ is zero if the η's are constants. The derivative with respect to t of $Y^{[-1]}$ is $Y + Y^{[1]}$ in the following sense:

$$Y + Y^{[1]} \;=\; \left[\, \frac{\partial}{\partial t}\,,\, Y^{[-1]} \,\right] \;=\; \frac{\partial}{\partial t}\, Y^{[-1]} \;-\; Y^{[-1]}\, \frac{\partial}{\partial t}. \tag{7.23}$$

Indeed it is obvious that for any smooth function h of x and t, one has

$$Y.h\,(t,x) \;+\; Y^{[1]}.h\,(t,x) \;=\; \frac{\partial}{\partial t}\,\Big(Y^{[-1]}.h\,(t,x)\Big) \;-\; Y^{[-1]}.\frac{\partial h}{\partial t}\,(t,x), \tag{7.24}$$

simply because $\frac{\partial}{\partial t}$ commutes with $\frac{\partial}{\partial^{|I|}x_I}$. Then we rewrite (7.23) in the following way:

$$\begin{aligned} Y &= \left[\, \frac{\partial}{\partial t}\,,\, Y^{[-1]} \,\right] \;-\; Y^{[1]} \\ &= F^{\varepsilon}\, Y^{[-1]} \;-\; \left(\sum_{r=1}^{P} \varepsilon^{-\frac{r-1}{r}} F_r^{\varepsilon}\right) Y^{[-1]} \;-\; Y^{[-1]}\, \frac{\partial}{\partial t} \;-\; Y^{[1]}. \end{aligned} \tag{7.25}$$

In order to prove that if (7.16) holds for q, it also holds for $q+1$, we apply the identity (7.25) with

$$\begin{aligned} Y &= F^{\varepsilon}_{p_1,\dots,p_q}, \\ Y^{[-1]} &= \varepsilon\, G^{\varepsilon}_{p_1,\dots,p_q}, \\ Y^{[1]} &= \varepsilon\, H^{\varepsilon}_{p_1,\dots,p_q}, \end{aligned}$$

for

$$(p_1,\dots,p_q) \neq (q,\dots,q) \;\text{ and }\; \frac{1}{p_1} + \dots + \frac{1}{p_q} \leq 1, \tag{7.26}$$

where $G^{\varepsilon}_{p_1,\dots,p_q}$ and $H^{\varepsilon}_{p_1,\dots,p_q}$ are given by

$$G^{\varepsilon}_{p_1,\dots,p_q} = \sum_{((j_1,s_1),\dots,(j_q,s_q)) \,\in\, I^q(p_1,\dots,p_q)} \frac{\eta_{j_1}^{s_1} \cdots \eta_{j_q}^{s_q} e^{i(\omega_{j_1}^{s_1} + \dots + \omega_{j_q}^{s_q})t/\varepsilon}\, X_{j_q}^{s_q} X_{j_{q-1}}^{s_{q-1}} \dots X_{j_1}^{s_1}}{i^q \omega_{j_1}^{s_1} (\omega_{j_1}^{s_1} + \omega_{j_2}^{s_2}) \dots (\omega_{j_1}^{s_1} + \dots + \omega_{j_q}^{s_q})} \tag{7.27}$$

$$H^{\varepsilon}_{p_1,\dots,p_q} = \sum_{((j_1,s_1),\dots,(j_q,s_q)) \in I^q(p_1,\dots,p_q)} \frac{\left(\frac{d}{dt}\left(\eta_{j_1}^{s_1} \cdots \eta_{j_q}^{s_q}\right)\right) e^{i(\omega_{j_1}^{s_1} + \dots + \omega_{j_q}^{s_q})t/\varepsilon}\, X_{j_q}^{s_q} X_{j_{q-1}}^{s_{q-1}} \dots X_{j_1}^{s_1}}{i^q \omega_{j_1}^{s_1} (\omega_{j_1}^{s_1} + \omega_{j_2}^{s_2}) \dots (\omega_{j_1}^{s_1} + \dots + \omega_{j_q}^{s_q})}. \tag{7.28}$$

Note that the denominators are nonzero because, from Lemma 7.1, the definition (7.14) of the set of indices $I^q(p_1,\dots,p_q)$ precisely removes the terms where the denominators would be zero.

Then (7.25) with the above expressions for Y, $Y^{[1]}$, and $Y^{[-1]}$ yields

$$F^{\varepsilon}_{p_1,\dots,p_q} \;=\; -\sum_{r=1}^{P} \varepsilon^{\frac{1}{r}} F_r^{\varepsilon} G^{\varepsilon}_{p_1,\dots,p_q} + \varepsilon\, F^{\varepsilon} G^{\varepsilon}_{p_1,\dots,p_q} - \varepsilon\, G^{\varepsilon}_{p_1,\dots,p_q} \frac{\partial}{\partial t} \;-\; \varepsilon\, H^{\varepsilon}_{p_1,\dots,p_q}. \tag{7.29}$$

From (7.27) and (7.11) we have

$$\begin{aligned} &F_r^{\varepsilon} G^{\varepsilon}_{p_1,\dots,p_q} \\ &\quad= \sum_{\substack{((j_1,s_1),\dots,(j_q,s_q)) \;\in\; I^q(p_1,\dots,p_q) \\ (j_{q+1},s_{q+1}) \,\in\, I_r}} \frac{\eta_{j_1}^{s_1} \cdots \eta_{j_{q+1}}^{s_{q+1}} e^{i(\omega_{j_1}^{s_1} + \dots + \omega_{j_{q+1}}^{s_{q+1}})t/\varepsilon}\, X_{j_{q+1}}^{s_{q+1}} X_{j_q}^{s_q} \dots X_{j_1}^{s_1}}{i^q \omega_{j_1}^{s_1} (\omega_{j_1}^{s_1} + \omega_{j_2}^{s_2}) \dots (\omega_{j_1}^{s_1} + \dots + \omega_{j_q}^{s_q})}. \end{aligned}$$

From (7.14), the $(q+1)$-tuples $((j_1,s_1),\dots,(j_{q+1},s_{q+1}))$, which are in $I^q(p_1,\dots,p_q)\times I_r$ but not in $I^{q+1}(p_1,\dots,p_q,r)$, are such that there exists a permutation τ of the set of integers $\{1,\dots,q+1\}$ for which

$$\begin{aligned}((j_{\tau(1)},s_{\tau(1)}),(j_{\tau(2)},s_{\tau(2)}),\dots,(j_{\tau(q+1)},s_{\tau(q+1)}))\\ =((j_1,-s_1),(j_2,-s_2),\dots,(j_{q+1},-s_{q+1})),\end{aligned}\tag{7.30}$$

and this is possible only if q is odd. Equations (7.8) (for X and ω, not for η) and (7.30) imply that the term corresponding to $((j_1,-s_1),(j_2,-s_2),\dots,(j_{q+1},-s_{q+1}))$ is equal to

$$\frac{\eta_{j_{\tau(1)}}^{s_{\tau(1)}}\cdots\eta_{j_{\tau(q+1)}}^{s_{\tau(q+1)}}e^{i(\omega_{j_{\tau(1)}}^{s_{\tau(1)}}+\cdots+\omega_{j_{\tau(q+1)}}^{s_{\tau(q+1)}})t/\varepsilon}\,X_{j_{q+1}}^{s_{q+1}}X_{j_q}^{s_q}\cdots X_{j_1}^{s_1}}{i^q(-\omega_{j_1}^{s_1})(-\omega_{j_1}^{s_1}-\omega_{j_2}^{s_2})\dots(-\omega_{j_1}^{s_1}-\cdots-\omega_{j_q}^{s_q})}$$

which, since q must be odd (if not, there is no such term), is equal to

$$-\frac{\left(\prod_{k=1}^{q+1}\eta_{j_{\tau(k)}}^{s_{\tau(k)}}\right)e^{i(\omega_{j_1}^{s_1}+\cdots+\omega_{j_{q+1}}^{s_{q+1}})t/\varepsilon}\,X_{j_{q+1}}^{s_{q+1}}X_{j_q}^{s_q}\cdots X_{j_1}^{s_1}}{i^q\omega_{j_1}^{s_1}(\omega_{j_1}^{s_1}+\omega_{j_2}^{s_2})\cdots(\omega_{j_1}^{s_1}+\cdots+\omega_{j_q}^{s_q})}$$

and τ gives the change of index in the product allowing us to say that this is the opposite of the term corresponding to $((j_1,s_1),(j_2,s_2),\dots,(j_{q+1},s_{q+1}))$. Hence these terms sum to zero in the above sum. From (7.14), this implies $F_r^\varepsilon G_{p_1,\dots,p_q}^\varepsilon = F_{p_1,\dots,p_q,r}^\varepsilon$. Substituting this in (7.29) yields (we rename r as p_{q+1})

$$F_{p_1,\dots,p_q}^\varepsilon = -\sum_{p_{q+1}=1}^{P}\varepsilon^{\frac{1}{p_{q+1}}}F_{p_1,\dots,p_q,p_{q+1}}^\varepsilon+\varepsilon\,F^\varepsilon G_{p_1,\dots,p_q}^\varepsilon-\varepsilon\,G_{p_1,\dots,p_q}^\varepsilon\frac{\partial}{\partial t}-\varepsilon\,H_{p_1,\dots,p_q}^\varepsilon.$$

Hence (7.16) yields

$$\begin{aligned}F^\varepsilon = \frac{\partial}{\partial t}+F_1+\sum_{p=2}^{q}(-1)^{p-1}F^\varepsilon_{\underbrace{p,p,\dots,p}_{p\text{ times}}}\\ +\varepsilon^{\gamma_{1,q}}\left(F^\varepsilon D_{1,q}^\varepsilon-D_1^\varepsilon\frac{\partial}{\partial t}\right)+\varepsilon^{\gamma_{2,q}}D_{2,q}^\varepsilon\\ +(-1)^q\sum_{\substack{(p_1,\dots,p_q)\in\{2,\dots,P\}^q\\ \frac{1}{p_1}+\cdots+\frac{1}{p_q}\le 1\\ (p_1,\dots,p_q)\neq(q,\dots,q)\\ p_{q+1}\in\{1,\dots,P\}}}\varepsilon^{-\left(1-\frac{1}{p_1}-\cdots-\frac{1}{p_{q+1}}\right)}F^\varepsilon_{p_1,\dots,p_q,p_{q+1}}\\ +(-1)^{q-1}\sum_{\substack{(p_1,\dots,p_q)\in\{2,\dots,P\}^q,\\ \frac{1}{p_1}+\cdots+\frac{1}{p_q}\le 1\\ (p_1,\dots,p_q)\neq(q,\dots,q)}}\varepsilon^{\frac{1}{p_1}+\cdots+\frac{1}{p_q}}\left(F^\varepsilon G^\varepsilon_{p_1,\dots,p_q}-G^\varepsilon_{p_1,\dots,p_q}\frac{\partial}{\partial t}-H^\varepsilon_{p_1,\dots,p_q}\right).\end{aligned}\tag{7.31}$$

The term corresponding to $(p_1,\dots,p_{q+1})=(q+1,\dots,q+1)$ in the sum on the third line is $(-1)^qF^\varepsilon_{q+1,\dots,q+1}$, it adds to the sum on the first line and this yields the first line of (7.16) for $q+1$. The other terms in this sum such that $\frac{1}{p_1}+\cdots+\frac{1}{p_{q+1}}\le 1$ yield exactly the third line of (7.16) for $q+1$, and the terms in this sum such that

$\frac{1}{p_1} + \cdots + \frac{1}{p_{q+1}} > 1$, as well as all the last sum, add up with the second line to give the second line (the "small" terms) of (7.16) for $q+1$. This proves (7.16) for $q+1$ and ends the proof by induction of Fact 1.

Proof of Fact 2. From the definition (7.13) of $F^\varepsilon_{p_1,p_2,\ldots,p_d}$, we have

(7.32)
$$F^\varepsilon_{\underbrace{p,\ldots,p}_{p \text{ times}}} = \sum_{\substack{((j_1,s_1),\ldots,(j_p,s_p)) \in I^p(p,\ldots,p) \\ \omega_{j_1}^{s_1}+\cdots+\omega_{j_p}^{s_p}=0}} \frac{\eta_{j_1}^{s_1}\eta_{j_2}^{s_2}\cdots\eta_{j_p}^{s_p} X_{j_p}^{s_p} X_{j_{p-1}}^{s_{p-1}} \ldots X_{j_1}^{s_1}}{i^{(p-1)}\omega_{j_1}^{s_1}(\omega_{j_1}^{s_1}+\omega_{j_2}^{s_2})\ldots(\omega_{j_1}^{s_1}+\cdots+\omega_{j_{p-1}}^{s_{p-1}})}$$
$$+ \sum_{\substack{((j_1,s_1),\ldots,(j_p,s_p)) \in I^p(p,\ldots,p) \\ \omega_{j_1}^{s_1}+\cdots+\omega_{j_p}^{s_p}\neq 0}} \frac{\eta_{j_1}^{s_1}\eta_{j_2}^{s_2}\cdots\eta_{j_p}^{s_p} e^{i(\omega_{j_1}^{s_1}+\cdots+\omega_{j_p}^{s_p})t/\varepsilon} X_{j_p}^{s_p} X_{j_{p-1}}^{s_{p-1}} \ldots X_{j_1}^{s_1}}{i^{(p-1)}\omega_{j_1}^{s_1}(\omega_{j_1}^{s_1}+\omega_{j_2}^{s_2})\ldots(\omega_{j_1}^{s_1}+\cdots+\omega_{j_{p-1}}^{s_{p-1}})}.$$

Now, apply (7.20), (7.21), (7.22), and (7.25) with Y equal to the second sum, and therefore

$$Y^{[-1]} \;=\; \varepsilon\, G^\varepsilon_{p,\ldots,p}, \quad Y^{[1]} \;=\; \varepsilon\, H^\varepsilon_{p,\ldots,p},$$

with

(7.33)
$$G^\varepsilon_{\underbrace{p,\ldots,p}_{p \text{ times}}} = \sum_{\substack{((j_1,s_1),\ldots,(j_p,s_p)) \in I^p(p,\ldots,p) \\ \omega_{j_1}^{s_1}+\cdots+\omega_{j_p}^{s_p}\neq 0}} \frac{\eta_{j_1}^{s_1}\cdots\eta_{j_p}^{s_p} e^{i(\omega_{j_1}^{s_1}+\cdots+\omega_{j_p}^{s_p})t/\varepsilon} X_{j_p}^{s_p} X_{j_{p-1}}^{s_{p-1}} \cdots X_{j_1}^{s_1}}{i^{p}\omega_{j_1}^{s_1}(\omega_{j_1}^{s_1}+\omega_{j_2}^{s_2})\ldots(\omega_{j_1}^{s_1}+\cdots+\omega_{j_p}^{s_p})};$$

(7.34)
$$H^\varepsilon_{\underbrace{p,\ldots,p}_{p \text{ times}}} = \sum_{\substack{((j_1,s_1),\ldots,(j_p,s_p)) \in I^p(p\ldots p) \\ \omega_{j_1}^{s_1}+\cdots+\omega_{j_p}^{s_p}\neq 0}} \frac{\left(\frac{\mathrm{d}}{\mathrm{d}t}\left(\eta_{j_1}^{s_1}\cdots\eta_{j_p}^{s_p}\right)\right) e^{i(\omega_{j_1}^{s_1}+\cdots+\omega_{j_p}^{s_p})t/\varepsilon} X_{j_p}^{s_p} X_{j_{p-1}}^{s_{p-1}} \cdots X_{j_1}^{s_1}}{i^{p}\omega_{j_1}^{s_1}(\omega_{j_1}^{s_1}+\omega_{j_2}^{s_2})\ldots(\omega_{j_1}^{s_1}+\cdots+\omega_{j_p}^{s_p})}.$$

This allows us to rewrite the second sum in (7.32) as

$$\varepsilon\, F^\varepsilon\, G^\varepsilon_{p,\ldots,p} \;-\; \varepsilon\, G^\varepsilon_{p,\ldots,p}\,\frac{\partial}{\partial t} \;-\; \varepsilon\, H^\varepsilon_{p,\ldots,p} \;-\; \sum_{r=1}^{P} \varepsilon^{\frac{1}{r}} F^\varepsilon_{p,\ldots,p,r}$$

with

(7.35)

$$F^{\varepsilon}_{\underbrace{p,\dots,p,r}_{p \text{ times}}} = \sum_{\substack{((j_1,s_1),\dots,(j_p,s_p)) \ \in \ I^p(p,\dots,p) \\ \omega_{j_1}^{s_1}+\cdots+\omega_{j_p}^{s_p}\neq 0 \\ (j_{p+1},s_{p+1})\in I_r}} \frac{\eta_{j_1}^{s_1}\cdots\eta_{j_{p+1}}^{s_{p+1}} e^{i(\omega_{j_1}^{s_1}+\cdots+\omega_{j_{p+1}}^{s_{p+1}})t/\varepsilon}\, X_{j_{p+1}}^{s_{p+1}} X_{j_p}^{s_p}\dots X_{j_1}^{s_1}}{i^p\omega_{j_1}^{s_1}(\omega_{j_1}^{s_1}+\omega_{j_2}^{s_2})\dots(\omega_{j_1}^{s_1}+\cdots+\omega_{j_p}^{s_p})}.$$

Let us now consider the *first* sum in (7.32). From Lemma 7.1 and the fact that, from (7.14), a p-tuple that is in $I^p(p,\dots,p)$ cannot be of the type described in the third item of this lemma, all the p-tuples $((j_1,s_1),\dots,(j_p,s_p))$ in $I^p(p,\dots,p)$ such that $\omega_{j_1}^{s_1}+\cdots+\omega_{j_p}^{s_p}=0$ are exactly of the form $((j,\sigma(1)),\dots,(j,\sigma(p)))$ or $((j,-\sigma(1)),\dots,(j,-\sigma(p)))$ with $\ell(j)=p$ and $\sigma\in\mathfrak{S}(p)$. Hence the first sum may be rewritten (recall that $X_j^{-s}=X_j^s$) as

$$2\sum_{j\in J_p}\Re\left(\frac{\eta_j^1\cdots\eta_j^p}{i^{p-1}}\right)C_j$$

with

$$C_j = \sum_{\sigma\in\mathfrak{S}(p)}\frac{X_j^{\sigma(p)}X_j^{\sigma(p-1)}\cdots X_j^{\sigma(1)}}{\omega_j^{\sigma(1)}(\omega_j^{\sigma(1)}+\omega_j^{\sigma(2)})\cdots(\omega_j^{\sigma(1)}+\cdots+\omega_j^{\sigma(p-1)})}. \tag{7.36}$$

If one replaces in the above sum σ by $\sigma\circ\tau$, where τ is the permutation that sends $(1,2,\dots,p)$ on $(p,p-1,\dots,1)$ (change of indices in the summation), one gets

$$C_j = \sum_{\sigma\in\mathfrak{S}(p)}\frac{X_j^{\sigma(1)}X_j^{\sigma(2)}\cdots X_j^{\sigma(p)}}{(\omega_j^{\sigma(p)}+\cdots+\omega_j^{\sigma(2)})(\omega_j^{\sigma(p)}+\cdots+\omega_j^{\sigma(3)})\cdots(\omega_j^{\sigma(p)}+\omega_j^{\sigma(p-1)})\omega_j^{\sigma(p)}}.$$

Since $\omega_j^1+\cdots+\omega_j^p=0$, the denominator may be transformed:

$$C_j = (-1)^{p-1}\sum_{\sigma\in\mathfrak{S}(p)}\frac{X_j^{\sigma(1)}X_j^{\sigma(2)}\dots X_j^{\sigma(p)}}{\omega_j^{\sigma(1)}(\omega_j^{\sigma(1)}+\omega_j^{\sigma(2)})\cdots(\omega_j^{\sigma(1)}+\cdots+\omega_j^{\sigma(p-1)})}.$$

Finally, a combinatorial computation in the free Lie algebra (see [8], or [9] in which this identity is also obtained but in a less computational way) gives

$$\sum_{\sigma\in\mathfrak{S}(p)}\frac{X_j^{\sigma(1)}X_j^{\sigma(2)}\dots X_j^{\sigma(p)}}{\omega_j^{\sigma(1)}(\omega_j^{\sigma(1)}+\omega_j^{\sigma(2)})\cdots(\omega_j^{\sigma(1)}+\cdots+\omega_j^{\sigma(p-1)})} = \frac{1}{p}\sum_{\sigma\in\mathfrak{S}(p)}\frac{[X_j^{\sigma(1)},[X_j^{\sigma(2)},[\cdots,X_j^{\sigma(p)}]\cdots]]}{\omega_j^{\sigma(1)}(\omega_j^{\sigma(1)}+\omega_j^{\sigma(2)})\cdots(\omega_j^{\sigma(1)}+\cdots+\omega_j^{\sigma(p-1)})}.$$

Hence $C_j = \frac{(-1)^{p-1}}{p} B_j$ with B_j given by (7.18). Substituting the above in (7.32) yields

$$F^{\varepsilon}_{\underbrace{p,\dots,p}_{p \text{ times}}} = \frac{2\,(-1)^{p-1}}{p} \sum_{j \in J_p} \Re\left(\frac{\eta_j^1 \cdots \eta_j^p}{i^{p-1}}\right) B_j + \varepsilon \left(F^{\varepsilon}\, G^{\varepsilon}_{\underbrace{p,\dots,p}_{p \text{ times}}} - G^{\varepsilon}_{\underbrace{p,\dots,p}_{p \text{ times}}} \frac{\partial}{\partial t} \right) - \varepsilon\, H^{\varepsilon}_{p,\dots,p} - \sum_{r=1}^{P} \varepsilon^{\frac{1}{r}} F^{\varepsilon}_{\underbrace{p,\dots,p,r}_{p \text{ times}}}. \tag{7.37}$$

This clearly yields (7.17), ends the proof of Fact 2, and hence ends the proof of Theorem 6.5.

7.3. Proof of Theorem 4.9. Let $F^{\varepsilon} = \frac{\partial}{\partial t} + f^{\varepsilon}$ with f^{ε} the vector field associated with the right-hand side of (4.21), and let $G = \frac{\partial}{\partial t} + g$ with g the vector field associated with the right-hand side of (4.17). First, we show that F^{ε} DO-converges (see Definition 6.1) to G as ε tends to zero.

Since (4.21) is the same as (4.16) with $u_{j,s} = u^{\varepsilon}_{j,s}$ given by (4.18), F^{ε} can be expressed in the form (6.15), with all X^s_j's homogeneous of degree zero because each X^s_j corresponds to one of the $b_{\tau^s_j} v^s_j$'s and, from Proposition 4.6, all vector fields $b_{\tau^s_j} v^s_j$ are homogeneous of degree zero. We can apply Theorem 6.5 because the sets $\Omega_{n,\alpha}$ ($n = 2,\dots,N$) in the construction of Theorem 4.7 are MC and linearly independent with respect to P (see [9, section 5]). It implies that F^{ε} DO-converges, as ε tends to zero, to a vector field $F^0 = \frac{\partial}{\partial t} + f^0$ of the form (6.17), and in the definition (6.1) of DO-convergence, all differential operators are homogeneous of degree zero. We claim that $G = F^0$. Indeed, from Proposition 6.2, the property of DO-convergence implies the uniform convergence of the trajectories on finite time intervals. Therefore, the trajectories of (4.21) converge to those of $\dot{x} = f^0(t,x)$; however, from Theorem 4.7 (recall that (4.21) is the same as (4.16)–(4.18)), they converge to the trajectories of (4.17). This implies that the systems $\dot{x} = g(t,x)$ and $\dot{x} = f^0(t,x)$ are the same because they have the same trajectories. Hence, $F^0 = G$.

Finally, since F^{ε} DO-converges to $G = \frac{\partial}{\partial t} + g$ (with g autonomous) and since all differential operators in the definition of DO-convergence are homogeneous of degree zero, the asymptotic stability of the origin of (4.21), for $\varepsilon > 0$ small enough, will follow from Proposition 6.2 if we can show that the origin of (4.17) is asymptotically stable. This is a direct consequence of (4.3) to (4.5). ∎

REFERENCES

[1] R. W. Brockett, *Asymptotic stability and feedback stabilization*, in Differential Geometric Control Theory, R. W. Brockett et al., eds., Birkhäuser, Boston, 1983.

[2] J.-M. Coron, *A necessary condition for feedback stabilization*, Systems Control Lett., 14 (1990), pp. 227–232.

[3] J.-M. Coron, *Global asymptotic stabilization for controllable systems without drift*, Math. Control Signals Systems, 5 (1992), pp. 295–312.

[4] J.-M. Coron, *On the stabilization in finite time of locally controllable systems by means of continuous time-varying feedback law*, SIAM J. Control Optim., 33 (1995), pp. 804–833.

[5] G. W. Haynes and H. Hermes, *Nonlinear controllability via Lie theory*, SIAM J. Control, 8 (1970), pp. 450–460.

[6] H. Hermes, *Nilpotent and high-order approximations of vector field systems*, SIAM Rev., 33 (1991), pp. 238–264.

[7] M. Kawski, *Homogeneous stabilizing feedback laws*, Control Theory Adv. Tech., 6 (1990), pp. 497–516.

[8] J. Kurzweil and J. Jarnik, *Iterated Lie brackets in limit processes in ordinary differential equations*, Results Math., 14 (1988), pp. 125–137.
[9] W. Liu, *An approximation algorithm for nonholonomic systems*, SIAM J. Control Optim., 35 (1997), pp. 1328–1365.
[10] C. Lobry, *Contrôlabilité des systèmes non linéaires*, SIAM J. Control, 8 (1970), pp. 573–605.
[11] R. T. M'Closkey and R. M. Murray, *Nonholonomic systems and exponential convergence: Some analysis tools*, in 32nd IEEE Conference on Decision and Control, San Antonio, TX, 1993.
[12] R. T. M'Closkey and R. M. Murray, *Exponential stabilization of driftless nonlinear control systems via time-varying homogeneous feedback*, in 33rd IEEE Conference on Decision and Control, Lake Buena Vista, FL, 1994.
[13] R. T. M'Closkey and R. M. Murray, *Exponential stabilization of driftless nonlinear control systems using homogeneous feedback*, IEEE Trans. Automat. Control, 42 (1997), pp. 614–628.
[14] P. Morin and C. Samson, *Applications of backstepping techniques to the time-varying exponential stabilization of chained-form systems*, European J. Control, 3 (1997), pp. 15–36.
[15] J.-B. Pomet, *Explicit design of time-varying stabilizing control laws for a class of controllable systems without drift*, Systems Control Lett., 18 (1992), pp. 147–158.
[16] J.-B. Pomet and C. Samson, *Exponential stabilization of nonholonomic systems in power form*, in IFAC Symposium on Robust Control Design, Rio de Janeiro, 1994, pp. 447–452.
[17] L. Rosier, *Homogeneous Lyapunov function for homogeneous continuous vector field*, Systems Control Lett., 19 (1992), pp. 467–473.
[18] C. Samson, *Velocity and Torque Feedback Control of a Nonholonomic Cart*, in Advanced Robot Control, Lecture Notes in Control and Inform. Sci. 162, Springer-Verlag, New York, 1991.
[19] J. A. Sanders and F. Verhulst, *Averaging Methods in Nonlinear Dynamical Systems*, Appl. Math. Sci. 56, Springer-Verlag, New York, 1985.
[20] R. Sépulchre, G. Campion, and V. Wertz, *Some remarks on periodic feedback stabilization*, in 2nd NOLCOS, Bordeaux, France, 1992, IFAC, pp. 418–423.
[21] E. D. Sontag and H. J. Sussmann, *Remarks on continuous feedback*, in 19th IEEE Conference on Decision and Control, Albuquerque, NM, 1980.
[22] G. Stefani, *On the local controllability of scalar-input control systems*, in Theory and Applications of Nonlinear Control Systems, C. I. Byrnes and A. Lindquist, eds., North-Holland, Amsterdam, 1986, pp. 167–179.
[23] H. J. Sussmann, *Subanalytic sets and feedback control*, J. Differential Equations, 31 (1979), pp. 31–52.
[24] H. J. Sussmann, *A general theorem on local controllability*, SIAM J. Control Optim., 25 (1987), pp. 158–194.
[25] H. J. Sussmann and W. Liu, *Limits of highly oscillatory controls and approximation of general paths by admissible trajectories*, in 30th IEEE Conference on Decision and Control, Brighton, UK, 1991.
[26] H. J. Sussmann and W. Liu, *Lie bracket extensions and averaging: The single bracket case*, in Nonholonomic motion planning, Z. Li and J. Canny, eds., Kluwer, Boston, 1993, pp. 107–147.
[27] A. R. Teel, R. M. Murray, and G. Walsh, *Nonholonomic control systems: From steering to stabilization with sinusoids*, in 31st IEEE Conference on Decision and Control, Tucson, AZ, 1992.

SIAM J. CONTROL OPTIM.
Vol. 38, No. 1, pp. 50–60

SENSITIVITY OF SOLUTIONS TO VARIATIONAL INEQUALITIES ON BANACH SPACES*

ADAM B. LEVY†

Abstract. We analyze the sensitivity of parameterized variational inequalities for convex polyhedric sets in reflexive Banach spaces. We compute a generalized derivative of the solution mapping where the formula for the derivative is given in terms of the solutions to an auxiliary variational inequality. These results are distinguished from other work in this area by the fact that they do not depend on the uniqueness of the solutions to the variational inequalities. To obtain our results, we use second-order epi-derivatives to analyze the second-order properties of polyhedric sets. We apply our results to sensitivity analyses of stationary points and KKT pairs associated with constrained infinite-dimensional optimization problems.

Key words. variational inequality, sensitivity analysis, polyhedric set, parameterized optimization, second-order epi-derivative, protoderivative

AMS subject classifications. 49K40, 90C31

PII. S036301299833985X

1. Introduction.

Variational inequalities have been studied widely in infinite-dimensional optimization, particularly because they can represent solutions or solution-multiplier pairs associated with optimal control problems. In this paper, we analyze the sensitivity of solutions x in a reflexive Banach space X to variational inequalities of the form

$$\langle v - F(x,u), c - x\rangle \leq 0 \quad \forall\, c \in C, \tag{1.1}$$

where v is a parameter in the dual space X^*, u is a parameter in a Banach space U, $F : X \times U \to X^*$ is a (single-valued) mapping, and $C \subseteq X$ is a closed, nonempty, convex polyhedric set.

DEFINITION 1.1. *A subset C of a Banach space is called* polyhedric at $\bar{x} \in C$ for $x^* \in X^*$ *if the identity holds that*

$$\overline{(x^*)^\perp \bigcap \cup_{\lambda>0}\lambda(C-\bar{x})} = (x^*)^\perp \bigcap \overline{\cup_{\lambda>0}\lambda(C-\bar{x})},$$

where the overbar denotes the strong closure of the set.

Polyhedric sets are an infinite-dimensional generalization of finite-dimensional polyhedral sets; sets that are polyhedric at all their points were first studied in a Hilbert space setting [10], [4], where projections onto these sets were shown to be semidifferentiable. It is a fact that the point x is the projection of $G(u)$ onto the set C if and only if the variational inequality $\langle G(u) - x, c - x\rangle \leq 0$ for all $c \in C$ is satisfied. The authors in [10] and [4] used this fact to apply their results about projections to analyze the sensitivity of variational inequalities (1.1) of the special form where $F(x,u) = G(u) - x$, $U = \mathbb{R}$, and $v = 0$. Notice that this form of variational inequality automatically has a unique solution $x(u)$ for each u, since the projection of $G(u)$ onto the convex set C is a unique point. From their results about the semidifferentiability of projections, these authors showed that the solution function

*Received by the editors June 3, 1998; accepted for publication (in revised form) April 19, 1999; published electronically November 10, 1999.

http://www.siam.org/journals/sicon/38-1/33985.html

†Department of Mathematics, Bowdoin College, Brunswick, ME 04011 (alevy@bowdoin.edu).

$x(u)$ is semidifferentiable with respect to u. In [14] and [8], the results in [4] and [10] were applied to the sensitivity analysis of solutions and solution-multiplier pairs associated with parameterized optimization problems on Hilbert spaces. This work is complemented by [9], where the sensitivity of variational inequalities associated with convex polyhedric cones C was analyzed under a "strong regularity" assumption which involves unique and Lipschitz continuous solutions to a linearized version of the variational inequality. All of these papers include examples of polyhedric sets arising in variational inequalities associated with optimal control problems. The main result of the present paper complements these results by quantifying the sensitivity of the solutions to (1.1) without relying on some of the strong conditions found in the works of previous authors.

THEOREM 1.1. *Consider fixed parameters $\bar{v} \in X^*$ and $\bar{u} \in U$ together with a solution $\bar{x}$ to the variational inequality* (1.1) *corresponding to these parameters. If C is a convex set that is polyhedric at $\bar{x} \in C$ for $\bar{v} - F(\bar{x}, \bar{u})$ and F is semidifferentiable at $(\bar{x}, \bar{u})$ (with derivative mapping $DF(\bar{x}, \bar{u}) : X \times U \to X^*$), then the solution mapping*

$$S(u, v) := \{x \in C : \langle v - F(x, u), c - x\rangle \leq 0 \quad \forall\, c \in C\}$$

is protodifferentiable at $(\bar{u}, \bar{v})$ for $\bar{x}$ with protoderivative mapping $DS(\bar{u}, \bar{v}|\bar{x}) : U \times X^* \rightrightarrows X$ given by

$$DS(\bar{u}, \bar{v}|\bar{x})(u, v) = \{x \in C' : \langle v - DF(\bar{x}, \bar{u})(x, u), c' - x\rangle \leq 0 \quad \forall\, c' \in C'\},$$

where $C' \subseteq X$ is defined by

$$C' := \overline{(\bar{v} - F(\bar{x}, \bar{u}))^{\perp} \bigcap \cup_{\lambda > 0} \lambda(C - \bar{x})}.$$

Theorem 1.1 is distinguished from previous work in this area in that it involves no explicit assumptions about the uniqueness (or even existence) of the solutions to the variational inequality. This is representative of a fundamental shift in how sensitivity analysis can be approached: By utilizing a generalized derivative like the protoderivative, we can focus on differential properties of solutions without having to simultaneously consider properties like uniqueness and continuity. These other important properties can then be studied separately and the results combined when a combination of the properties is desired. Our approach is fruitful since the protoderivative gives important sensitivity information even when uniqueness and continuity are not present. In particular, we show that the images of the protoderivative mapping are the sets of "directional derivatives" obtained by considering different curves tangential to a direction. Moreover, nothing is lost with this approach as we show that the protoderivative is the same as the usual directional derivative for continuous functions (here called the "semiderivative").

As a precursor to our main result, we analyze the second-order properties of convex polyhedric sets C by considering their associated "indicator functions" $\delta_C : X \to \mathbb{R} \cup \{\infty\}$ defined by

$$\delta_C(x) = \begin{cases} 0 & \text{if } x \in C, \\ \infty & \text{otherwise.} \end{cases}$$

We show that δ_C is "twice Mosco epi-differentiable at $\bar{x}$ for x^*" with second-order epi-derivative equal to the indicator function associated with the set

$$\overline{(x^*)^{\perp} \bigcap \cup_{\lambda > 0} \lambda(C - \bar{x})},$$

but the second-order directional derivative along a fixed direction is the indicator function associated with the set

$$(x^*)^\perp \bigcap \cup_{\lambda>0}\lambda(C-\bar{x}).$$

These results are interesting in their own right, since they quantify the "straightness" of the boundaries of polyhedric sets in a way that distinguishes them from finite-dimensional polyhedral sets.

In the final two sections, we apply our results to abstract constrained optimization problems and we obtain results that complement those in [14] and [8]. Because our sensitivity analysis does not depend on the uniqueness of solutions, we avoid some of the strong conditions needed in [14] and [8]. Of course, the stronger conditions used by other authors give stronger consequences too. (For example, in [8] the existence, uniqueness, and Lipschitz continuity of solutions is obtained.) The abstract models that we study cover many important problems in infinite-dimensional optimization, including problems in parameterized optimal control.

2. Second-order analysis of convex polyhedric sets. The results in this section hold when X is any Banach space (not necessarily reflexive). To analyze the second-order properties of a closed, nonempty, convex polyhedric set $C \subseteq X$ at a point $\bar{x} \in C$, we fix a point x^* in the dual space X^* and construct the second-order difference quotient functions $\Delta_t^2(\bar{x}|x^*) : X \to \mathbb{R} \cup \{\infty\}$ as follows:

$$\Delta_t^2(\bar{x}|x^*)(x) := \frac{\delta_C(\bar{x}+tx) - \delta_C(\bar{x}) - t\langle x^*, x\rangle}{t^2/2}.$$

Notice that the term $\delta_C(\bar{x})$ contributes nothing since the point $\bar{x}$ is in the set C; however, we include this term so that the difference quotient will look familiar. The generalized second-order derivatives that we employ in this paper are defined as the Mosco epi-limits of the difference quotients $\Delta_t^2(\bar{x}|x^*)$. Recall from [2] that a family of functions $\Delta_t : X \to \mathbb{R} \cup \{\infty\}$ *Mosco epi-converges as* $t \downarrow 0$ to $\Delta : X \to \mathbb{R} \cup \{\infty\}$ if for every sequence of scalars $t_n \downarrow 0$, and every element $x \in X$, the two conditions hold that

$$\Delta(x) \leq \inf_{x_n \overset{w}{\to} x} \liminf_n \Delta_{t_n}(x_n),$$

$$\Delta(x) \geq \inf_{x_n \overset{s}{\to} x} \limsup_n \Delta_{t_n}(x_n),$$

where the superscripts w and s indicate weak and strong convergence, respectively. We say that the indicator function $\delta_C : X \to \mathbb{R} \cup \{\infty\}$ is *twice Mosco epi-differentiable at* $\bar{x}$ *for* x^* with second-order epi-derivative function $\delta_C''(\bar{x}|x^*) : X \to \mathbb{R} \cup \{\infty\}$ if the second-order difference quotients $\Delta_t^2(\bar{x}|x^*)$ Mosco epi-converge as $t \downarrow 0$ to $\delta_C''(\bar{x}|x^*)$ with $\delta_C''(\bar{x}|x^*)(0) = 0$. The second-order epi-derivative detects the curvature of the set C, as the following finite-dimensional example from [12] illustrates.

Example. Consider the half-disk in $\mathbb{R}^2$

$$C = \{x = (x_1, x_2) \in \mathbb{R}^2 : x_1^2 + x_2^2 \leq 2, x_1 - x_2 \leq 0\}$$

and its corner point $\bar{x} = (1,1)$. If we consider the point $x^* = (2,2)$ which solves the variational inequality $\langle x^*, c - \bar{x}\rangle \leq 0$ for all $c \in C$, we see that the second-order epi-derivative at $\bar{x}$ for x^* is

$$\delta_C''(\bar{x}|x^*)(x) = (x_1^2 + x_2^2) + \delta_{\{\lambda(-1,1):\lambda\geq 0\}}(x).$$

The $(x_1^2 + x_2^2)$ term is evidence that the second-order epi-derivative for the vector x^* in this case detects the curving on the semicircular boundary of C. In contrast, consider instead the point $x^* = (2, -2)$ which also solves the variational inequality $\langle x^*, c - \bar{x}\rangle \leq 0$ for all $c \in C$. In this case, the second-order epi-derivative at $\bar{x}$ for x^* is

$$\delta''_C(\bar{x}|x^*)(x) = \delta_{\{\lambda(-1,-1):\lambda\geq 0\}}(x).$$

The second-order epi-derivative for this vector x^* detects only the straight side of C.

Recall that a convex set C in $\mathbb{R}^n$ is called *polyhedral* if it can be specified by finitely many linear constraints. The boundaries of such sets thus consist of finitely many straight pieces. It can be deduced from [12, Exercise 13.17] that for any convex polyhedral set $C \subseteq \mathbb{R}^n$, the indicator function δ_C is twice (Mosco) epi-differentiable at any point $\bar{x} \in C$ for any point $x^* \in X^*$ solving the variational inequality $\langle x^*, c-x\rangle \leq 0$ for all $c \in C$. Moreover, the second-order epi-derivative in this case is the indicator function associated with the set

$$\overline{(x^*)^\perp \bigcap \cup_{\lambda>0}\lambda(C - \bar{x})},$$

so the second-order epi-derivative in this case detects the (lack of) curvature of the set C. The following theorem shows that this property extends to polyhedric sets on Banach spaces, a result which supports the claim that polyhedric sets are a reasonable generalization of polyhedral sets.

THEOREM 2.1. *Consider a closed, nonempty, convex set C in a Banach space X that is polyhedric at $\bar{x} \in C$ and a point $x^* \in X^*$ satisfying $\langle x^*, c-\bar{x}\rangle \leq 0$ for all $c \in C$. Then the indicator function $\delta_C : X \to \mathbb{R} \cup \{\infty\}$ is twice Mosco epi-differentiable at $\bar{x}$ for x^* and the second-order epi-derivative $\delta''_C(\bar{x}|x^*)$ satisfies*

$$\delta''_C(\bar{x}|x^*)(x) = \begin{cases} 0 & \text{if } x \in \overline{(x^*)^\perp \bigcap \cup_{\lambda>0}\lambda(C - \bar{x})}, \\ \infty & \text{otherwise.} \end{cases}$$

Proof. To prove this result, we need to verify that for any $t_n \downarrow 0$ and any $x_n \xrightarrow{w} x$, the difference quotient functions $\Delta_t^2(\bar{x}|x^*)$ satisfy

$$\liminf_n \Delta^2_{t_n}(\bar{x}|x^*)(x_n) \geq \begin{cases} 0 & \text{if } x \in \overline{(x^*)^\perp \bigcap \cup_{\lambda>0}\lambda(C - \bar{x})}, \\ \infty & \text{otherwise.} \end{cases} \tag{2.1}$$

From the definition of the difference quotient functions, it is clear that $\Delta^2_{t_n}(\bar{x}|x^*)(x_n)$ equals infinity unless $\bar{x} + t_n x_n \in C$, so if there is no sequence $\{x_n\}$ weakly converging to x and satisfying $\bar{x} + t_n x_n \in C$, then the limit inferior in (2.1) is infinity for all weakly converging sequences $x_n \xrightarrow{w} x$, and the inequality (2.1) is verified. We thus assume that there exists a sequence $x_n \xrightarrow{w} x$ with $\bar{x} + t_n x_n \in C$. In this case, according to the assumption about the point x^*, the inner product $\langle x^*, t_n x_n\rangle$ is nonpositive. It follows that the value of the difference quotient for t_n evaluated at x_n always satisfies

$$\Delta^2_{t_n}(\bar{x}|x^*)(x_n) \geq 0. \tag{2.2}$$

Thus, the only part of inequality (2.1) that needs to be verified is that the limit inferior of $\Delta^2_{t_n}(\bar{x}|x^*)(x_n)$ equals infinity when $x \notin \overline{(x^*)^\perp \bigcap \cup_{\lambda>0}\lambda(C - \bar{x})}$. Since C is polyhedric at $\bar{x}$ for x^*, the set $\overline{(x^*)^\perp \bigcap \cup_{\lambda>0}\lambda(C - \bar{x})}$ is the same as the set

$(x^*)^\perp \bigcap \overline{\cup_{\lambda>0}\lambda(C-\bar{x})}$, so to verify the inequality (2.1) in this case, we assume either that x is not perpendicular to x^* or that x is not an element of the set $\overline{\cup_{\lambda>0}\lambda(C-\bar{x})}$. We show first that this latter case cannot occur.

Case. $x \notin \overline{\cup_{\lambda>0}\lambda(C-\bar{x})}$ cannot occur. Since we have assumed throughout that $\bar{x}+t_n x_n \in C$ with $x_n \stackrel{w}{\to} x$, it follows that $x_n \in \cup_{\lambda>0}\lambda(C-\bar{x})$ so that x is in the weak closure of the set $\cup_{\lambda>0}\lambda(C-\bar{x})$. However, the set $\cup_{\lambda>0}\lambda(C-\bar{x})$ is convex (since C is convex) so its weak closure is the same as its strong closure $\overline{\cup_{\lambda>0}\lambda(C-\bar{x})}$. We conclude that $x \in \overline{\cup_{\lambda>0}\lambda(C-\bar{x})}$.

Case. $\langle x^*, x\rangle \neq 0$. Since $\bar{x}+t_n x_n \in C$, the difference quotient for t_n evaluated at x_n satisfies

$$\Delta^2_{t_n}(\bar{x}|x^*)(x_n) = \frac{-t_n\langle x^*, x_n\rangle}{t_n^2/2}. \tag{2.3}$$

However, since the term $\langle x^*, x_n\rangle$ converges to $\langle x^*, x\rangle$ which is nonzero in this case, it follows from (2.3) and the inequality (2.2) that the limit inferior of the difference quotient for t_n evaluated at x_n is infinity.

This establishes the inequality (2.1), so we will be finished if we can also verify that for any $t_n \downarrow 0$, there exists $x_n \stackrel{s}{\to} x$ satisfying the inequality

$$\limsup_n \Delta^2_{t_n}(\bar{x}|x^*)(x_n) \leq \begin{cases} 0 & \text{if } x \in \overline{(x^*)^\perp \bigcap \cup_{\lambda>0}\lambda(C-\bar{x})}, \\ \infty & \text{otherwise.} \end{cases} \tag{2.4}$$

It follows immediately that we need only consider the case

$$x \in \overline{(x^*)^\perp \bigcap \cup_{\lambda>0}\lambda(C-\bar{x})},$$

which implies that there exists a sequence $\xi_n \stackrel{s}{\to} x$ satisfying $\langle x^*, \xi_n\rangle = 0$ and with $\xi_n \in \lambda_n(C-\bar{x})$ for some positive scalars λ_n. For any $t \in [0, 1/\lambda_n]$, we can write

$$\bar{x}+t\xi_n = (1-t\lambda_n)\bar{x} + t\lambda_n(\bar{x}+\xi_n/\lambda_n),$$

which is a convex combination of elements in C. From the convexity of C, we conclude that any such points $\bar{x}+t\xi_n$ are contained in C. We construct a sequence of points x_n from the sequence $\{\xi_n\}$ in the following manner. We set $x_1 = \xi_1$, $x_2 = \xi_1, \ldots, x_n = \xi_1$ for all n until t_n is smaller than $1/\lambda_2$, at which point we begin setting $x_n = \xi_2$ for all n until t_n is less than $1/\lambda_3$ and continue this process ad infinitum. The resulting sequence $\{x_n\}$ converges strongly to x, satisfies $\langle x^*, x_n\rangle = 0$, and has $\bar{x}+t_n x_n \in C$ for all n larger than the one where we began using ξ_2. It follows that the difference quotient for t_n evaluated at these x_n equals zero, so that its limit superior satisfies the inequality (2.4).

The second-order epi-derivative proposed in this case equals zero when evaluated at zero, so the proof is complete. □

In order to see how the second-order epi-derivative compares to a more classical notion of second-order directional derivative, we consider a point $\bar{x} \in C$ and a point $x^* \in X^*$ satisfying $\langle x^*, c-\bar{x}\rangle \leq 0$ for all $c \in C$, and take the limit as $t \downarrow 0$ of the difference quotient functions $\Delta^2_t(\bar{x}|x^*)$ in a fixed direction x:

$$\lim_{t\downarrow 0} \frac{\delta_C(\bar{x}+tx) - \delta_C(\bar{x}) - t\langle x^*, x\rangle}{t^2/2}. \tag{2.5}$$

PROPOSITION 2.1. *Consider a closed, nonempty, convex set C in the Banach space X, a point $\bar{x} \in C$ and a point $x^* \in X^*$ satisfying $\langle x^*, c - \bar{x}\rangle \leq 0$ for all $c \in C$. Then the second-order directional derivative along any fixed direction $x \in X$ satisfies*

$$\lim_{t \downarrow 0} \frac{\delta_C(\bar{x} + tx) - \delta_C(\bar{x}) - t\langle x^*, x\rangle}{t^2/2} = \begin{cases} 0 & \text{if } x \in (x^*)^\perp \bigcap \cup_{\lambda>0}\lambda(C - \bar{x}), \\ \infty & \text{otherwise.} \end{cases}$$

Proof. Just as in the proof of Theorem 2.1, if $x \in \lambda(C - \bar{x})$, then we know that $\bar{x} + tx \in C$ for all $t \in [0, 1/\lambda]$. It follows that if x is an element of the intersection $(x^*)^\perp \bigcap \cup_{\lambda>0}\lambda(C - \bar{x})$, then the limit (2.5) is equal to zero. Moreover, the difference quotient in (2.5) is always nonnegative (again according to an argument similar to the one in the proof of Theorem 2.1), so if x is not perpendicular to x^*, then the limit (2.5) is infinite. Finally, if x is not in the set $\cup_{\lambda>0}\lambda(C - \bar{x})$ it follows that the term $\delta_C(\bar{x} + tx)$ is infinite for all $t > 0$, so the limit (2.5) is infinite too. ∎

It follows from Theorem 2.1 and Proposition 2.1 that the second-order epi-derivative is the same as the second-order directional derivative along a fixed direction if and only if the set $(x^*)^\perp \bigcap \cup_{\lambda>0}\lambda(C - \bar{x})$ is closed. This always occurs when C is a polyhedral set in $\mathbb{R}^n$, so that the two second-order derivatives agree in the case of polyhedral C. However, this is not necessarily true for polyhedric sets in Banach spaces, as the following example shows.

Example. Consider the Banach space X of continuous functions on the interval $[-1, 1]$ under the supremum norm

$$\| x \| := \sup_{\tau \in [-1,1]} |x(\tau)|,$$

the set $C = \{x \in X : x(\tau) \leq 0 \text{ for all } \tau \in [-1, 1]\}$, and the function $\bar{x} \in C$ defined by $\bar{x}(\tau) = -\tau^4$. This set C is trivially polyhedric at $\bar{x}$ for $x^* = 0 \in X^*$ (any set is polyhedric at all of its points for $x^* = 0$). Now consider the function $x \in X$ defined by $x(\tau) = \tau^2$. In order for x to be an element of the set $\cup_{\lambda>0}\lambda(C - \bar{x})$, there must be some $\lambda > 0$ with $\tau^2 \leq \lambda\tau^4$ for all $\tau \in [-1, 1]$. Since no such λ exists, we conclude that x is not in the set $\cup_{\lambda>0}\lambda(C - \bar{x})$. However, if we define the continuous functions x_n on $[-1, 1]$ by

$$x_n(\tau) := \begin{cases} n\tau^4 & \text{for } \tau \in \left[-\sqrt{\frac{1}{n}}, \sqrt{\frac{1}{n}}\right], \\ \tau^2 & \text{for } \tau < -\sqrt{\frac{1}{n}} \text{ or } \tau > \sqrt{\frac{1}{n}}, \end{cases}$$

then these converge to x as $n \to \infty$ and satisfy $x_n \in n(C - \bar{x})$. Therefore, x is contained in $\overline{\cup_{\lambda>0}\lambda(C - \bar{x})}$. It follows that the second-order epi-derivative satisfies $\delta''_C(\bar{x}|x^*)(x) = 0$, while the second-order directional derivative along the fixed direction x (2.5) is infinite.

This example highlights the fact that polyhedric sets in Banach spaces have properties distinct from polyhedral sets in $\mathbb{R}^n$. In particular, polyhedric sets C in Banach spaces have "straight" boundaries as detected by the second-order epi-derivative $\delta''_C(\bar{x}|x^*)$. However, this straightness is not the same as it is for polyhedral sets whose boundary "straightness" is also detected by the second-order directional derivative along fixed directions (2.5).

3. Protoderivatives of solutions to variational inequalities. It is convenient to represent the solutions to the variational inequality (1.1) as the values of the multifunction $S : U \times X^* \rightrightarrows X$ defined by

$$S(u, v) := \{x : v - F(x, u) \in N_C(x)\}, \tag{3.1}$$

where the normal cone mapping $N_C : X \rightrightarrows X^*$ associated with the set C is defined for $x \in X$ by

$$N_C(x) := \begin{cases} \{x^* : \langle x^*, c - x\rangle \le 0 \text{ for all } c \in C\} & \text{if } x \in C, \\ \emptyset & \text{otherwise.} \end{cases} \tag{3.2}$$

"Protoderivatives" of these kinds of solution multifunctions have been studied in [6], where they are shown to depend on the protodifferentiability of the normal cone mapping N_C. The protoderivative of any multifunction $S : Z \rightrightarrows Y$ between two Banach spaces can be defined in terms of the contingent cone to the graph of S at $(\bar{z}, \bar{y}) \in \operatorname{gph} S$:

$$T_{\operatorname{gph} S}(\bar{z}, \bar{y}) := \limsup_{t \downarrow 0} \frac{\operatorname{gph} S - (\bar{z}, \bar{y})}{t},$$

defined in terms of "$\limsup_{t \downarrow 0}$" denoting the outer set limit as $t \downarrow 0$, which is the set of all points obtained as limits of points in the sets $\big(\operatorname{gph} S - (\bar{z}, \bar{y})\big)/t_n$ for some sequence $t_n \downarrow 0$. The multifunction $S : Z \rightrightarrows Y$ is *protodifferentiable at $\bar{z}$ for $\bar{y}$* if every element in the contingent cone $T_{\operatorname{gph} S}(\bar{z}, \bar{y})$ can actually be obtained as a limit

$$\lim_{t \downarrow 0} \frac{\xi(t) - (\bar{z}, \bar{y})}{t} \tag{3.3}$$

for some selection mapping $\xi : [0, \epsilon] \to \operatorname{gph} S$ with $\epsilon > 0$. Under these circumstances, the protoderivative of S at $\bar{z}$ for $\bar{y}$ is the multifunction $DS(\bar{z}|\bar{y}) : Z \rightrightarrows Y$ whose graph is the contingent cone $T_{\operatorname{gph} S}(\bar{z}, \bar{y})$. The following is a consequence of the second-order Mosco epi-differentiability demonstrated in Theorem 2.1.

THEOREM 3.1. *Consider a closed, nonempty, convex set C in a reflexive Banach space X that is polyhedric at $\bar{x} \in C$ and a point $x^* \in X^*$ satisfying $\langle x^*, c - \bar{x}\rangle \le 0$ for all $c \in C$. Then the normal cone mapping $N_C : X \rightrightarrows X^*$ is protodifferentiable at $\bar{x}$ for x^* with protoderivative equal to the normal cone mapping associated with the set*

$$C' := \overline{(x^*)^\perp \bigcap \cup_{\lambda > 0} \lambda(C - \bar{x})}.$$

Proof. This follows from Theorem 2.1 and [11, Theorem 5] since the normal cone mapping associated with a convex set is the convex subdifferential of the indicator function. □

Remark. The result [11, Theorem 5] used in the proof of Theorem 3.1 is based on Attouch's theorem [1], which connects the epi-convergence of convex functions on reflexive Banach spaces with the graphical convergence of their subgradient mappings. Because of this connection, Theorem 3.1 itself must be stated in terms of reflexive Banach spaces instead of the more general Banach spaces considered in section 2. Our main result, Theorem 1.1, follows immediately from results now obtained.

Proof of Theorem 1.1. This follows immediately from Theorem 3.1 and [6, Theorem 4.1]. The latter result gave the protodifferentiability of general solution mappings of the form

$$S(u, v) := \{x : v - F(x, u) \in M(x, u)\}$$

as long as F is semidifferentiable and the set-valued mapping M is protodifferentiable. According to Theorem 3.1, the normal cone mapping N_C studied here is one example

of a protodifferentiable set-valued mapping M, so the protodifferentiability of the solution mapping (3.1) follows from [6, Theorem 4.1]. Moreover, a formula was given in [6, Theorem 4.1] for the protoderivative of the general solution mapping, and this formula specializes in the case of (3.1) to the formula claimed in the statement of Theorem 1.1. □

Remark. Note that it is important that the parameter $v \in X^*$ appear in the variational inequality (1.1) if we are to deduce protodifferentiability from [6, Theorem 4.1]. If the solution mapping (3.1) does not depend explicitly on v, we can obtain estimates for related generalized derivatives called "outer graphical derivatives" via [5].

Protoderivatives for multifunctions between two Euclidean spaces have been studied widely (see [12] for a survey), where they have been shown to have many useful properties. In the case when S is a (single-valued) continuous function, the protoderivative is related closely to the "semiderivative." A continuous function $S : Z \to Y$ between two Banach spaces is *semidifferentiable at $\bar{z}$ with semiderivative $DS(\bar{z}) : Z \to Y$* if for every $\phi : \mathbb{R}^+ \to Z$ for which $(\phi(t) - \bar{z})/t$ converges to some point $z \in Z$, the limit exists

$$DS(\bar{z})(z) = \lim_{t \downarrow 0} \frac{S(\phi(t)) - S(\bar{z})}{t}.$$

(See [13] for a discussion of equivalent definitions of this derivative which the author calls the "Hadamard" directional derivative.) It was shown in [7] that a locally Lipschitzian mapping $S : \mathbb{R}^n \to \mathbb{R}^m$ is protodifferentiable at $\bar{z}$ for $\bar{y} = S(\bar{z})$ if and only if it is semidifferentiable at $\bar{z}$. (An equivalent derivative called the "B-derivative" was used in [7].) The following proposition clarifies the relationship between the protoderivative and the semiderivative in the case of continuous functions between two Banach spaces.

PROPOSITION 3.1. (i) *If the continuous function $S : Z \to Y$ is semidifferentiable at $\bar{z}$, then S is protodifferentiable at $\bar{z}$ for $\bar{y} = S(\bar{z})$ with protoderivative $DS(\bar{z}|\bar{y})$ equal to the semiderivative $DS(\bar{z})$.*

(ii) *If the multifunction $S : Z \to Y$ is protodifferentiable at $\bar{z}$ for $\bar{y} = S(\bar{z})$, then for any $\phi : \mathbb{R}^+ \to Z$ with $(\phi(t) - \bar{z})/t \to z$ and $\big(S(\phi(t)) - S(\bar{z})\big)/t \to y$ as $t \downarrow 0$, the limit y is in the image of the protoderivative $DS(\bar{z}|\bar{y})(z)$.*

Proof. (i) We consider any pair (z, y) in the contingent cone to the graph of S at $(\bar{z}, \bar{y})$. Then there exist sequences $t_n \downarrow 0$ and $(z_n, y_n) \to (z, y)$ which satisfy

$$y_n \in \frac{S(\bar{z} + t_n z_n) - \bar{y}}{t_n}.$$

Since S is semidifferentiable at $\bar{z}$, it follows that y must be equal to the semiderivative evaluated in the direction z, so the protoderivative (if it exists) is equal to the semiderivative. Moreover, if we define the selection mapping $\xi(t) := (\bar{z} + tz, S(\bar{z} + tz))$, the limit (3.3) satisfies

$$\lim_{t \downarrow 0} \frac{\xi(t) - (\bar{z}, S(\bar{z}))}{t} = \left(z, \lim_{t \downarrow 0} \frac{S(\bar{z} + tz) - S(\bar{z})}{t} \right) = (z, y),$$

so S is protodifferentiable at $\bar{z}$ for $\bar{y} = S(\bar{z})$.

(ii) For any sequence $t_n \downarrow 0$, the sequences $z_n := (\phi(t_n) - \bar{z})/t_n$ and $y_n := \big(S(\phi(t_n) - S(\bar{z})\big)/t_n$ satisfy

$$(z_n, y_n) \in \frac{\operatorname{gph} S - (\bar{z}, \bar{y})}{t_n}$$

and converge to z and y, respectively. It follows that the pair (z, y) is in the contingent cone to $\operatorname{gph} S$ at $(\bar{z}, \bar{y})$, which is the same as the graph of the protoderivative $DS(\bar{z}|\bar{y})$. □

According to the definition of the protoderivative, each element in the set $DS(\bar{z}|\bar{y})(z)$ can be obtained as the limit of $\big(S(\phi(t)) - \bar{z}\big)/t$ for some ϕ as in (ii) of Theorem 3.1. This fact together with part (ii) of Theorem 3.1 shows that the image of the protoderivative $DS(\bar{z}|\bar{y})(z)$ is precisely the set of "directional derivatives" obtained by considering different curves ϕ tangential to the direction z.

4. Sensitivity of stationary points. We consider the optimization problem

$$\min\ g_0(x, u) - \langle v, x\rangle \text{ over all } x \in C, \tag{4.1}$$

where C is a closed, nonempty, convex set in a reflexive Banach space X, and the functional $g_0 : X \times U \to \mathbb{R}$ is continuously differentiable with respect to x. The stationary points associated with this problem for parameters $u \in U$ and $v \in X^*$ are the solutions to the variational inequality

$$v - \nabla_x g_0(x, u) \in N_C(x),$$

this being a first-order necessary condition for optimality in (4.1).

THEOREM 4.1. *Consider the parameters $\bar{v} \in X^*$ and $\bar{u} \in U$ together with a stationary point $\bar{x} \in C$ satisfying $\bar{v} - \nabla_x g_0(\bar{x}, \bar{u}) \in N_C(\bar{x})$. If the gradient mapping $\nabla_x g_0$ is semidifferentiable at $(\bar{x}, \bar{u})$ and the set C is polyhedric at $\bar{x}$ for $\bar{v} - \nabla_x g_0(\bar{x}, \bar{u})$, then the stationary point mapping*

$$S(u, v) := \{x \in C : v - \nabla_x g_0(x, u) \in N_C(x)\}$$

is protodifferentiable at $(\bar{u}, \bar{v})$ for $\bar{x}$ with protoderivative mapping $DS(\bar{u}, \bar{v}|\bar{x}) : U \times X^* \rightrightarrows X$ given by

$$DS(\bar{u}, \bar{v}|\bar{x})(u, v) = \big\{x \in C' : v - D\big(\nabla_x g_0\big)(\bar{x}, \bar{u})(x, u) \in N_{C'}(x)\big\},$$

where $C' \subseteq X$ is defined by

$$C' := \overline{(\bar{v} - \nabla_x g_0(\bar{x}, \bar{u}))^\perp \bigcap \cup_{\lambda>0} \lambda(C - \bar{x})}.$$

Proof. This follows immediately from Theorem 1.1. □

Theorem 4.1 compares to [14, Theorem 2], where an abstract optimal control problem is shown to have semidifferentiable solutions. In contrast to Theorem 4.1, however, strong second-order conditions were needed in [14, Theorem 2] to ensure unique solutions to the variational inequality.

5. Sensitivity of Karush–Kuhn–Tucker pairs. Our results also apply to solution-multiplier pairs associated with optimization problems like

$$\min\ g_0(x, u) - \langle v_1, x\rangle \text{ over all } x \in C(u, v_2), \tag{5.1}$$

where the constraint set $C(u, v_2)$ is defined for parameters $u \in U$ and $v_2 \in Y^*$ by

$$C(u, v_2) := \{x \in C : v_2 + g(x, u) \in K\}$$

for a mapping $g : X \times U \to Y^*$ that is continuously differentiable with respect to x, a closed, nonempty, convex set C in a reflexive Banach space X and a closed, nonempty,

convex cone K (with vertex at the origin) in a reflexive Banach space Y. For such an optimization problem, we can construct the Lagrangian

$$L(x,u,y) := g_0(x,u) + \langle y, g(x,u)\rangle,$$

in terms of which we can represent the Karush–Kuhn–Tucker (KKT) pairs $(x,y) \in C \times K^*$ (where K^* is polar cone to K, containing the elements $y^* \in Y^*$ which satisfy $\langle y^*, y\rangle \le 0$ for all $y \in K$) for the problem (5.1) which satisfy

$$v_1 - \nabla_x L(x,u,y) \in N_C(x), \quad v_2 + g(x,u) \in N_{K^*}(y). \tag{5.2}$$

The set of KKT pairs satisfying (5.2) for parameters (u, v_1, v_2) is the same as the image of the mapping

$$\begin{aligned} S(u,v_1,v_2) := \{(x,y) \in C \times K^* : (v_1,v_2) \\ - \big(\nabla_x L(x,u,y), -g(x,u)\big) \in N_{C\times K^*}(x,y)\big\}. \end{aligned} \tag{5.3}$$

This follows since the product set $N_C(x) \times N_{K^*}(y)$ is the same as the normal cone of the product $C \times K^*$ at the point (x,y).

THEOREM 5.1. *Consider the parameters $\bar v = (\bar v_1, \bar v_2) \in V := X^* \times Y^*$ and $\bar u \in U$ together with a KKT pair $(\bar x, \bar y) \in C \times K^*$. If the gradient mapping $\nabla_x L$ is semidifferentiable at $(\bar x, \bar u, \bar y)$, the mapping g is semidifferentiable at $(\bar x, \bar u)$, and the set $C \times K^*$ is polyhedric at $(\bar x, \bar y)$ for $\bar v - \big(\nabla_x L(\bar x, \bar u, \bar y), -g(\bar x, \bar u)\big)$, then the KKT-pair mapping* (5.3) *is protodifferentiable at $(\bar u, \bar v)$ for $(\bar x, \bar y)$ with protoderivative $DS(\bar u, \bar v|\bar x, \bar y) : U \times V \rightrightarrows X \times Y$ given by*

$$\begin{aligned} &DS(\bar u, \bar v|\bar x, \bar y)(u,v) \\ &\quad = \Big\{(x,y) \in C' : v - \Big(D\big(\nabla_x L\big)(\bar x, \bar u, \bar y)(x,u,y), -Dg(\bar x, \bar u)(x,u)\Big) \in N_{C'}(x,y)\Big\}, \end{aligned}$$

where $C' \subseteq X \times Y$ is defined by

$$C' := \overline{\Big(\bar v - \big(\nabla_x L(\bar x, \bar u, \bar y), -g(\bar x, \bar u)\big)\Big)^{\perp} \bigcap \cup_{\lambda>0} \lambda\big(C \times K^* - (\bar x, \bar y)\big)}.$$

Proof. This follows from Theorem 1.1 since under our assumptions, the mapping

$$(x,u,y) \mapsto \big(\nabla_x L(x,u,y), -g(x,u)\big)$$

is semidifferentiable at $(\bar x, \bar u, \bar y)$ with semiderivative

$$\Big(D\big(\nabla_x L\big)(\bar x, \bar u, \bar y)(x,u,y), -Dg(\bar x, \bar u)(x,u)\Big). \qquad \square$$

This result compares to results in [8] and [9], where second-order conditions and regularity conditions were used to guarantee unique KKT pairs that are Lipschitz continuous and semidifferentiable. Related results are given in [3], where similar but weaker conditions were used to study the directional stability of nearly optimal solution selections.

REFERENCES

[1] H. Attouch, *Convergence de fonctions convexes, des sous-différentials et semi-groupes associés*, C.R. Acad. Sci. Paris Sér I Math., 284 (1977), pp. 539–542.
[2] H. Attouch, *Variational Convergence of Functions and Operators*, Pitman, Boston, MA, 1984.
[3] J.F. Bonnans and A. Shapiro, *Optimization problems with perturbations: A guided tour*, SIAM Rev., 40 (1998), pp. 228–264.
[4] A. Haraux, *How to differentiate the projection on a convex set in Hilbert space. Some applications to variational inequalities*, J. Math. Soc. Japan, 29 (1977), pp. 615–631.
[5] A.B. Levy, *Implicit multifunction theorems for the sensitivity analysis of variational conditions*, Math. Programming, 74 (1996), pp. 333–350.
[6] A.B. Levy and R.T. Rockafellar, *Sensitivity analysis of solutions to generalized equations*, Trans. Amer. Math. Soc., 345 (1994), pp. 661–671.
[7] A.B. Levy and R.T. Rockafellar, *Protoderivatives and the geometry of solution mappings in nonlinear programming*, in Nonlinear Optimization and Applications, Plenum Press, New York, 1996, pp. 343–365.
[8] K. Malanowski, *Second-order conditions and constraint qualifications in stability and sensitivity analysis of solutions to optimization problems in Hilbert spaces*, Appl. Math. Optim., 25 (1992), pp. 51–79.
[9] K. Malanowski, *Two-norm approach in stability and sensitivity analysis of optimization and optimal control problems*, Adv. Math. Sci. Appl., 2 (1993), pp. 397–443.
[10] F. Mignot, *Contrôle dans les inéquations variational elliptiques*, J. Funct. Anal., 22 (1976), pp. 130–185.
[11] R.T. Rockafellar, *Nonsmooth analysis and parametric optimization*, in Methods of Nonconvex Analysis, A. Cellina, ed., Lecture Notes in Math. 1446, Springer-Verlag, New York, 1990.
[12] R.T. Rockafellar and R.J.-B. Wets, *Variational Analysis*, Springer-Verlag, New York, 1998.
[13] A. Shapiro, *On concepts of directional differentiability*, J. Optim. Theory Appl., 66 (1990), pp. 477–487.
[14] J. Sokolowski, *Differential stability of solutions to constrained optimization problems*, Appl. Math. Optim., 13 (1985), pp. 97–115.

SIAM J. CONTROL OPTIM.
Vol. 38, No. 1, pp. 61–78

RISK-SENSITIVE CONTROL OF DISCRETE-TIME MARKOV PROCESSES WITH INFINITE HORIZON*

G. B. DI MASI† AND L. STETTNER‡

Abstract. In this paper we study existence of solutions to the Bellman equation corresponding to risk-sensitive ergodic control of discrete-time Markov processes using three different approaches. Also, for particular classes of systems, asymptotics for vanishing risk factor is investigated, showing that in the limit the optimal value for an average cost per unit time is obtained.

Key words. controlled discrete-time Markov processes, exponential ergodic performance criterion, Bellman equation

AMS subject classifications. Primary, 93E20; Secondary, 60J05, 93C55

PII. S0363012997320614

1. Introduction.

On a probability space $(\Omega, \mathcal{F}, P)$ consider a controlled Markov process $X = (x_n)$ taking values on a complete separable metric state space E endowed with the Borel σ-algebra $\mathcal{E}$. Assume that x_n has a controlled transition operator $P^{a_n}(x_n, \cdot)$, where a_n is the control at time n taking values on a compact metric space U and adapted to the σ-algebra $\sigma\{x_0, x_1, \ldots, x_n\}$.

Let $c : E \times U \to R^+$ be a continuous and bounded function. Our aim is to minimize the following exponential ergodic performance criterion:

$$J_x\big((a_n)\big) = \lim_{n\to\infty} \sup \frac{1}{n} \log E_x \left\{ \exp \left\{ \sum_{i=0}^{n-1} c(x_i, a_i) \right\} \right\}. \tag{1.1}$$

Consider the following assumption:

(B1) $\forall f \in C(E)$ the mapping $E \times U \ni (x, a) \mapsto P^a f(x)$ is continuous.

In (B1) and in what follows we shall denote by $C(E)$ the space of continuous bounded real valued functions on E endowed with the uniform norm $||\cdot||$.

The following result that is proved in [10, Theorem 2.1] for countable state space can be easily extended to our general state space model.

PROPOSITION 1.1. *If there exists a function $w \in C(E)$ and a constant λ such that for $x \in E$,*

$$e^{w(x)+\lambda} = \inf_{a\in U} \left[e^{c(x,a)} \int_E e^{w(y)} P^a(x, dy) \right], \tag{1.2}$$

then under (B1)

$$\lambda = \inf_{(a_n)} J_x\big((a_n)\big) = J_x\big(u(x_n)\big), \tag{1.3}$$

*Received by the editors April 30, 1997; accepted for publication (in revised form) March 19, 1999; published electronically November 10, 1999.

http://www.siam.org/journals/sicon/38-1/32061.html

†Universitá di Padova, Dipartimento di Matematica, Pura ed Applicata, Via Belzoni 7, 35131 Padova, Italy (dimasi@galileo.math.unipd.it).

‡Institute of Mathematics, Polish Academy of Sciences, Sniadeckich 8, 00-950 Warsaw, Poland (stettner@impan.gov.pl). The research of this author was partially supported by KBN grant 2 P03A 05309.

where $u : E \to U$ is a Borel function for which the inf *in* (1.2) *is attained.*

Infinite horizon risk-sensitive problems were studied for linear diffusions with exponential quadratic cost in a series of papers (see, e.g., [15], [8], [13], and references therein). General diffusion models were studied in [11] and [7]. Risk-sensitive control of Markov processes with cost criterion (1.1) was studied only in [9] and [6].

Our work was motivated by [9], where the existence of solutions to the Bellman equation (1.2) is proved for countable state space under the assumption of the existence of a uniformly (with respect to all stationary Markov controls and initial states) positive recurrent state.

In this paper we show the existence of solutions to the Bellman equation under different assumptions and using different approaches. In particular, in section 2 we show the existence of a unique solution using a certain span-norm contraction. In section 3 we study the discounted exponential cost criterion and prove the convergence of the solution to the discounted Bellman equation to the solution of (1.2) when the discount factor approaches 1. In section 4 we show the existence of solutions to the Bellman equation using the so-called stochastic discounted game approach. We then complete the paper studying in section 5 the asymptotic behavior for vanishing risk factor and showing that the limit corresponds to the average cost per unit time problem. Although we formulate our results for general states, the assumptions imposed in this paper seem to be rather restrictive if the state space is not compact. We also assume that the cost function $c(x, a)$ is bounded, similarly as in the span approach to risk neutral problems in section 3.3 of [10]. Nevertheless we think that this paper could motivate further research on extending the results under weaker assumptions.

2. Span contraction approach. Notice first that due to Lemma 3.3 in [9] or Proposition 2.3 in [3], (1.2) can be written in the following equivalent form:

$$w(x) + \lambda = \inf_{a \in U} \sup_{\mu \in \mathcal{P}(E)} \left[c(x,a) + \int_E w(y)\mu(dy) - I\big(\mu, P^a(x,\cdot)\big) \right], \tag{2.1}$$

where $\mathcal{P}(E)$ is the space of probability measures on E and $I(\mu, \nu)$, called mutual entropy of μ and ν is defined as follows:

$$I(\mu,\nu) = \begin{cases} \displaystyle\int_E \log \tfrac{d\mu}{d\nu}\, \mu\,(dx) & \text{when } \mu \text{ is absolutely continuous with respect to } \nu, \\ +\infty & \text{otherwise.} \end{cases}$$

Moreover, the sup in (2.1) is attained for

$$\mu^*(dz) = \frac{e^{w(z)} P^a(x, dz)}{\int_E e^{w(y)} P^a(x, dy)}.$$

Remark 2.1. The function w in (2.1) corresponds to the upper value in a two person stochastic dynamic game when the first player (minimizer) chooses at each step a control parameter a while the second player chooses the system dynamics, given by the measure μ, so as to maximize the reward function

$$\limsup_{n\to\infty} n^{-1} E_x \left\{ \sum_{i=0}^{n-1} \Big(c(x_i, a_i) - I\big(\mu_i, P^{a_i}(x_i, \cdot)\big) \Big) \right\} \tag{2.2}$$

(see [9], where a similar discounted game was considered).

The following assumption (A1) is frequently imposed for studying average cost per unit time problems (see, e.g., [10]):

(A1) $\exists_{\delta<1}$ such that $\forall_{x,x'\in E}$, $\forall_{a,a'\in U}$, $\forall_{B\in\mathcal{E}}$

$$P^a(x,B) - P^{a'}(x',B) \le \delta.$$

For $g \in C(E)$ define the operator

$$Tg(x) = \inf_{a\in U}\left[c(x,a) + \log\int_E e^{g(y)}P^a(x,dy)\right] \tag{2.3}$$

which, by Lemma 3.3 in [9] or Proposition 2.3 in [3], takes the equivalent form

$$Tg(x) = \inf_{a\in U}\sup_{\mu\in\mathcal{P}(E)}\left[c(x,a) + \int_E g(y)\mu(dy) - I\big(\mu, P^a(x,\cdot)\big)\right]. \tag{2.4}$$

The sup in (2.4) is attained for

$$\mu^*(B) = \frac{\int_B e^{g(z)}P^a(x,dz)}{\int_E e^{g(z)}P^a(x,dz)}. \tag{2.5}$$

PROPOSITION 2.2. *Under* (A1) *and* (B1) *the operator* T *is a local contraction in* $C(E)$ *endowed with the span norm* $||g||_{sp} = \sup_{x\in E} g(x) - \inf_{x\in E} g(x)$*; namely, for each* $M > 0$*, there exists a constant* $\alpha(M) < 1$ *such that for each* $g_1, g_2 \in C(E)$ *with* $||g_1||_{sp} \le M$ *and* $||g_2||_{sp} \le M$ *we have*

$$||Tg_1 - Tg_2||_{sp} \le \alpha(M)||g_1 - g_2||_{sp}. \tag{2.6}$$

Proof. Because of assumption (B1) formula (2.3) shows that T transforms $C(E)$ into itself and the inf is attained.

For given functions g_1, g_2 in $C(E)$ and $x_1, x_2 \in E$, let a_1, a_2 be such that

$$Tg_i(x_i) = \sup_{\mu\in\mathcal{P}(E)}\left\{c(x_i,a_i) + \int g_i(y)\mu(dy) - I\big(\mu, P^{a_i}(x_i,\cdot)\big)\right\}.$$

Moreover, let

$$\mu_1(B) = \frac{\int_B e^{g_2(z)}P^{a_1}(x_1,dz)}{\int_E e^{g_2(z)}P^{a_1}(x_1,dz)},$$
$$\mu_2(B) = \frac{\int_B e^{g_1(z)}P^{a_2}(x_2,dz)}{\int_E e^{g_1(z)}P^{a_2}(x_2,dz)}.$$

Then

$$
\begin{aligned}
&Tg_1(x_2) - Tg_2(x_2) - \big(Tg_1(x_1) - Tg_2(x_1)\big) \\
&\le \sup_{\mu\in\mathcal{P}(E)} \left\{ c(x_2,a_2) + \int_E g_1(y)\mu(dy) - I\big(\mu, P^{a_2}(x_2,\cdot)\big) \right\} \\
&- \sup_{\mu\in\mathcal{P}(E)} \left\{ c(x_2,a_2) + \int_E g_2(y)\mu(dy) - I\big(\mu, P^{a_2}(x_2,\cdot)\big) \right\} \\
&- \sup_{\mu\in\mathcal{P}(E)} \left\{ c(x_1,a_1) + \int_E g_1(y)\mu(dy) - I\big(\mu, P^{a_1}(x_1,\cdot)\big) \right\} \\
&+ \sup_{\mu\in\mathcal{P}(E)} \left\{ c(x_1,a_1) + \int_E g_2(y)\mu(dy) - I\big(\mu, P^{a_1}(x_1,\cdot)\big) \right\} \\
&\le c(x_2,a_2) + \int_E g_1(y)\mu_2(dy) - I\big(\mu_2, P^{a_2}(x_2,\cdot)\big) \qquad (2.7)\\
&-c(x_2,a_2) - \int_E g_2(y)\mu_2(dy) + I\big(\mu_2, P^{a_2}(x_2,\cdot)\big) \\
&-c(x_1,a_1) - \int_E g_1(y)\mu_1(dy) + I\big(\mu_1, P^{a_1}(x_1,\cdot)\big) \\
&+c(x_1,a_1) + \int_E g_2(y)\mu_1(dy) - I\big(\mu_1, P^{a_1}(x_1,\cdot)\big) \\
&= \int_E \big(g_1(y) - g_2(y)\big)(\mu_2-\mu_1)(dy) \le \sup_{y\in E}\big(g_1(y) - g_2(y)\big)(\mu_2-\mu_1)(\Gamma) \\
&+ \inf_{y\in E}\big(g_1(y) - g_2(y)\big)(\mu_2-\mu_1)(\Gamma^c) = ||g_1 - g_2||_{sp}(\mu_2-\mu_1)(\Gamma),
\end{aligned}
$$

where the set Γ comes from the Hahn–Jordan decomposition of $\mu_2 - \mu_1$.

Consequently, taking sup over $x_1, x_2 \in E$ in (2.7) we have

$$
||Tg_1 - Tg_2||_{sp} \le ||g_1 - g_2||_{sp} \sup_{B\in\mathcal{E}} \sup_{x,x'\in E} \sup_{a,a'\in U} (\mu_{xag_1} - \mu_{x'a'g_2})(B), \tag{2.8}
$$

where

$$
\mu_{xag}(B) = \frac{\int_B e^{g(z)} P^a(x,dz)}{\int_E e^{g(z)} P^a(x,dz)}.
$$

To complete the proof it then remains to show that

$$
\sup_{g_1,g_2;||g_1||_{sp},||g_2||_{sp}\le M} \sup_{B\in\mathcal{E}} \sup_{x,x'\in E} \sup_{a,a'\in U} (\mu_{xag_1} - \mu_{x'a'g_2})(B) = \alpha(M) < 1. \tag{2.9}
$$

Suppose (2.9) does not hold. Then there exist sequences (g_{1n}), (g_{2n}) with $||g_{1n}||_{sp}$, $||g_{2n}||_{sp} \le M$, (B_n), $B_n \in \mathcal{E}$, (x_n), (x'_n) and (a_n), (a'_n) such that

$$
\left(\mu_{x_n a_n g_{1n}} - \mu_{x'_n a'_n g_{2n}}\right)(B_n) \to 1 \quad \text{as} \quad n \to \infty.
$$

Therefore

$$
\mu_{x_n a_n g_{1n}}(B_n) \to 1
$$

and

$$
\mu_{x'_n a'_n g_{2n}}(B_n) \to 0 \quad \text{as} \quad n \to \infty.
$$

Since

$$e^{-||g||_{sp}} P^a(x, B) \le \mu_{xag}(B),$$

we have

$$P^{a'_n}(x'_n, B_n) \to 0 \quad \text{and} \quad P^{a_n}(x_n, B_n^c) \to 0.$$

Consequently

$$\lim_{n\to\infty} P^{a_n}(x_n, B_n) - P^{a'_n}(x'_n, B_n) = 1,$$

which contradicts assumption (A1). □

We get immediately the following corollary.

COROLLARY 2.3. *Under* (A1) *and* (B1) *there exists at most one (up to an additive constant) function* $w \in C(E)$ *and a unique constant* λ *for which the Bellman equation* (1.2) *is satisfied.* □

The following assumption (A2) will be shown to guarantee that in a suitable subset of $C(E)$ the operator T is in fact a global contraction.

(A2) There exists $\eta \in \mathcal{P}(E)$ and a Borel function $E \times E \times A \ni (x, y, a) \mapsto p(x, y, a)$ such that $\forall_{x\in E}$, $\forall_{a\in U}$, $\forall_{B\in\mathcal{E}}$,

$$P^a(x, B) = \int_B p(x, y, a)\eta(dy)$$

and

$$\sup_{x,x'\in E} \sup_{y\in E} \sup_{a\in U} \frac{p(x, y, a)}{p(x', y, a)} = \mathcal{K} < \infty. \tag{2.10}$$

Remark 2.4. If in addition to the existence of a density $p(x, y, a)$ with respect to a probability measure η we have that

(B2) $E \times U \ni (x, a) \mapsto p(x, \cdot, a) \in L^1(\eta)$ is continuous,

then

(B3) the mapping $E \times U \ni (x, a) \mapsto P^a(x, \cdot)$ is continuous in the variation norm of $\mathcal{P}(E)$ and, in particular, (B1) also holds.

Furthermore, if

(A3) $p(x, y, a)$ in (A2) satisfies the stronger form of (2.10), namely,

$$\sup_{x,x'\in E} \sup_{y\in E} \sup_{a,a'\in U} \frac{p(x, y, a)}{p(x', y, a')} < +\infty,$$

then (A1) holds.

Remark 2.5. The assumptions (A2) and (A3) require mutual equivalence of transition probabilities for different values of initial states and controls and existence of a uniformly bounded away from 0 their Radon–Nikodým derivatives. These kinds

of assumption restrict the class of models in locally compact state space; however, they appear acceptable in the case of compact or countable state spaces.

THEOREM 2.6. *Under* (B1), (A1), *and* (A2) *the operator* T *defined in* (2.3) *and* (2.4) *maps* $C(E)$ *into* $C_L(E)$, *where* $C_L(E) \subset C(E)$ *is the set of continuous functions with span norm bounded by* $L = \log \mathcal{K} + ||c||$, *and consequently there is a unique (up to an additive constant) function* $w \in C(E)$ *which is in fact in* $C_L(E)$ *and a constant* λ *such that Bellman equation* (1.2) *is satisfied.*

Proof. Notice that for $g \in C(E)$ we have

$$Tg(x) - Tg(x') \le \sup_{a\in U}\left[c(x,a) - c(x',a) + \log \frac{\int_E e^{g(y)}P^a(x,dy)}{\int_E e^{g(y)}P^a(x',dy)}\right]$$

$$\le ||c|| + \sup_{a\in U} \log \frac{\int_E e^{g(y)}\frac{p(x,y,a)}{p(x',y,a)}p(x',y,a)\eta(dy)}{\int_E e^{g(y)}p(x',y,a)\eta(dy)} \le ||c|| + \log \mathcal{K}.$$

Therefore $||Tg||_{sp} \le L$ and T is a global contraction in the span norm in $C_L(E)$, so that it has a fixed point w in $C_L(E)$ and w, which is (up to an additive constant) unique, and the constant $\lambda = w - Tw$ are solutions to (1.2). □

Let $b\mathcal{B}(E)$ denote the set of bounded Borel measurable functions on E, and let $\mathcal{U} = \{u : E \to U, \text{ Borel measurable}\}$.

Given $u \in \mathcal{U}$, for $g \in b\mathcal{B}(E)$ define

$$T_u g(x) = c\big(x, u(x)\big) + \log \int_E e^{g(y)}P^{u(x)}(x,dy) \tag{2.11}$$

or in its equivalent form (see [9, Lemma 3.3] or [2, Proposition 2.3])

$$T_u g(x) = \sup_{\mu\in\mathcal{P}(E)}\left[c\big(x,u(x)\big) + \int_E g(y)\mu(dy) - I\big(\mu, P^{u(x)}(x,\cdot)\big)\right], \tag{2.12}$$

where the sup is attained for

$$\mu_u^*(B) = \frac{\int_B e^{g(y)}P^{u(x)}(x,dy)}{\int_E e^{g(y)}P^{u(x)}(x,dy)}.$$

PROPOSITION 2.7. *Under* (A1) *and* (A2) *for any* $u \in \mathcal{U}$ *the operator* T_u *in* (2.11) *and* (2.12) *transforms* $b\mathcal{B}(E)$ *into* $b\mathcal{B}_L(E)$, *where* $b\mathcal{B}_L(E) \subset b\mathcal{B}(E)$ *consists of functions with span norm bounded by* $L = \log \mathcal{K} + ||c||$. *Furthermore* T_u *is a global contraction in* $b\mathcal{B}_L(E)$, *and consequently there exists a unique (up to an additive constant) function* $w_u \in b\mathcal{B}_L(E)$ *and a unique constant* λ_u *such that*

$$w_u(x) + \lambda_u = \left[c\big(x,u(x)\big) + \log \int_E e^{w_u(y)}P^{u(x)}(x,dy)\right]. \tag{2.13}$$

Proof. Let $g_1, g_2 \in b\mathcal{B}(E)$, $x_1, x_2 \in E$. Then

$$\begin{aligned} &T_u g_1(x_2) - T_u g_2(x_2) - \big(T_u g_1(x_1) - T_u g_2(x_1)\big) \\ &= \sup_{\mu\in\mathcal{P}(E)}\left[\int_E g_1(y)\mu(dy) - I\big(\mu, P^{u(x_2)}(x_2,\cdot)\big)\right] \\ &- \sup_{\mu\in\mathcal{P}(E)}\left[\int_E g_2(y)\mu(dy) - I\big(\mu, P^{u(x_2)}(x_2,\cdot)\big)\right] \end{aligned}$$

$$
\begin{aligned}
&- \sup_{\mu\in\mathcal{P}(E)} \left[\int_E g_1(y)\mu(dy) - I\big(\mu, P^{u(x_1)}(x_1,\cdot)\big)\right] \\
&+ \sup_{\mu\in\mathcal{P}(E)} \left[\int_E g_2(y)\mu(dy) - I\big(\mu, P^{u(x_1)}(x_1,\cdot)\big)\right] \\
&\le \int_E (g_1-g_2)(y)\left(\mu_{x_2u(x_2)g_1} - \mu_{x_1u(x_1)g_2}\right)(dy) \\
&\le ||g_1-g_2||_{sp} \sup_{B\in\mathcal{E}} \left(\mu_{x_2u(x_2)g_1} - \mu_{x_1u(x_1)g_2}\right)(B),
\end{aligned}
$$

where $\mu_{xu(x)g}$ is as defined below (2.8).

Similarly to what has been done in the proof of Proposition 2.2, it is possible to show that

$$
\sup_{x_1,x_2\in E} \sup_{B\in\mathcal{E}} \left(\mu_{x_2u(x_2)g_1} - \mu_{x_1u(x_1)g_2}\right)(B) \le \alpha(M) < 1
$$

provided that $||g_1||_{sp}$, $||g_2||_{sp} \le M$.

Furthermore, by arguments similar to those used in the proof of Theorem 2.6, it is easily seen that under (A2) the operator T_u transforms $b\mathcal{B}(E)$ into $b\mathcal{B}_L(E)$ and consequently T_u is a global contraction in the space $b\mathcal{B}_L(E)$. Existence and uniqueness of the solution to (2.13) then follow as in Theorem 2.6. □

Two important consequences of Proposition 2.7 are given below.

COROLLARY 2.8. *Under* (A1) *and* (A2) *for each* $\varepsilon > 0$ *there exists a positive integer* n_0 *such that for* $n \ge n_0$

$$
\sup_{x\in E} \sup_{u\in\mathcal{U}} \left| n^{-1}\log E_x\left\{\exp\left\{\sum_{i=0}^{n-1} c\big(x_i,u(x_i)\big)\right\}\right\} - \lambda_u \right| < \varepsilon. \tag{2.14}
$$

Proof. Recalling that by Proposition 2.7, $||w_u||_{sp} \le L$, and iterating (2.13), namely,

$$
e^{w_u(x)} = E_x\left\{\exp\left\{\sum_{i=0}^{n-1}\Big(c\big(x_i,u(x_i)\big) - \lambda_u\Big)\right\} e^{w_u(x_n)}\right\},
$$

we obtain

$$
e^{-L} \le e^{-||w_u||_{sp}} \le E_x\left\{\exp\left\{\sum_{i=0}^{n-1}\Big(c\big(x_i,u(x_i)\big) - \lambda_u\Big)\right\}\right\} \le e^{||w_u||_{sp}} \le e^L
$$

and therefore

$$
-\frac{L}{n} \le n^{-1}\log E_x\left\{\exp\left\{\sum_{i=0}^{n-1} c\big(x_i,u(x_i)\big)\right\}\right\} - \lambda_u \le \frac{L}{n}
$$

from which the result follows. □

COROLLARY 2.9. *If* E *and* U *are finite, under* (A1) *and* (A2) *the policy iteration algorithm works and can be performed through the following steps:*

(1) *Fix* $z \in E$ *and take any* $u \in \mathcal{U}$.

(2) *Solve* (2.13) *with* $w_u(z) = 0$.

(3) *If for a point* $\overline{x} \in E$ *we have*

$$w_u(\overline{x}) + \lambda_u > \inf_{a \in U} \left[c(\overline{x}, a) + \log \int_E e^{w_u(y)} P^a(\overline{x}, dy) \right], \tag{2.15}$$

where the inf *is attained at* $\overline{a}$*, then put* $\overline{u}(\overline{x}) = \overline{a}$*,* $\overline{u}(x) = u(x)$ *for* $x \neq \overline{x}$*, and go to point* 2 *with* $u = \overline{u}$*. If* $\forall x \in E$ (2.15) *holds as an equality, then* u *is an optimal control.*

Proof. It follows from the fact that for E and U finite the class $\mathcal{U}$ is also finite, taking into account that by (2.13) and (2.15) we have $\forall x \in E$

$$w_u(x) + \lambda_u \geq c\big(x, \overline{u}(x)\big) + \log \int e^{w_u(y)} P^{\overline{u}(x)}(x, dy)$$

from which by interation we obtain

$$w_u(x) + n\lambda_u \geq \log E_x \left\{ \exp \left\{ \sum_{i=0}^{n-1} c\big(x_i, \overline{u}(x_i)\big) \right\} e^{w_u(x_n)} \right\}$$

and consequently $\lambda_{\overline{u}} \leq \lambda_u$. □

Remark 2.10. The previous corollaries provide us with two feasible methods to solve risk-sensitive ergodic problems. In particular, Corollary 2.8 suggests a finite horizon approximation, which is uniform in the class of stationary Markov controls $\mathcal{U}$, for general state and control spaces.

On the other hand, for finite state and control spaces, Corollary 2.9 provides a version of the well-known Howard improvement algorithm (see, e.g., [10]), which in a finite number of steps leads to an optimal control.

3. Discounted cost asymptotics. For $\beta \in (0, 1)$ and $\gamma \in [0, d]$ consider the exponential discounted cost criterion (see, e.g., [2])

$$J^{\beta}_{x,\gamma}\big((a_n)\big) = E_x \left\{ \exp \left\{ \gamma \sum_{i=0}^{\infty} \beta^i c(x_i, a_i) \right\} \right\}. \tag{3.1}$$

Let $w_\beta(x, \gamma)$ be the corresponding value function

$$w_\beta(x, \gamma) = \inf_{(a_n)} E_x \left\{ \exp \left\{ \gamma \sum_{i=0}^{\infty} \beta^i c(x_i, a_i) \right\} \right\}. \tag{3.2}$$

Proposition 3.1. *Under* (B1) *there exists a unique* $w_\beta \in C(E \times [0, d])$ *such that the equation*

$$w_\beta(x, \gamma) = \inf_{a \in U} \left[e^{\gamma c(x,a)} \int_E w_\beta(y, \gamma\beta) P^a(x, dy) \right] \tag{3.3}$$

with boundary condition

$$\lim_{\gamma \to 0} \sup_{x \in E} |w_\beta(x, \gamma) - 1| = 0 \tag{3.4}$$

holds.

Moreover $w_\beta(x, \gamma)$ *is the optimal value for the cost function* (3.1) *defined in* (3.2).

Proof. We split the proof into several steps.

Step 1. Let for $g \in C(E \times [0,d])$

$$T_\beta g(x,\gamma) = \inf_{a\in U}\left[e^{\gamma c(x,a)}\int_E g(y,\gamma\beta)P^a(x,dy)\right]. \tag{3.5}$$

Notice that by (B1) T_β transforms $C(E\times[0,d])$ into itself. Furthermore, if g satisfies the boundary condition

$$\lim_{\gamma\to 0}\sup_{x\in E}|g(x,\gamma)-1| = 0, \tag{3.6}$$

then (3.6) holds also for $T_\beta g(x,\gamma)$ since

$$\begin{aligned}|T_\beta g(x,\gamma)-1| &\le \sup_{a\in U}\left[|e^{\gamma c(x,a)}-1|\int_E |g(y,\gamma\beta)|P^a(x,dy)\right]\\ &+\sup_{a\in U}\left[\int_E |g(y,\gamma\beta)-1|P^a(x,dy)\right]\to 0\end{aligned}$$

uniformly in x as $\gamma\to 0$.

Step 2. Notice that T_β preserves the order in the sense that if $\forall_{x\in E}$, $\forall_{\gamma\in[0,d]}$ $g_1(x,\gamma)\ge g_2(x,\gamma)$, then $T_\beta g_1(x,\gamma)\ge T_\beta g_2(x,\gamma)$.

Therefore since for $\underline{g}(x,\gamma)\equiv 1$, $T_\beta\underline{g}(x,\gamma)\ge 1\equiv\underline{g}(x,\gamma)$ we have that $T_\beta^n\underline{g}(x,\gamma)$ is an increasing sequence. On the other hand, since for $\overline{g}(x,\gamma)=e^{\gamma\frac{||c||}{1-\beta}}$

$$T_\beta\overline{g}(x,\gamma)=\inf_{a\in U}\left[e^{\gamma c(x,a)}e^{\gamma\beta\frac{||c||}{1-\beta}}\right]\le e^{\gamma\frac{||c||}{1-\beta}}=\overline{g}(x,\gamma),$$

we have that $T_\beta^n\overline{g}(x,\gamma)$ is a decreasing sequence and

$$1\le T_\beta^n\underline{g}(x,\gamma)\le T_\beta^n\overline{g}(x,\gamma)\le e^{\gamma\frac{||c||}{1-\beta}}. \tag{3.7}$$

Step 3. For $g\in C(E\times[0,d])$ we have the representation formula for $n=1,2,\dots,$

$$T_\beta^n g(x,\gamma)=\inf_{(a_n)}E_x\left\{\exp\left\{\gamma\sum_{i=0}^{n-1}\beta^i c(x_i,a_i)\right\}g(x_n,\gamma\beta^n)\right\}, \tag{3.8}$$

and therefore uniformly in $x\in E$ and $\gamma\in[0,d]$

$$\begin{aligned}&|T_\beta^n\overline{g}(x,\gamma)-T_\beta^n\underline{g}(x,\gamma)|\\ &\le\sup_{(a_n)}E_x\left\{e^{\gamma\sum_{i=0}^{n-1}\beta^i c(x_i,a_i)}\left|e^{\frac{\gamma\beta^n||c||}{1-\beta}}-1\right|\right\}\to 0\end{aligned}$$

as $n\to\infty$.

Thus there exists a function $w_\beta(x,\gamma)$ which is a uniform (in $x\in E$ and $\gamma\in[0,d]$) limit of $T_\beta^n\overline{g}(x,\gamma)$ and $T_\beta^n\underline{g}(x,\gamma)$. Consequently $w_\beta(x,\gamma)\in C(E\times[0,d])$ and is a solution to (3.3) with boundary condition (3.4) (using Step 1). The uniqueness is guaranteed by the representation (3.8) together with the boundary condition (3.4). □

Remark 3.2. An optimal strategy (a_n) for the cost functional $J^\beta_{x,\gamma}$ in (3.1) is of the form $a_n=u(x_n,\gamma\beta^n)$, where $u: E\times[0,d]\to U$ is a function for which the inf in (3.3) is attained. Notice that the above strategy is not a stationary one since it depends on time n through the power n of β.

We shall need the following assumption.

(A4) There exists $\eta \in \mathcal{P}(E)$ and a positive integer N such that for any Markov strategy $V = (a_n)$, where $a_n = u_n(x_n)$ with $u_n \in \mathcal{U}$ and for any $x \in E$ the measures $P_x^V\{x_N \in \cdot\}$ are absolutely continuous with respect to η with densities $p_N^V(x,y)$ satisfying

$$\sup_V \sup_{x,x' \in E} \sup_{y \in E} \frac{p_N^V(x,y)}{p_N^V(x',y)} \le \overline{\mathcal{K}} < \infty.$$

Clearly (A2) implies (A4) with $N = 1$.

Define for fixed $z \in E$

$$h_\beta(x,\gamma) = \frac{w_\beta(x,\gamma)}{w_\beta(z,\gamma)};$$

then we have the following theorem.

THEOREM 3.3. *Under* (A1), (A4), *and* (B3), *there exists a unique constant* λ *and for fixed* $z \in E$ *a unique function* $w \in C(E)$ *with* $w(z) = 0$ *which satisfy the Bellman equation*

$$w(x) + \lambda = \inf_a \left[\gamma c(x,a) + \log \int_E e^{w(y)} P^a(x,dy)\right]. \tag{3.9}$$

Furthermore, for each $x \in E$ *and each* $\gamma \in [0,d]$,

$$\lim_{\beta \uparrow 1} \frac{w_\beta(x,\gamma)}{w_\beta(x,\gamma\beta)} = e^\lambda,$$
$$\lim_{\beta \uparrow 1} h_\beta(x,\gamma) = e^{w(x)},$$

where the last convergence is uniform on compact sets of E.

Proof. For the optimal Markov strategy $V = (a_n)$ with $a_n = u_n(x_n, \gamma\beta^n)$ (see Remark 2.10) we obtain using (A4)

$$\begin{aligned}
\frac{w_\beta(x,\gamma)}{w_\beta(z,\gamma)} &= \frac{E_x^V\left\{\exp\left\{\gamma \sum_{i=0}^{N-1} c(x_i,a_i)\beta^i\right\} w_\beta(x_N,\gamma\beta^N)\right\}}{E_z^V\left\{\exp\left\{\gamma \sum_{i=0}^{N-1} c(x_i,a_i)\beta^i\right\} w_\beta(x_N,\gamma\beta^N)\right\}} \\
&\le e^{\gamma N||c||} \frac{\int_E w_\beta(y,\gamma\beta^N) p_N^V(z,y) \frac{p_N^V(x,y)}{p_N^V(z,y)} \eta(dy)}{\int_E w_\beta(y,\gamma\beta^N) p_N^V(z,y) \eta(dy)} \\
&\le \overline{\mathcal{K}} e^{\gamma N||c||}.
\end{aligned}$$

Similarly

$$\frac{w_\beta(x,\gamma)}{w_\beta(z,\gamma)} \ge \overline{\mathcal{K}}^{-1} e^{-\gamma N||c||}$$

so that

$$\overline{\mathcal{K}}^{-1} e^{-\gamma N||c||} \le h_\beta(x,\gamma) \le \overline{\mathcal{K}} e^{\gamma N||c||}. \tag{3.10}$$

By (3.3) we then have

$$h_\beta(x,\gamma) \cdot \frac{w_\beta(z,\gamma)}{w_\beta(z,\gamma\beta)} = \inf_{a \in U} \left[e^{\gamma c(x,a)} \int_E h_\beta(y,\gamma\beta) P^a(x,dy)\right]. \tag{3.11}$$

Using (3.10), we obtain from (3.11)

$$1 \le \frac{w_\beta(x,\gamma)}{w_\beta(x,\gamma\beta)} \le e^{\gamma||c||}\frac{||h_\beta(\cdot,\gamma\beta)||}{h_\beta(x,\gamma\beta)}$$
$$\le e^{\gamma||c||}\overline{\mathcal{K}}^2 e^{2\gamma\beta N||c||} \le \overline{\mathcal{K}}^2 e^{\gamma||c||(2N+1)}.$$

By (B3) and (3.10), for each $\gamma \in [0,d]$ and any positive integer m, $T_\beta h_\beta(x,\gamma\beta^m)$ is continuous in x, uniformly in β, and is bounded.

For a fixed γ, using a diagonal procedure, by the Ascoli–Arzelà theorem [12, Theorem 8.33] one can find a sequence $\beta_n \uparrow 1$, constants λ_m, and functions $w_m(x,\gamma)$ such that for $m = 0,1,2,\ldots,$

$$\frac{w_{\beta_n}(z,\gamma\beta_n^m)}{w_{\beta_n}(z,\gamma\beta_n^{m+1})} \to e^{\lambda_m} \tag{3.12}$$

and

$$T_{\beta_n} h_{\beta_n}(x,\gamma\beta_n^m) \to e^{w_m(x,\gamma)+\lambda_m} \tag{3.13}$$

uniformly in x on compact subsets of E.

By (3.11) we then have for $m = 0,1,2,\ldots,$

$$\lim_{n\to\infty} h_{\beta_n}(x,\gamma\beta_n^{m+1}) = e^{w_m(x,\gamma)} \tag{3.14}$$

uniformly in x on compact subsets of E.

Furthermore, from (3.10) and (3.14),

$$-\log\overline{\mathcal{K}} - N\gamma||c|| \le w_m(x,\gamma) \le \log\overline{\mathcal{K}} + N\gamma||c||. \tag{3.15}$$

Consequently, from (3.13) and (3.14), we have

$$e^{w_m(x,\gamma)}e^{\lambda_m} = \lim_{n\to\infty}\left\{\inf_{a\in U}\left[e^{\gamma\beta_n^m c(x,a)}\int_E h_{\beta_n}(y,\gamma\beta_n^{m+1})P^a(x,dy)\right]\right\}$$
$$= \inf_{a\in U}\left[e^{\gamma c(x,a)}\int_E e^{w_m(y,\gamma)}P^a(x,dy)\right]. \tag{3.16}$$

Notice that $w_m(z,\gamma) = 0$ and $||w_m(\cdot,\gamma)||_{sp} \le 2(\log\overline{\mathcal{K}} + N\gamma||c||)$.

Let, for $g \in C(E\times[0,d])$,

$$Tg(x,\gamma) = \inf_{a\in U}\left[\gamma c(x,a) + \log\int_E e^{g(y,\gamma)}P^a(x,dy)\right].$$

From (3.16) it is clear that $Tw_m = w_m + \lambda_m$.

Since $w_m(z,\gamma) = w_{m+1}(z,\gamma) = 0$ by Corollary 2.3, we have $w_m(z,\gamma) = w(z,\gamma)$ for $m = 0,1,2,\ldots,$ and consequently $\lambda_0 = \lambda_1 = \cdots = \lambda_m = \lambda$.

Since by Corollary 2.3, adapting the procedure described above, from any sequence $\beta_n \uparrow 1$ one can choose a further subsequence β_{n_k} (to simplify notation denoted again by β_n) such that

$$\lim_{n\to\infty}\log\frac{w_{\beta_n}(z,\gamma)}{w_{\beta_n}(z,\gamma\beta_n)} = \lambda$$

and

$$\lim_{n\to\infty} \log h_{\beta_n}(x,\gamma) = w(x),$$

where λ and $w(x)$ are unique solutions of the Bellman equation (3.9), then the above convergences hold for any sequence $\beta \uparrow 1$.

Finally, since λ is unique (see Corollary 2.3), it does not depend on the choice of the initial point z as stated in the theorem. □

Remark 3.4. Following the proof of Theorem 3.3, notice that we can obtain the existence of a constant λ and $w \in C(E)$ for which (3.9) is satisfied using only assumptions (A4) and (B3). From Proposition 1.1, λ is then the optimal cost of the risk-sensitive problem (see (1.3)) and is therefore unique.

4. Discounted game approach. For $\beta \in (0,1)$ consider the following equation:

$$e^{w_\beta(x)} = \inf_{a\in U}\left\{e^{c(x,a)}\int_E e^{\beta w_\beta(y)}P^a(x,dy)\right\}. \tag{4.1}$$

The function w_β in (4.1) can be interpreted as the upper value of a certain discounted stochastic dynamic game (see [9] for details). We have the following proposition.

PROPOSITION 4.1. *Under* (B1), *for* $\beta \in (0,1)$ *there exists a unique* $w_\beta \in C(E)$ *for which the equation is satisfied. Moreover*

$$0 \le w_\beta(x) \le \frac{||c||}{1-\beta}. \tag{4.2}$$

Proof. Notice that under (B1) the operator

$$T^\beta g(x) = \inf_{a\in U}\left[c(x,a) + \log\int_E e^{\beta g(y)}P^a(x,dy)\right] \tag{4.3}$$

transforms $C(E)$ into itself. Furthermore, for $g_1, g_2 \in C(E)$,

$$T^\beta g_1(x) - T^\beta g_2(x) \le \sup_{a\in U}\left[\log\frac{\int_E e^{\beta g_1(y)}P^a(x,dy)}{\int_E e^{\beta g_2(y)}P^a(x,dy)}\right]$$
$$\le \sup_{a\in U}\left[\log\frac{\int_E e^{\beta||g_1-g_2||+\beta g_2(y)}P^a(x,dy)}{\int_E e^{\beta g_2(y)}P^a(x,dy)}\right] \le \beta||g_1-g_2||$$

and changing the role of g_1 and g_2 we obtain that

$$||T^\beta g_1 - T^\beta g_2|| \le \beta||g_1-g_2||,$$

which means that T^β is a contraction in $C(E)$. In addition for $\underline{g}(x) \equiv 0$ and $\overline{g}(x) = \frac{||c||}{1-\beta}$ we have $T^\beta\underline{g} \ge \underline{g}$ and $T^\beta\overline{g} \le \overline{g}$ so that by the contraction principle, analogous to what has been done in Proposition 3.1, $(T^\beta)^n\underline{g}$ and $(T^\beta)^n\overline{g}$ approximate w_β from below and from above, respectively, from which (4.2) follows. □

Fix $z \in E$ and let $k_\beta(x) = w_\beta(x) - w_\beta(z)$.

From (4.1) we then have

$$k_\beta(x) + (1-\beta)w_\beta(z) = T^\beta k_\beta(x) \tag{4.4}$$

with the operator T_β defined in (4.3).

THEOREM 4.2. *Under* (A2) *and* (B3) *there exist a function* $w \in C(E)$ *and a unique constant* λ *for which* (1.2) *holds. Moreover*

$$\lim_{\beta\uparrow 1}(1-\beta)w_\beta(x) = \lambda. \tag{4.5}$$

Assuming additionally that (A1) *holds, we have then that* w *is the unique function in* $C(E)$ *such that* $w(z) = 0$ *for which* (1.2) *holds and furthermore*

$$\lim_{\beta\uparrow 1} k_\beta(x) = w(x) \tag{4.6}$$

uniformly on compact subsets of E.

Proof. Let us notice first that similar to the proof of Theorem 2.6, under (A2) and (B3) the operator T^β transforms $C(E)$ into $C_L(E)$. Consequently, since $w_\beta = T_\beta w_\beta$ we have that $||k_\beta|| \le L$. Furthermore, by (B3) the function $T^\beta k_\beta(x)$ is uniformly in $\beta \in (0,1)$ continuous and bounded.

Therefore, taking into account (4.2), there exist $\lambda \in R$ and $w \in C(E)$ and a sequence $\beta_n \uparrow 1$ such that

$$\lim_{n\to\infty}(1-\beta_n)w_{\beta_n}(z) = \lambda$$

and

$$\lim_{n\to\infty} T^{\beta_n}k_{\beta_n}(x) = w(x) + \lambda$$

uniformly in x from compact subsets of E.

Then from (4.4)

$$\lim_{n\to\infty} k_{\beta_n}(x) = w(x)$$

uniformly on compact subsets of E and consequently from (4.3) and (2.3)

$$\lim_{n\to\infty} T^{\beta_n}k_{\beta_n}(x) = Tw(x)$$

so that, letting $\beta_n \uparrow 1$ in (4.4) we have that $w(x)$ and λ satisfy (1.2).

The uniqueness of λ as well as the convergence (4.5) for all $x \in E$ is guaranteed by Proposition 1.1. The uniqueness of $w(x)$ is guaranteed by Corollary 2.3 and therefore we have (4.6). □

5. Risk-sensitive asymptotics. In this section we shall study the asymptotics of the value functions corresponding to the cost functionals

$$J_{x,\gamma}\big((a_n)\big) = \gamma^{-1}\limsup_{n\to\infty} n^{-1}\log\left(E_x\left\{\exp\left\{\sum_{i=0}^{n-1}\gamma c(x_i,a_i)\right\}\right\}\right) \tag{5.1}$$

as the risk factor $\gamma \downarrow 0$, assuming that E is a locally compact separable metric space.

Notice that by the Jensen inequality we have

$$J_{x,\gamma}\big((a_n)\big) \ge \limsup_{n\to\infty} n^{-1}E_x\left\{\sum_{i=0}^{n-1} c(x_i,a_i)\right\} := \overline{J}_x\big((a_n)\big). \tag{5.2}$$

Furthermore, again by the Jensen inequality, for $\gamma_1 \le \gamma_2$ we have

$$J_{x,\gamma_1}\big((a_n)\big) \le J_{x,\gamma_2}\big((a_n)\big), \tag{5.3}$$

which means that the optimal values of the cost $J_{x,\gamma}$ are decreasing as $\gamma \to 0$ and greater than the optimal value of $\overline{J}_x$. We shall in fact show that the latter value is reached in the limit.

We impose the following assumption.

(A5) There exists $\eta \in \mathcal{P}(E)$ such that for $x \in E$, $a \in U$, $B \in \mathcal{E}$

$$P^a(x,B) = \int_B p(x,y,a)\eta(dy),$$

where $p(x,y,a) > 0$, the mapping $(x,y,a) \to p(x,y,a)$ is continuous and (A3) holds, i.e.,

$$\sup_{x,x' \in E} \sup_{y \in E} \sup_{a,a' \in U} \frac{p(x,y,a)}{p(x',y,a')} = \tilde{\mathcal{K}} < +\infty.$$

For the proof of the main theorem we shall need the following result.

LEMMA 5.1. *Under* (A5) *we have that*

$$\overline{\lambda} = \inf_{u \in \mathcal{U}_c} \overline{J}_x\big(u(x_n)\big), \tag{5.4}$$

where $\mathcal{U}_c$ *is the subclass of* $\mathcal{U}$ *consisting of continuous functions* $u : E \to U$, *and* $\overline{\lambda} = \inf_{(a_n)} \overline{J}_x\big((a_n)\big)$.

Proof. Since E is locally compact there is an increasing sequence of compact sets K_m such that $\bigcup_{m=1}^{\infty} K_m = E$. We can also require that $\eta(\partial K_m) = 0$. Consider now a sequence of partitions $(E_1^m, E_2^m, \dots, E_{d_m}^m)$, $m = 1,2,\dots$, of E with representative elements $\{e_1^m, e_2^m, \dots, e_{d_m}^m\}$ such that $E = \bigcup_{i=1}^{d_m} E_i^m$, $E_i^m \cap E_j^m = \emptyset$ for $i \neq j$, $\eta(\partial E_i^m) = 0$ the diameter of E_i^m is not greater than $\frac{1}{m}$ for $i = 1,2,\dots,d_m - 1$, $E_{d_m}^m = E \setminus K_m$; furthermore $\{E_1^{m+1}, E_2^{m+1}, \dots, E_{d_{m+1}}^{m+1}\}$ is a partition finer than $\{E_1^m, E_2^m, \dots, E_{d_m}^m\}$ and $\{e_1^m, e_2^m, \dots, e_{d_m}^m\} \subset \{e_1^{m+1}, \dots, e_{d_{m+1}}^{m+1}\}$.

Let $P_m^a(x,\cdot) = P^a(e_l^m,\cdot)$, for $x \in E_l^m$, $l = 1,2,\dots,d_m, m = 1,2,\dots$

Consider a Markov process $X^m = (x_n^m)$ with transition kernel $P_m^a(x,\cdot)$. For $u \in \mathcal{U}$ define

$$\lambda_m^u = \limsup_{n\to\infty} n^{-1} E_x^u \left\{ \sum_{i=0}^{n-1} c((x_i^m), u(x_i^m)) \right\}.$$

Similarly for a Markov process $X = (x_n)$ with transition kernel $P^a(x,\cdot)$ let

$$\lambda^u = \limsup_{n\to\infty} n^{-1} E_x^u \left\{ \sum_{i=0}^{n-1} c((x_i), u(x_i)) \right\}.$$

By (A5), using the ergodicity results of section 3.3 of [10] we obtain that there are a positive integer N and a constant M such that for $n \geq N$ and $m = 1,2,\dots$ we have

$$\sup_u \sup_{x \in E} \left| \lambda_m^u - n^{-1} E_x^u \left\{ \sum_{i=0}^{n-1} c((x_i^m), u(x_i^m)) \right\} \right| \leq \frac{M}{n}$$

and

$$\sup_u \sup_{x\in E} \left| \lambda^u - n^{-1} E_x^u \left\{ \sum_{i=0}^{n-1} c((x_i), u(x_i)) \right\} \right| \le \frac{M}{n} \tag{5.5}$$

with supremum taken over all $u \in \mathcal{U}$.

Since

$$\sup_{a\in U} \; ||P_m^a(x,\cdot) - P^a(x,\cdot)||_{var} \to 0$$

as $m \to \infty$ uniformly in x on compact subsets, where $||\ ||_{var}$ stands for variation norm, we have that

$$\sup_u |\lambda_m^u - \lambda^u| \to 0$$

as $m \to \infty$.

For each controlled Markov process X^m piecewise constant controls (i.e., the controls constant on each set of the partition $E_1^m, \dots, E_{d_m}^m$) are optimal. Therefore the class of piecewise constant controls is also nearly optimal for the controlled Markov process X. Consequently to complete the proof of Lemma 5.1 it remains to show that each piecewise constant (on our partition) function u can be approximated by continuous functions $u_l : E \to U$ in the sense that

$$|\lambda^{u_l} - \lambda^u| \to 0.$$

Choose a sequence u_l of continuous functions from E into U such that

$$\eta(\{z \in E : u_l(z) \neq u(z)\}) \to 0$$

as $l \to \infty$.

Therefore

$$\eta\{z \in E : \exists_{y\in E}\ p(z,y,u_l(z)) \neq p(z,y,u(z))\} \to 0$$

as $l \to \infty$, or in other words,

$$\eta\{z \in E :\ ||P^{u_l(z)}(z,\cdot) - P^{u(z)}(z,\cdot)||_{var} > 0\} \to 0$$

as $l \to \infty$.

Consequently for each n

$$\left| E_x^{u_l} \left\{ \sum_{i=0}^{n-1} c(x_i, u_l(x_i)) \right\} - E_x^u \left\{ \sum_{i=0}^{n-1} c(x_i, u(x_i)) \right\} \right| \to 0$$

as $l \to \infty$ for η almost all $x \in E$.

By (5.5) (as above) we obtain then that $|\lambda^{u_l} - \lambda^u| \to 0$, as $l \to \infty$, which completes the proof. □

The main result of this section can be stated as follows.

THEOREM 5.2. *Under* (A5) *we have*

$$\lim_{\gamma\downarrow 0} \inf_{(a_n)} \sup_{x\in E} J_{x,\gamma}\big((a_n)\big) = \overline{\lambda}. \tag{5.6}$$

Proof. Notice first that by (5.2), for $\gamma > 0$,

$$\inf_{(a_n)} \sup_{x\in E} J_{x,\gamma}\big((a_n)\big) \geq \overline{\lambda}$$

so that, using also the monotonicity property (5.3),

$$\lim_{\gamma\downarrow 0} \inf_{(a_n)} \sup_{x\in E} J_{x,\gamma}\big((a_n)\big) \geq \overline{\lambda}. \tag{5.7}$$

By Theorem 6.9 and Corollary 7.21 of [14] we have that for $u \in \mathcal{U}_c$

$$\begin{aligned} &\lim_{n\to\infty} \sup_{x\in E} \frac{1}{n}\frac{1}{\gamma} \log E\left\{\exp\left\{\sum_{i=0}^{n-1} \gamma c\big(x_i, u(x_i)\big)\right\}\right\} \\ &\leq \sup_{\nu\in\mathcal{P}(E)} \left[\int_E c\big(z, u(z)\big)\nu(dz) - \frac{1}{\gamma} I^u(\nu)\right], \end{aligned} \tag{5.8}$$

where

$$I^u(\nu) = -\inf_{h\in H} \int_E \log \frac{P^u h(x)}{h(x)} \nu(dx),$$

$P^u h(x) = \int h(y) P^{u(x)}(x, dy)$, and H is the set of all bounded functions $h : E \to R^+$ such that $\frac{1}{h(x)}$ is also bounded.

Since I^u is lower semicontinuous, if E is compact there is for each γ a $\nu_\gamma \in \mathcal{P}(E)$ such that

$$\begin{aligned} &\sup_{\nu\in\mathcal{P}(E)} \left[\int_E c\big(z, u(z)\big)\nu(dz) - \frac{1}{\gamma} I^u(\nu)\right] \\ &= \int_E c\big(z, u(z)\big)\nu_\gamma(dz) - \frac{1}{\gamma} I^u(\nu_\gamma). \end{aligned} \tag{5.9}$$

If E is compact, there is a sequence $\gamma_m \to 0$ and a measure $\overline{\nu} \in P(E)$ such that $\nu_{\gamma_m} \to \overline{\nu}$ weakly. Since

$$\frac{1}{n}\frac{1}{\gamma}\left|\log E\left\{\exp\left\{\sum_{i=0}^{n-1} \gamma c\big(x_i, u(x_i)\big)\right\}\right\}\right| \leq ||c||,$$

by (5.8) and (5.9) we have that $\limsup_{m\to\infty} I^u(\nu_{\gamma_m}) = 0$.

By lower semicontinuity of I^u we then have that $\liminf_{m\to\infty} I^u(\nu_{\gamma_m}) \geq I^u(\overline{\nu})$ and therefore $I^u(\overline{\nu}) = 0$.

By Lemma 2.5 of [4], $I^u(\overline{\nu}) = 0$ if and only if $\overline{\nu} = \pi_u$, the unique invariant measure corresponding to the transition operator $P^{u(x)}(x, dy)$.

Consequently

$$\begin{aligned} &\lim_{m\to\infty} \left[\int_E c\big(z, u(z)\big)\nu_{\gamma_m}(dz) - \frac{1}{\gamma_m} I^u(\nu_{\gamma_m})\right] \\ &\leq \int_E c\big(z, u(z)\big)\pi_u(dz) = \overline{J}_x\big(u(x_n)\big) \end{aligned}$$

and taking into account (5.2) and (5.8) we have that

$$\lim_{m\to\infty}\lim_{n\to\infty}\sup_{x\in E}\frac{1}{n}\frac{1}{\gamma_m}\log E\left\{\exp\left\{\sum_{i=0}^{n-1}\gamma_m c\big(x_i,u(x_i)\big)\right\}\right\}$$

(5.10)

$$=\lim_{m\to\infty}\left[\int_E c\big(z,u(z)\big)\nu_{\gamma_m}(dz)-\frac{1}{\gamma_m}I^u(\nu_{\gamma_m})\right]=\overline{J}_x\big(u(x_n)\big).$$

Since from any sequence $\gamma_m\downarrow 0$ one can choose a further subsequence $\gamma_{m_k}\downarrow 0$ such that we have the convergence result (5.10) with γ_m replaced by γ_{m_k}, we finally have

$$\lim_{m\to\infty}\lim_{n\to\infty}\sup_{x\in E}\frac{1}{n}\frac{1}{\gamma_m}\log E\left\{\exp\left\{\sum_{i=0}^{n-1}\gamma_m c\big(x_i,u(x_i)\big)\right\}\right\}=\overline{J}_x\big(u(x_n)\big)$$

from which, by (5.7) and (5.8), we obtain (5.6) in the case of compact E.

In the case when E is only locally compact we construct a continuous function $\psi: E\to R$, $\psi(x)\geq 1$ for $x\in E$ such that the mapping

$$E\ni x\mapsto\int_E\psi(y)P^{u(x)}(x,dy)$$

is bounded on compact subsets of E and for each $m>0$ the set

$$K_m^u=\left\{x\in E:\frac{\psi(x)}{\int_E\psi(y)P^{u(x)}(x,dy)}\leq m\right\}$$

is compact.

Indeed, by local compactness of E for fixed $\overline{x}\in E$ and $\overline{a}\in U$ one can find an increasing sequence of compact subsets $K_n\subset K_{n+1}$, $\partial K_n\subset \mathrm{Int}K_{n+1}$ such that

$$\int_{K_n^c}p(\overline{x},y,\overline{a})\eta(dy)\leq\frac{1}{(n+1)^3}$$

and a continuous function ψ taking values from the interval $[n,n+1]$, for $x\in K_{n+1}\setminus K_n$. Then by (A5) for $x\in E$ and $u\in\mathcal{U}$ we have

$$0<\frac{1}{\tilde{\mathcal{K}}}E_{\overline{x}}^{\overline{a}}\{\psi(x_q)\}\leq E_x^u\{\psi(x_1)\}$$

$$\leq\tilde{\mathcal{K}}E_{\overline{x}}^{\overline{a}}\{\psi(x_1)\}\leq\tilde{\mathcal{K}}\sum_{n=1}^{\infty}(n+1)\frac{1}{(n+1)^3}<\infty$$

and therefore K_m^u is a compact set.

Consequently, by Lemma 4.2 of [5], for each $m>0$ the set

$$C_m^u=\{\nu\in\mathcal{P}(E):I^u(\nu)\leq m\}$$

is compact in $\mathcal{P}(E)$ and we can adapt the proof of the case when E is compact. □

REFERENCES

[1] R. Cavazos-Cadena and E. Fernández-Gaucherand, *Controlled Markov chains with risk-sensitive criteria: Average costs, optimality equations, and optimal solutions*, Math. Methods Oper. Res., 49 (1999), pp. 299–324.

[2] K. J. Chung and M. J. Sobel, *Discounted MDPs: Distribution functions and exponential utility maximization*, SIAM J. Control Optim., 25 (1987), pp. 49–62.

[3] P. Dai Pra, L. Meneghini, and W. Runggaldier, *Connections between stochastic control and dynamic games*, Math. Control Signals Systems, 9 (1996), pp. 303–326.

[4] M. D. Donsker and S. R. S. Varadhan, *Asymptotic evaluation of certain Markov process expectations for large time* I, Comm. Pure Appl. Math, 28 (1975), pp. 1–47.

[5] M. D. Donsker and S. R. S. Varadhan, *Asymptotic evaluation of certain Markov process expectations for large time* III, Comm. Pure Appl. Math, 29 (1977), pp. 389–461.

[6] W. H. Fleming and D. Hernández-Hernández, *Risk-sensitive control of finite state machines on an infinite horizon* I, SIAM J. Control Optim., 35 (1997), pp. 1790–1810.

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

[8] K. Glover and J. C. Doyle, *State space formulae for all stabilizing controllers that satisfy an H^∞-norm bound and relation to risk sensitivity*, Systems Control Lett., 11 (1986), pp. 167–172.

[9] D. Hernandez-Hernandez and S. J. Marcus, *Risk sensitive control of Markov processes in countable state space*, Systems Control Lett., 29 (1996), pp. 147–155.

[10] O. Hernandez-Lerma, *Adaptive Control Processes*, Springer-Verlag, New York, 1989.

[11] H. Nagai, *Bellman equations of risk sensitive control*, SIAM J. Control Optim., 34 (1996), pp. 74–101.

[12] H. L. Royden, *Real Analysis*, MacMillan, New York, 1968.

[13] T. Runolfsson, *Stationary risk-sensitive LQG control and its relation to LQG and H-infinity control*, in Proceedings, 29th CDC Conference, Honolulu, HI, 1990, pp. 1018–1023.

[14] D. W. Stroock, *An Introduction to the Theory of Large Deviations*, Springer-Verlag, New York, 1984.

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

SIAM J. CONTROL OPTIM.
Vol. 38, No. 1, pp. 79–93

SAMPLE-PATH OPTIMALITY AND VARIANCE-MINIMIZATION OF AVERAGE COST MARKOV CONTROL PROCESSES*

ONÉSIMO HERNÁNDEZ-LERMA†, OSCAR VEGA-AMAYA‡, AND GUADALUPE CARRASCO§

Abstract. This paper studies several average-cost criteria for Markov control processes on *Borel spaces* with possibly *unbounded costs*. Under suitable hypotheses we show (i) the existence of a *sample-path average cost* (SPAC-) optimal stationary policy; (ii) a stationary policy is SPAC-optimal if and only if it is *expected average cost* (EAC-) optimal; and (iii) within the class of stationary SPAC-optimal (equivalently, EAC-optimal) policies there exists one with a minimal limiting *average variance*.

Key words. (discrete-time) Markov control processes, average cost criteria, sample-path average cost, expected average cost, canonical policies, average variance

AMS subject classifications. 93E20, 90C40

PII. S0363012998340673

1. Introduction.

We study several average cost (AC) criteria for Markov control processes on *Borel spaces* with possibly *unbounded costs*. Under suitable hypotheses, we show (i) the existence of a *sample-path average cost* (SPAC-) optimal stationary policy; (ii) a stationary policy is SPAC-optimal if and only if it is *expected average cost* (EAC-) optimal; (iii) within the class of stationary SPAC-optimal policies there exists one with minimal limiting *average variance*.

Discrete-time Markov control processes (MCPs) with AC criteria are among the most widely studied stochastic control problems. However, as can be seen in [1, 2, 4, 7, 8, 15, 16, 23, 32, 33, 34] and their extensive bibliographies, most of the related literature is concentrated on the EAC criterion, especially for MCPs with denumerable state space and/or bounded costs. In contrast, for the SPAC case there are a lot fewer works. For instance, for *finite state* MCPs, we should mention the pioneering works by Mandl [28, 29, 30]; see also [23, pp. 161–162]. For the *countable state* case, see [3, 4, 5], whereas for MCPs on *Borel spaces* we know only of [1, Theorem 6.3(v), (vi)] for bounded costs and [26, 35, 36] for unbounded costs. Finally, to the best of our knowledge, for the variance minimization the only previous works are those by Mandl [28, 29, 30] for *finite state* MCPs and by Kurano [24] for the Borel case with *bounded costs*. The reader should be warned, however, that several authors have studied a "variance minimization problem" for a quantity which is not the "real" variance—see, for instance, [21] and [32, p. 408]. Here, our definition (12) [see also (45)] of limiting average variance is the one used in the *central limit theorem for Markov chains*, as in [7, Corollaire 7.III.2] or [31, Theorem 17.0.1].

*Received by the editors June 17, 1998; accepted for publication (in revised form) May 5, 1999; published electronically December 3, 1999. This research was partially supported by the Consejo Nacional de Ciencia y Tecnología (CONACYT, México) grants 3115P-E9608 and 28309E.

http://www.siam.org/journals/sicon/38-1/34067.html

†Departamento de Matemáticas, CINVESTAV-IPN, Apartado Postal 14-740, México D.F. 07000, México (ohernand@math.cinvestav.mx).

‡Departamento de Matemáticas, Universidad de Sonora, Blvd. Transversal y Rosales s/n, Hermosillo, Sonora, México (ovega@gauss.mat.uson.mx).

§Departamento de Matemáticas, Facultad de Ciencias, UNAM, Ciudad Universitaria, México D.F. 04510, México (gcl@hp.fciencias.unam.mx).

To obtain the results (i), (ii), (iii) mentioned in the first paragraph, we introduce hypotheses (Assumptions 3.1, 3.2, 3.4) already used in [18]—see also [36]—to study several undiscounted cost criteria, including EAC-optimality. It turns out that the results (i), (ii), (iii)—stated precisely in Theorems 3.7 and 3.8—can be obtained by adding a mild "second order" condition, Assumption 3.6. This setting combines several tools and proof techniques of common use in stochastic control and probability theory. For example, the "Lyapunov condition" in Assumption 3.2(b) is a suitable variant of conditions previously used by many authors, for instance, [5, 6, 7, 9, 12, 13, 17, 18, 19, 20, 22, 23, 25, 31]. Moreover, the "Lyapunov function" W in Assumptions 3.1 and 3.2 is also required to act as a "weight" (or "bounding") function, which allows us to introduce appropriate weighted norms for functions and measures, as in (14) and (58). The use of weight functions in stochastic control problems goes back (at least) to the 1970s—for earlier references see, for example, [6, 12, 13, 16, 17, 20]. On the other hand, a key step in our proof of SPAC-optimality (see Lemma 4.4) uses the law of large numbers for martingales, also known as the martingale stability theorem [7, 14, 23, 27], which is in the spirit of the proof in [1, Theorem 6.3; 24, 28, 29, 30]. This is different from the "tightness" approach followed by other authors, in which an important intermediate step is to give conditions for a certain set, say M, of occupation (or empirical) measures to be tight. To make sure that M is tight, one can, for instance, compactify the underlying space(s), as in [3, 4] (see also [5]) or impose conditions on the cost function, as in [26, 35, 36]. This has also been done in other works for different kinds of control problems—see, for example, the references in [10, 11, 16, p. 122], where the tightness approach is referred to as the "direct approach."

The remainder of the paper is organized as follows. In section 2 we introduce the Markov control model and the AC criteria we are interested in. In section 3 we introduce our hypotheses and state our main results, Theorems 3.7 and 3.8, which are proved in section 4 and section 5, respectively. We close in section 6 with some important remarks. In particular, in Remark 6.1 we mention hypotheses that can be used in lieu of the "stability" Assumptions 3.2 and 3.4.

2. The Markov control processes. The material in this section is quite standard—see [1, 4, 8, 15, 16, 32, 33, 34].

We shall consider the usual discrete-time, stationary, Markov control model $(\mathbf{X}, \mathbf{A}, \{A(x) : x \in \mathbf{X}\}, Q, C)$ with state space $\mathbf{X}$ and control (or action) set $\mathbf{A}$, both assumed to be Borel spaces with σ-algebras $\mathcal{B}(\mathbf{X})$ and $\mathcal{B}(\mathbf{A})$, respectively. For each state $x \in \mathbf{X}$, $A(x) \in \mathcal{B}(\mathbf{A})$ is the (nonempty) set of feasible control actions in x. We assume that the set

$$\mathbf{K} := \{(x, a) : x \in \mathbf{X}, a \in A(x)\}$$

of feasible state-action pairs is a Borel subset of $\mathbf{X} \times \mathbf{A}$. Finally, Q [or $Q(B|x, a)$ for $B \in \mathcal{B}(\mathbf{X})$ and $(x, a) \in \mathbf{K}$] denotes the transition law, and $C : \mathbf{K} \to \mathbb{R}$ is a measurable function that stands for the cost-per-stage function.

The class of measurable functions $f : \mathbf{X} \to \mathbf{A}$ such that $f(x)$ is in $A(x)$ for every $x \in \mathbf{X}$ is denoted by $\mathbf{F}$, and we suppose that it is nonempty.

Control policies. Let $\mathbf{H}_0 := \mathbf{X}$ and $\mathbf{H}_n := \mathbf{K}^n \times \mathbf{X}$ for $n = 1, 2, \ldots$. A *control policy* is a sequence $\pi = \{\pi_n\}$ of stochastic kernels on $\mathbf{A}$ given $\mathbf{H}_n$ satisfying the constraint $\pi_n(A(x_n)|h_n) = 1$ for every "history" $h_n = (x_0, a_0, \ldots, x_{n-1}, a_{n-1}, x_n)$ in $\mathbf{H}_n$, and $n = 0, 1, \ldots$. The class of all policies is denoted by Π.

A policy $\pi = \{\pi_n\}$ is said to be a (deterministic) *stationary policy* if there exists $f \in \mathbf{F}$ such that $\pi_n(\cdot|h_n)$ is concentrated at $f(x_n)$ for each history $h_n \in \mathbf{H}_n$ and $n =$

$0, 1, \ldots$. Following a standard convention, we identify $\mathbf{F}$ with the class of stationary policies.

For notational ease, for a stationary policy $f \in \mathbf{F}$ we write

$$C_f(x) := C(x, f(x)) \quad \text{and} \quad Q_f(\cdot|x) := Q(\cdot|x, f(x)) \quad \forall x \in \mathbf{X}. \tag{1}$$

Let $(\Omega, \mathcal{F})$ be the (canonical) measurable space consisting of the sample space $\Omega := (\mathbf{X} \times \mathbf{A})^\infty$ and its product σ-algebra $\mathcal{F}$. Then, for each policy π and "initial state" $x \in \mathbf{X}$, a stochastic process $\{(x_n, a_n)\}$ and a probability measure P_x^π are defined on $(\Omega, \mathcal{F})$ in a canonical way, where x_n and a_n represent the state and control at time n, $n = 0, 1, \ldots$. The expectation operator with respect to P_x^π is denoted by E_x^π.

AC criteria. For each $n = 1, 2, \ldots$, let

$$J_n^0(\pi, x) := \sum_{t=0}^{n-1} C(x_t, a_t) \tag{2}$$

be the *n-stage sample-path cost* when using the policy π, given the initial state $x \in \mathbf{X}$. The long-run SPAC is then defined as

$$J^0(\pi, x) := \limsup_{n\to\infty} \frac{1}{n} J_n^0(\pi, x). \tag{3}$$

DEFINITION 2.1. *A policy $\pi^* \in \Pi$ is said to be SPAC-optimal if there exists a constant $\widehat{\rho}$ such that*

$$J^0(\pi^*, x) = \widehat{\rho} \quad P_x^{\pi^*} - a.s. \quad \forall x \in \mathbf{X},$$

and

$$J^0(\pi, x) \geq \widehat{\rho} \quad P_x^\pi - a.s. \quad \forall \pi \in \Pi, x \in \mathbf{X}.$$

The constant $\widehat{\rho}$ is called the optimal SPAC.

The "expected" analogs of (2) and (3) are, respectively, the *n-stage expected cost*

$$J_n(\pi, x) := E_x^\pi \sum_{t=0}^{n-1} C(x_t, a_t)$$

and the long-run EAC

$$J(\pi, x) := \limsup_{n\to\infty} \frac{1}{n} J_n(\pi, x). \tag{4}$$

Among the EAC-related optimality concepts we are interested in are the following.

DEFINITION 2.2. (a) *A policy $\pi^* \in \Pi$ is said to be EAC-optimal if*

$$J(\pi^*, x) = J^*(x) \quad \forall x \in \mathbf{X}, \tag{5}$$

where

$$J^*(x) := \inf_{\pi\in\Pi} J(\pi, x) \quad \textit{for} \quad x \in \mathbf{X} \tag{6}$$

is the optimal expected average cost.

(b) *A stationary policy $f_* \in \mathbf{F}$ is called canonical (or, more explicitly, EAC-canonical) if there exist a constant ρ_* and a measurable function $h_* : \mathbf{X} \to \mathbb{R}$ such that*

$$\rho_* + h_*(x) = \min_{a \in A(x)} \left[C(x,a) + \int_{\mathbf{X}} h_*(y) Q(dy|x,a) \right] \quad \forall x \in \mathbf{X}, \tag{7}$$

and $f_*(x) \in A(x)$ attains the minimum on the right-hand side of (7) *for every $x \in \mathbf{X}$, that is (using the notation* (1)*),*

$$\rho_* + h_*(x) = C_{f_*}(x) + \int_{\mathbf{X}} h_*(y) Q_{f_*}(dy|x) \quad \forall x \in \mathbf{X}. \tag{8}$$

If (7) and (8) are satisfied, it is then said that (ρ_*, h_*, f_*) is a *canonical triplet*, a concept introduced by Yushkevich [37] (see also [1, 8, 12, 16, 18]). The connection between parts (a) and (b) in Definition 2.2 is that if (ρ_*, h_*, f_*) is a canonical triplet and, in addition, h_* satisfies that

$$\lim_{n \to \infty} \frac{1}{n} E_x^\pi h_*(x_n) = 0 \quad \forall \pi \in \Pi, x \in \mathbf{X}, \tag{9}$$

then f_* *is EAC-optimal and ρ_* is the optimal expected average cost*, i.e.,

$$J(f_*, x) = J^*(x) = \rho_* \quad \forall x \in \mathbf{X}. \tag{10}$$

Thus, in this case we have

$$\mathbf{F}_{cp} \subset \mathbf{F}_{eac}, \tag{11}$$

where $\mathbf{F}_{cp}$ is the class of canonical policies and $\mathbf{F}_{eac} \subset \mathbf{F}$ is the class of stationary EAC-optimal policies.

In the following sections we show, among other things, the existence of stationary policies that are optimal in the sense of Definitions 2.1 and 2.2, and, moreover, the existence of a stationary policy that minimizes the limiting average variance in the class $\mathbf{F}_{eac}$. In other words, for each $f \in \mathbf{F}$ and $x \in \mathbf{X}$, consider the *n-stage variance*

$$V_n(f,x) := \text{var}\ \left[J_n^0(f,x) \right] = E_x^f \left[J_n^0(f,x) - J_n(f,x) \right]^2$$

and the limiting *average variance*

$$V(f,x) := \limsup_{n \to \infty} \frac{1}{n} V_n(f,x). \tag{12}$$

Then we shall prove that there exists a stationary policy $\widehat{f}$ such that $\widehat{f}$ is EAC-optimal and

$$V(\widehat{f}, x) = \inf_{f \in \mathbf{F}_{eac}} V(f,x) \quad \forall x \in \mathbf{X}. \tag{13}$$

3. Main results. We first require two sets of hypotheses. The first one, Assumption 3.1, is a combination of the usual continuity/compactness requirements (to ensure, for instance, the existence of "measurable minimizers") together with a growth condition on the one-step cost C.

ASSUMPTION 3.1. *For each state $x \in \mathbf{X}$,*

(a) $A(x)$ *is a compact subset of* $\mathbf{A}$*;*

(b) $C(x,\cdot)$ *is lower semicontinuous on* $A(x)$*;*

(c) $\int_{\mathbf{X}} u(y)Q(dy|x,\cdot)$ *is continuous on* $A(x)$ *for every bounded measurable function* u *on* $\mathbf{X}$;

(d) *there exists a measurable function* $W \geq 1$ *on* $\mathbf{X}$ *and a constant* r_1 *such that*

(d1) $|C(x,a)| \leq r_1 W(x) \quad \forall (x,a) \in \mathbf{K}$ *and*

(d2) $\int_{\mathbf{X}} W(y)Q(dy|x,\cdot)$ *is continuous on* $A(x)$.

The second set of hypotheses we need is to guarantee that the MCP has a nice "stable" behavior uniformly in $\mathbf{F}$. Here, to fix ideas we shall impose Assumptions 3.2 and 3.4 below, which are an adaptation to MCPs of a Lyapunov-like condition used in Markov chain theory—see [9] or [31, p. 367]. However, the reader should keep in mind that, as noted in Remark 6.1, there are other hypotheses that yield the same stable behavior.

ASSUMPTION 3.2. *For each stationary policy* $f \in \mathbf{F}$,

(a) *there exists positive constants* $B_f < 1$ *and* $b_f < \infty$, *and a petite subset* K_f *of* $\mathbf{X}$ *such that (using the notation* (1)*)*

$$\int_{\mathbf{X}} W(y)Q_f(dy|x) \leq B_f W(x) + b_f \mathbf{I}_{K_f}(x) \quad \forall x \in \mathbf{X},$$

where $W \geq 1$ *is the function in Assumption* 3.1(d), *and* $\mathbf{I}_K(\cdot)$ *denotes the indicator function of* K;

(b) *the state processes* $\{x_n\}$*—which under* $f \in \mathbf{F}$ *is a Markov chain with transition kernel* Q_f [16, *Proposition* 2.3.5]*—is* φ*-irreducible and aperiodic for some* σ*-finite measure* φ *on* $\mathbf{X}$.

To state some consequences of Assumption 3.2, let us first introduce the following notation: $B_W(\mathbf{X})$ denotes the normed linear space of measurable functions u on $\mathbf{X}$ with a finite W*-norm* $||u||_W$, which is defined as

$$||u||_W := \sup_{x \in \mathbf{X}} |u(x)|/W(x). \tag{14}$$

We shall write $\int_{\mathbf{X}} u(y)\mu(dy)$ as $\mu(u)$, i.e.,

$$\mu(u) := \int_{\mathbf{X}} u(y)\mu(dy).$$

REMARK 3.3 (see [31, Theorem 16.0.1]). *Under Assumption* 3.2, *for each stationary policy* $f \in \mathbf{F}$ *we have the following:*

(a) *The Markov chain* $\{x_n\}$ *induced by* f *is positive Harris-recurrent, and, moreover, its unique invariant probability measure* μ_f *satisfies that* $\mu_f(W) < \infty$;

(b) $\{x_n\}$ *is* W*-geometrically ergodic; that is, there exist positive constants* $\gamma < 1$ *and* $M_f < \infty$ *such that*

$$\left| \int_{\mathbf{X}} u(y)Q_f^n(dy|x) - \mu_f(u) \right| \leq ||u||_W M_f \gamma_f^n W(x) \tag{15}$$

for each $u \in B_W(\mathbf{X}), x \in \mathbf{X}$, *and* $n = 0, 1, \ldots$.

The next assumption concerns the constants M_f and γ_f in (15).

ASSUMPTION 3.4. $M := \sup_f M_f$ *and* $\gamma := \sup_f \gamma_f$ *are such that* $M < \infty$ *and* $\gamma < 1$.

Assumptions 3.1, 3.2, and 3.4 have been used in [18] and [36] to study several undiscounted cost criteria, such as overtaking optimality, bias optimality, and others. In particular, the following result was established.

REMARK 3.5 (see [18, Theorem 3.5], [36, Theorem 4.5.3]). *Under Assumptions* 3.1, 3.2, *and* 3.4, *there exists a canonical triplet* (ρ_*, h_*, f_*), *where* h_* *is a function in* $B_W(\mathbf{X})$ *that satisfies* (9)*; hence* (10) *holds.*

In others words, we already have a canonical policy $f_* \in \mathbf{F}_{cp} \subset \mathbf{F}_{eac}$ [see (11)]. It turns out that by suitably strengthening Assumption 3.1 we can also obtain SPAC-optimal policies with minimal average variance in $\mathbf{F}_{eac}$. Thus, consider the following.

ASSUMPTION 3.6. *There exists a constant* r_2 *such that*

$$C^2(x,a) \le r_2 W(x) \quad \forall (x,a) \in \mathbf{K}. \tag{16}$$

We can now state our first main result, which is proved in section 4.

THEOREM 3.7. *Suppose that Assumptions* 3.1, 3.2, 3.4, *and* 3.6 *are satisfied, and let* ρ_* *be as in* (10). *Then*

(a) *for each* $\pi \in \Pi$ *and* $x \in \mathbf{X}$

$$J^0(\pi,x) \ge \liminf_{n\to\infty} \frac{1}{n} J_n^0(\pi,x) \ge \rho_* \quad P_x^\pi - a.s.; \tag{17}$$

(b) *a stationary policy is EAC-optimal if and only if it is SPAC-optimal; hence (by* (17) *and Remark* 3.5*) there exists a SPAC-optimal policy* $f_* \in \mathbf{F}$, *and* ρ_* *is the optimal sample path average cost. That is,* $\rho_* = \widehat{\rho}$ *where* $\widehat{\rho}$ *is the constant in Definition* 2.1.

It is worth noting that Theorem 3.7(b) and the second inequality in (17) state that $f_* \in \mathbf{F}$ is in fact *strong SPAC-optimal*, where "strong" means that $f_* = \pi^*$ satisfies Definition 2.1 (with $\widehat{\rho} = \rho_*$) when the "lim-sup" SPAC in (3) is replaced by *lim-inf SPAC*

$$\underline{J}^0(\pi,x) := \liminf_{n\to\infty} \frac{1}{n} J_n^0(\pi,x). \tag{18}$$

On the other hand, denoting by $\mathbf{F}_{spac} \subset \mathbf{F}$ the class of SPAC-optimal stationary policies, we may rewrite the first statement in Theorem 3.7(b) as

$$\mathbf{F}_{eac} = \mathbf{F}_{spac}. \tag{19}$$

The reader should note that a priori neither one of the relations $\mathbf{F}_{eac} \subset \mathbf{F}_{spac}$ and $\mathbf{F}_{eac} \supset \mathbf{F}_{spac}$ is obvious!

To state our second main result we need some notation: For each $x \in \mathbf{X}$, let $A^*(x) \subset A(x)$ the set of control actions $a \in A(x)$ that attain the minimum in (7), i.e.,

$$A^*(x) := \left\{ a \in A(x) : \rho_* + h_*(x) = C(x,a) + \int_{\mathbf{X}} h_*(y) Q(dy|x,a) \right\}. \tag{20}$$

Observe that, by (8), a *policy* $f \in \mathbf{F}$ *is canonical* ($f \in \mathbf{F}_{cp}$) *if and only if* $f(x) \in A^*(x)$ $\forall\, x \in \mathbf{X}$. Moreover, consider the function Φ on $\mathbf{K}$ defined as

$$\Phi(x,a) := \int_{\mathbf{X}} h_*^2(y) Q(dy|x,a) - \left[\int_{\mathbf{X}} h_*(y) Q(dy|x,a) \right]^2. \tag{21}$$

As in (1), for $f \in \mathbf{F}$ and $x \in \mathbf{X}$ we write

$$\Phi_f(x) := \Phi(x, f(x)).$$

With this notation we can state our variance-minimization result as follows. (The proof is given in section 5.)

THEOREM 3.8. *Suppose that Assumptions* 3.1, 3.2, 3.4, *and* 3.6 *are satisfied. Then there exists a constant* $\sigma_*^2 \geq 0$, *a canonical policy* $f_* \in \mathbf{F}_{cp}$, *and a function* $V^*(\cdot)$ *in* $B_W(\mathbf{X})$ *such that, for each* $x \in \mathbf{X}$,

$$\begin{aligned} \sigma_*^2 + V^*(x) &= \min_{a \in A^*(x)} \left[\Phi(x,a) + \textstyle\int_{\mathbf{X}} V^*(y) Q(dy|x,a)\right] \\ &= \Phi_{f_*}(x) + \textstyle\int_{\mathbf{X}} V^*(y) Q_{f_*}(dy|x). \end{aligned} \tag{22}$$

Furthermore, f_* *satisfies* (13) *and* $V(f_*, \cdot) = \sigma_*^2$; *in fact,*

$$V(f_*, x) = \mu_{f_*}(\Phi_{f_*}) = \sigma_*^2 \quad \forall x \in X \tag{23}$$

and

$$\sigma_*^2 \leq V(f,x) \quad \forall f \in \mathbf{F}_{eac}, x \in \mathbf{X}. \tag{24}$$

Note the similarity between (22) and (7)–(8). This similarity is crucial in the proof of Theorem 3.8 (see section 5).

4. Proof of Theorem 3.7. The assumptions of Theorem 3.7 are supposed to hold throughout the following.

The proof of Theorem 3.7 requires some preliminary results, stated in the following lemmas. Let us first note that by Assumption 3.1(d1), the function C_f in (1) is in $B_W(\mathbf{X})$ for every $f \in \mathbf{F}$. Hence, as $\mu_f(W) < \infty$ (Remark 3.3(a)), the (finite) constants

$$J_f := \mu_f(C_f) \quad \text{for} \quad f \in \mathbf{F}$$

are well defined.

LEMMA 4.1. *Let* $f \in \mathbf{F}$ *be an arbitrary stationary policy. Then for each* $x \in \mathbf{X}$

$$J(f,x) = \lim_{n\to\infty} \frac{1}{n} J_n(f,x) = J_f \tag{25}$$

and also

$$J^0(f,x) = \lim_{n\to\infty} \frac{1}{n} J_n^0(f,x) = J_f \quad P_x^f - a.s. \tag{26}$$

Moreover, the function

$$h_f(x) := E_x^f \sum_{t=0}^{\infty} [C_f(x_t) - J_f] \tag{27}$$

belongs to $B_W(\mathbf{X})$ *and is such that the pair* (J_f, h_f) *satisfies the Poisson equation*

$$J_f + h_f(x) = C_f(x) + \int_{\mathbf{X}} h_f(y) Q_f(dy|x) \quad \forall x \in \mathbf{X}. \tag{28}$$

Proof. In (15) replace $u(\cdot)$ with $C_f(\cdot)$. Then (15) and Assumption 3.4 yield

$$\left|E_x^f C_f(x_n) - J_f\right| \leq r_1 M \gamma^n W(x) \quad \forall x \in \mathbf{X}, n = 0, 1, \ldots, \tag{29}$$

where r_1 is the constant in Assumption 3.1(d1), and clearly (25) follows from (29) and the definition (4) of the long-run EAC. Observe, on the other hand, that (29) yields

$$|h_f(x)| \leq r_1 M(1-\gamma)^{-1} W(x) \quad \forall x \in \mathbf{X},$$

so that h_f is indeed in $B_W(\mathbf{X})$, whereas writing (27) as

$$h_f(x) = C_f(x) - J_f + E_x^f \sum_{t=1}^{\infty} [C_f(x_t) - J_f]$$

and then using the Markov property, we obtain (28).

Finally, (26) is a consequence of Remark 3.3(a) and the strong law of large numbers for Markov chains—see, for instance, [31, p. 411]. □

Let us now consider the function $w(x) := W(x)^{1/2}$. By Jensen's inequality and Assumption 3.2(a),

$$\int_{\mathbf{X}} w(y) Q_f(dy|x) \leq B_f' w(x) + b_f' \mathbf{I}_{K_f}(x) \quad \forall f \in \mathbf{F}, x \in \mathbf{X}, \tag{30}$$

with $B_f' := B_f^{1/2} < 1$ and $b_f' := b_f^{1/2} < \infty$. Therefore, by Assumptions 3.2 and 3.4, we see the following from Remark 3.3(b).

LEMMA 4.2. *For each stationary policy $f \in \mathbf{F}$,*

(a) $\{x_n\}$ *is w-geometrically ergodic; that is,*

$$\left| \int_{\mathbf{X}} u(y) Q_f^n(dy|x) - \mu_f(u) \right| \leq ||u||_w M \gamma^n w(x) \tag{31}$$

for each $x \in \mathbf{X}, n = 0, 1, \ldots,$ and $u \in B_w(\mathbf{X})$, where $B_w(\mathbf{X})$ is the normed linear space of measurable functions u on $\mathbf{X}$ such that

$$||u||_w := \sup_{x \in \mathbf{X}} |u(x)|/w(x) < \infty;$$

(b) *the function h_f in* (27) *belongs to $B_w(\mathbf{X})$ (since* (16*) yields the w-analogue of Assumption* 3.1(d1) : $|C(x,a)| \leq r_2^{1/2} w(x)$ $\forall$ $(x,a) \in \mathbf{K}$).

The next two lemmas are crucial for the proof of (17).

LEMMA 4.3. *For each policy $\pi \in \Pi$ and initial state $x \in \mathbf{X}$ we have*

(a) $E_x^\pi \sum_{t=1}^{\infty} t^{-2} W(x_t) < \infty$;

hence the following statements hold $P_x^\pi-$ a.s.:

(b) $\sum_{t=1}^{\infty} t^{-2} W(x_t) < \infty$;

(c) $t^{-2} W(x_t) \to 0$;

(d) $t^{-1} w(x_t) \to 0$.

Proof. As $W \geq w \geq 1$, it is evident that $(a) \Rightarrow$ (b) $\Rightarrow$ (c) $\Rightarrow$ (d). Thus, it suffices to prove (a).

To prove (a), let us note that by Assumption 3.1 and a well-known measurable selection theorem (see, for instance, [16, Proposition D.5, p. 182]), there exists $g \in \mathbf{F}$ such that

$$\sup_{a \in A(x)} \int_{\mathbf{X}} W(y) Q(dy|x,a) = \int_{\mathbf{X}} W(y) Q_g(dy|x) \quad \forall x \in \mathbf{X},$$

which together with Assumption 3.2(a) yields

$$\int_{\mathbf{X}} W(y)Q(dy|x,a) \le B_g W(x) + b_g \mathbf{I}_{K_g}(x) \quad \forall (x,a) \in \mathbf{K}. \tag{32}$$

Now let $\pi \in \Pi$ and $x_0 = x \in \mathbf{X}$ be arbitrary, and let $\mathcal{F}_t$ be the $\sigma-$algebra generated by the state and control variables up to time t, that is,

$$\mathcal{F}_t := \sigma\{x_0, a_0, \cdots, x_t, a_t\} \quad \text{for} \quad t = 0, 1, \ldots. \tag{33}$$

Then, by the properties of the induced probability measure P_x^π [15, p. 4], [16, p. 16]

$$\begin{aligned} E_x^\pi \left[W(x_{t+1})|\mathcal{F}_t\right] &= \textstyle\int_{\mathbf{X}} W(y)Q(dy|x_t, a_t) \\ &\le B_g W(x_t) + b_g \quad [\text{by (32)}]. \end{aligned} \tag{34}$$

This implies

$$E_x^\pi W(x_{t+1}) \le B_g E_x^\pi W(x_t) + b_g \quad \forall t = 0, 1, \ldots.$$

Therefore

$$E_x^\pi W(x_t) \le B_g^t W(x) + b_g/(1 - B_g) \quad \forall t = 0, 1, \ldots, \tag{35}$$

and (a) follows. □

As in the previous proof, let $\pi \in \Pi$ and $x_0 = x \in \mathbf{X}$ be arbitrary, and let $\mathcal{F}_t$ be the σ-algebra in (33). Moreover, let h_* be the function in Remark 3.5 (see (7)), and define the random variables

$$Y_t(\pi, x) := h_*(x_t) - E_x^\pi \left[h_*(x_t)|\mathcal{F}_{t-1}\right] \quad \text{for} \quad t = 1, 2, \ldots \tag{36}$$

and

$$M_n(\pi, x) := \sum_{t=1}^{n} Y_t(\pi, x) \quad \text{for} \quad t = 1, 2, \ldots. \tag{37}$$

LEMMA 4.4. *For each policy $\pi \in \Pi$ and each initial state $x_0 = x \in \mathbf{X}$, the sequence $\{M_n(\pi, x)\}$ is a P_x^π-martingale with respect to the filtration $\{\mathcal{F}_n\}$, and*

$$\lim_{n\to\infty} \frac{1}{n} M_n(\pi, x) = 0 \quad P_x^\pi - a.s. \tag{38}$$

Proof. Choose an arbitrary policy $\pi \in \Pi$ and an arbitrary initial state x. Now note that (ρ_*, h_*) is a solution to the Poisson equation (8) (cf. (28)), so that, by Lemma 4.2(b),

$$h_* \text{ is in } B_w(\mathbf{X}); \text{ that is, } |h_*(\cdot)| \le ||h_*||_w w(\cdot). \tag{39}$$

Then, by (36),

$$\begin{aligned} |Y_t(\pi, x)| &\le |h_*(x_t)| + E_x^\pi \left[|h_*(x_t)| \,|\mathcal{F}_{t-1}\,\right] \\ &\le ||h_*||_w \left\{w(x_t) + E_x^\pi \left[w(x_t) \mid \mathcal{F}_{t-1}\right]\right\}, \end{aligned} \tag{40}$$

and it follows that

$$E_x^\pi |Y_t(\pi, x)| \leq 2||h_*||_w E_x^\pi w(x_t).$$

This inequality and (35) show that $M_n(\pi, x)$ is P_x^π-integrable for every n. On the other hand, it is clear that $M_n(\pi, x)$ is $\mathcal{F}_n$-measurable and that

$$E_x^\pi \left[M_{n+1}(\pi, x) - M_n(\pi, x) \mid \mathcal{F}_n \right] = 0 \quad P_x^\pi - \text{a.s.}, \ \forall n;$$

therefore, $\{M_n(\pi, x)\}$ is a martingale.

Thus, (38) will follow from the law of large numbers for martingales (or martingale stability theorem) [14, Theorem 2.18, p. 35], [23, p. 161], [27, p. 53] provided that

$$\sum_{t=1}^{\infty} t^{-2} E_x^\pi \left[|Y_t(\pi, x)|^2 \mid \mathcal{F}_{t-1} \right] < \infty \quad P_x^\pi - \text{a.s.} \tag{41}$$

To prove (41), first we use the elementary inequality

$$(a + b)^2 \leq 2(a^2 + b^2) \quad \forall a, b \in \mathbb{R}$$

and the fact that $w^2 := W$ to see that (40) yields

$$|Y_t(\pi, x)|^2 \leq 2||h_*||_w^2 \left\{ W(x_t) + E_x^\pi \left[W(x_t) \mid \mathcal{F}_{t-1} \right] \right\}$$

so that

$$E_x^\pi \left[|Y_t(\pi, x)|^2 \mid \mathcal{F}_{t-1} \right] \leq 4||h_*||_w^2 E_x^\pi \left[W(x_t) \mid \mathcal{F}_{t-1} \right].$$

Thus, by (34),

$$E_x^\pi \left[|Y_t(\pi, x)|^2 \mid \mathcal{F}_{t-1} \right] \leq 4||h_*||_w^2 (B_g + b_g) W(x_{t-1}).$$

This inequality and Lemma 4.3(b) yield (41) because

$$\sum_{t=1}^{\infty} t^{-2} W(x_t) \leq W(x_0) + \sum_{t=2}^{\infty} (t-1)^{-2} W(x_{t-1}) < \infty \qquad P_x^\pi - \text{a.s.}$$

This completes the proof of Lemma 4.4. □

We are now ready to prove Theorem 3.7 itself.

Proof of Theorem 3.7. (a). Fix $\pi \in \Pi$ and $x_0 = x \in \mathbf{X}$ arbitrary, and consider Mandl's [30] *discrepancy function* $D : \mathbf{K} \to \mathbb{R}$ defined as

$$D(x, a) := C(x, a) + \int_{\mathbf{X}} h_*(y) Q(dy|x, a) - h_*(x) - \rho_*.$$

By Remark 3.5 and (7), D is *nonnegative*. On the other hand, rewriting (36) as

$$Y_t(\pi, x) = h_*(x_t) - \int_{\mathbf{X}} h_*(y) Q(dy|x_{t-1}, a_{t-1}),$$

we see that (37) becomes

$$M_n(\pi, x) = h_*(x_n) - h_*(x_0) - \sum_{t=0}^{n-1} D(x_t, a_t) + J_n^0(\pi, x) - n\rho_*,$$

and so

$$J_n^0(\pi, x) \geq n\rho_* + M_n(\pi, x) - h_*(x_n) + h_*(x_0). \tag{42}$$

Finally, multiply both sides of (42) by $1/n$ and then take lim inf as $n \to \infty$ to obtain the second inequality in (17) [from (38), (39), and Lemma 4.3(d)]. As the first inequality in (17) is obvious, (a) follows.

(b) Suppose that $f \in \mathbf{F}$ is EAC-optimal. Then from (26) we get

$$J^0(f, x) = \rho_* \qquad P_x^f - \text{a.s.} \quad \forall x \in \mathbf{X},$$

which together with (17) yields that f is SPAC-optimal. (In other words, $\mathbf{F}_{eac} \subset \mathbf{F}_{spac}$—see (19).)

Conversely, suppose that $f \in \mathbf{F}$ is SPAC-optimal. That is,

$$J^0(f, x) = J_f \qquad P_x^f - \text{a.s.} \quad \forall x \in \mathbf{X}$$

and

$$J^0(\pi, x) \geq J_f \quad P_x^\pi - \text{a.s.} \quad \forall \pi \in \Pi, x \in \mathbf{X}.$$

In particular, the latter inequality and (26) yield

$$J_g \geq J_f \quad \forall g \in \mathbf{F}.$$

Hence, by Remark 3.5, $J_f = \rho_*$; that is, f is EAC-optimal. (In other words, $\mathbf{F}_{spac} \subset \mathbf{F}_{eac}$.) □

5. Proof of Theorem 3.8. We shall first state some preliminary facts.

By Lemma 4.2(b), the function h_f is in $B_w(\mathbf{X})$ for each $f \in \mathbf{F}$, and so the function

$$\Psi_f(x) := \int_{\mathbf{X}} h_f^2(y) Q_f(dy|x) - \left[\int_{\mathbf{X}} h_f(y) Q_f(dy|x)\right]^2 \quad \text{for} \quad x \in \mathbf{X} \tag{43}$$

belongs to $B_W(\mathbf{X})$. Thus, as $\mu_f(W) < \infty$ (Remark 3.3(a)), the (finite) constants

$$\sigma_f^2 := \mu_f(\Psi_f) \quad \text{for} \quad f \in \mathbf{F} \tag{44}$$

are well defined, and they coincide with the limiting average variance in (12). More explicitly, by the central limit theorem for Markov chains we have the following (see, for instance, [7, p. 302]; [31, pp. 411, 436]).

LEMMA 5.1. *For each* $f \in \mathbf{F}$

$$V(f, x) = \lim_{n \to \infty} \frac{1}{n} E_x^f \sum_{t=0}^{n-1} \Psi_f(x_t) = \sigma_f^2 \quad \forall x \in \mathbf{X}. \tag{45}$$

On the other hand, if $f \in \mathbf{F}$ is a *canonical policy*, then the corresponding solution $(J_f, h_f) = (\rho_*, h_f)$ to the Poisson equation (28) is such that h_f coincides with h_* except perhaps for an additive constant, that is [18, 36],

$$h_f(\cdot) = h_*(\cdot) + k_f \quad \forall f \in \mathbf{F}_{cp} \tag{46}$$

for some constant k_f. Therefore, comparing (43) and (21), we obtain

$$\Phi_f(\cdot) = \Psi_f(\cdot) \quad \forall f \in \mathbf{F}_{cp}. \tag{47}$$

We will next prove, in particular, that (46) and (47) hold μ_f-a.e. for all stationary EAC-optimal policies $f \in \mathbf{F}_{eac}$. (Recall (11).)

LEMMA 5.2. *Let f be a stationary policy in $\mathbf{F}_{eac}\backslash\mathbf{F}_{cp}$. Then*

(a) $h_f(\cdot) = h_*(\cdot) + k_f \quad \mu_f$*-a.e. for some constant* k_f;

(b) *there exists a canonical policy $\widehat{f} \in \mathbf{F}_{cp}$ such that $\mu_{\widehat{f}} = \mu_f$, and*

$$V(\widehat{f}, x) = V(f, x) = \sigma_f^2 \quad \forall x \in \mathbf{X}, \tag{48}$$

where σ_f^2 is the constant in (44), (45).

Proof. (a) From the so-called average cost optimality equation (7), we have

$$\rho_* + h_*(x) \le C_f(x) + \int_{\mathbf{X}} h_*(y) Q_f(dy|x) \quad \forall x \in \mathbf{X}. \tag{49}$$

On the other hand, we also have $J_f = \rho_*$ because f is EAC-optimal, and so the Poisson equation (28) for f becomes

$$\rho_* + h_f(x) = C_f(x) + \int_{\mathbf{X}} h_f(y) Q_f(dy|x) \quad \forall x \in \mathbf{X}. \tag{50}$$

From (49) and (50), it follows that the function $u(\cdot) = h_f(\cdot) - h_*(\cdot)$ is super-harmonic with respect to Q_f, i.e.,

$$u(x) \ge \int_{\mathbf{X}} u(y) Q_f(dy|x) \quad \forall x \in \mathbf{X}.$$

Iterating this inequality, we see that

$$u(x) \ge \int_{\mathbf{X}} u(y) Q_f^n(dy|x) \quad \forall x \in \mathbf{X}, n = 0, 1, \ldots,$$

and if we let $n \to \infty$, (31) yields

$$u(x) \ge \int_{\mathbf{X}} u(y) \mu_f(dy) \quad \forall x \in \mathbf{X}. \tag{51}$$

Now let $k_f := \inf_{x \in \mathbf{X}} u(x)$. Then, by (51), $\int_{\mathbf{X}} u(y)\mu_f(dy) = k_f$, which implies

$$u(\cdot) = k_f \quad \mu_f\text{-a.e.}$$

Thus, as $u := h_f - h_*$, (a) follows with $k_f = \inf_{x \in \mathbf{X}} u(x) = \int_{\mathbf{X}} u(y)\mu_f(dy)$.

(b) By (a), there exists a Borel set $N \in \mathcal{B}(\mathbf{X})$ such that $\mu_f(N) = 0$ and

$$h_f(x) = h_*(x) + k_f \quad \forall x \in N^c := \mathbf{X}\backslash N. \tag{52}$$

Now let $g \in \mathbf{F}_{cp}$ be a canonical policy (whose existence is ensured by Remark 3.5), and define a new policy $\widehat{f}$ as $\widehat{f} := g$ on N and $\widehat{f} := f$ on N^c. Then, $\widehat{f}$ is canonical, and $Q_f(\cdot|x) = Q_{\widehat{f}}(\cdot|x)$ on N^c, that is,

$$Q_f(\cdot|x) = Q_{\widehat{f}}(\cdot|x) \quad \mu_f - \text{a.e.} \tag{53}$$

Hence, on the one hand, (53) yields

$$\mu_f(\cdot) = \mu_{\widehat{f}}(\cdot), \tag{54}$$

and, on the other hand, (53) and (52) give (from (43) and (21))

$$\Psi_f(\cdot) = \Phi_f(\cdot) \quad \mu_f\text{-a.e.} \tag{55}$$

Therefore, (48) follows from (44)–(45) and (54)–(55). □

We are now ready for the proof of Theorem 3.8.

Proof of Theorem 3.8. Let $A^*(x)$ and $\Phi(x,a)$ be as in (20) and (21), respectively, and consider the new Markov control model

$$(\mathbf{X}, \mathbf{A}, \{A^*(x) : x \in \mathbf{X}\}, Q, \widehat{C})$$

with $\widehat{C}(x,a) := \Phi(x,a)$. It is easy to check that this control model satisfies Assumptions 3.1, 3.2, and 3.4, replacing C and $A(\cdot)$ with $\widehat{C}$ and $A^*(\cdot)$, respectively. Consequently, by Remark 3.5, there exists a constant $\sigma_*^2 \geq 0$, a function $V^*(\cdot)$ in $B_W(\mathbf{X})$, and a canonical policy $f_* \in \mathbf{F}_{cp}$ that satisfy (22). Then, by standard arguments it follows that

$$\sigma_*^2 = \mu_{f_*}(\Phi_{f_*}) = V(f_*, x) \quad \forall x \in \mathbf{X} \tag{56}$$

and

$$\sigma_*^2 \leq \mu_f(\Phi_f) = V(f,x) \quad \forall f \in \mathbf{F}_{cp}, x \in \mathbf{X}, \tag{57}$$

where the last equality in (56) and (57) follows from (47) and (45). Finally, from (57) and Lemma 5.2(b) we conclude that

$$\sigma_*^2 \leq \sigma_f^2 = V(f,x) \quad \forall f \in \mathbf{F}_{eac}, x \in \mathbf{X}.$$

This completes the proof of Theorem 3.8. □

6. Additional comments. In this section we briefly discuss alternative forms of the stability Assumptions 3.2 and 3.4, as well as examples related to our main results.

In addition to the space of functions $B_W(\mathbf{X})$ with the W-norm (14), where $W \geq 1$ is the "weight" function in Assumption 3.1(d), we shall consider the normed linear space $M_W(\mathbf{X})$ of the finite signed measures with a finite W-norm

$$||\mu||_W := \int_{\mathbf{X}} W(y)|\mu|(dy), \tag{58}$$

where $|\mu| := \mu^+ + \mu^-$ denotes the *total variation of* μ.

REMARK 6.1. *Suppose that Assumption* 3.1 *holds. Then from the proofs in* [18, *Theorem* 3.5] *and* [36, *Theorem* 4.5.3] *one can see that the result mentioned in Remark* 3.5 *holds provided that*

(I) *the* W*-geometric ergodicity* (15) *holds, with constants* M_f *and* γ_f *that satisfy Assumption* 3.4; *and*

(II) *The transition kernel* Q_f *is* φ*-irreducible for each* $f \in \mathbf{F}$, *where* φ *is a* σ*-finite measure on* X *independent of* $f \in \mathbf{F}$.

These two conditions were obtained in section 3 from Assumptions 3.2 and 3.4. However, there are other ways of getting I. For example, Assumptions 3.2 and 3.4 may be replaced by the following hypotheses used in [12, 13].

(a) For each stationary policy $f \in \mathbf{F}$ the kernel Q_f (in (1)) admits a unique invariant probability measure μ_f.

(b) There exists a probability measure ν in $M_W(\mathbf{X})$ and positive constants $\alpha < \infty$ and $\beta < 1$ for which the following holds: for each $f \in \mathbf{F}$ there exists a measurable function $0 \leq l_f \leq 1$ such that

(b1) $Q_f(B|x) \geq l_f(x)\nu(B) \quad \forall x \in \mathbf{X}, B \in \mathcal{B}(\mathbf{X})$;

(b2) $\nu(l_f) \geq \alpha$, and $\nu(W) = ||\nu||_W < \infty$;

(b3) $\int_{\mathbf{X}} W(y)Q_f(dy|x) \leq \beta W(x) + l_f(x)\nu(W) \quad \forall x \in \mathbf{X}$.

These conditions, (a) and (b), are an adaptation to MCPs of ideas used by Kartashov [22] to obtain the W-geometric ergodicity in (15). In fact, under (a) and (b), our present Assumption 3.4 is satisfied and one can also obtain estimates of the constants M and γ—see [22, Theorem 3.6] and [12, Lemmas 3.3 and 3.4]—in Assumption 3.4.

Similarly, instead of (a) and (b), one could adapt to MCPs the "contraction" property in [22, Corollary 2.1], which in our notation would be of the form $||\theta Q_f||_W \leq \rho||\theta||_W$ for every signed measure θ in $M_W(\mathbf{X})$ with $\theta(\mathbf{X}) = 0$, for some positive constant $\rho < 1$. On the other hand, as was already noted in [18, Remark 2.10], if the cost function $C(x, a)$ is *bounded*, then the "weight" function $W \geq 1$ may be bounded and (15) can be obtained from *Doeblin's condition*—see [31, Theorem 16.0.2].

To conclude, we should mention that the examples in [13] and [18] (see also [36, Chapter 4]) also hold in our present case. For instance, the example in [18, section 6], consists of an inventory system with state space $\mathbf{X} := [0, \infty)$ and a compact control set $\mathbf{A} \subset \mathbb{R}$, in which the one-step cost $C(x, a)$ is *piecewise-linear* in $x \in \mathbf{X}$ and $a \in \mathbf{A}$. Therefore, as the weight function is *exponential*, say

$$W(x) := k \exp(rx) \quad \forall x \in \mathbf{X},$$

with $k, r > 0$, our Assumption 3.6 will trivially hold for some $r_2 > 0$ sufficiently large. A similar comment holds for the queueing system in [13, section 5], except that this reference uses the conditions (a) and (b) in Remark 6.1 in lieu of Assumptions 3.2 and 3.4.

REFERENCES

[1] A. Arapostathis, V.S. Borkar, E. Fernández-Gaucherand, M.K. Ghosh, and S.I. Marcus, *Discrete-time controlled Markov processes with average cost criterion: A survey*, SIAM J. Control Optim., 31 (1993), pp. 282–344.

[2] D.P. Bertsekas, *Dynamic Programming: Deterministic and Stochastic Control Models*, Prentice–Hall, Englewood Cliffs, NJ, 1987.

[3] B.S. Borkar, *Control of Markov chains with long-run average cost criterion*, in Stochastic Differential Systems, Stochastic Control Theory and Applications, W. Fleming and P.L. Lions, eds., IMA Vol. Math. Appl. 10, Springer-Verlag, Berlin, 1988, pp. 57–77.

[4] V.S. Borkar, *Topics in Controlled Markov Processes*, Pitman Res. Notes Math. Ser. 240, Longman, Harlow, UK, 1991.

[5] R. Cavazos-Cadena and E. Fernández-Gaucherand, *Denumerable controlled Markov chains with average reward criterion: Sample path optimality*, Math. Methods Oper. Res., 41 (1995), pp. 89–108.

[6] R. Dekker and A. Hordijk, *Recurrence conditions for average and Blackwell optimality in denumerable state Markov decision chains*, Math. Oper. Res., 17 (1992), pp. 271–289.

[7] M. Duflo, *Méthodes Récursives Aléatoires*, Masson, Paris, 1990.

[8] E.B. Dynkin and A.A. Yushkevich, *Controlled Markov Processes*, Springer-Verlag, New York, 1979.

[9] P.W. Glynn and S.P. Meyn, *A Lyapunov bound for solutions of the Poisson equation*, Ann. Probab., 24 (1996), pp. 916–931.

[10] J. González-Hernández and O. Hernández-Lerma, *Envelopes of sets of measures, tightness, and Markov control processes*, Appl. Math. Optim., 40 (1999), pp. 377–392.

[11] J. González-Hernández and O. Hernández-Lerma, *Constrained Markov control processes in Borel spaces: The discounted case*, Math. Methods Oper. Res., submitted.
[12] E. Gordienko and O. Hernández-Lerma, *Average cost Markov control processes with weighted norms: Existence of canonical policies*, Appl. Math. (Warsaw), 23 (1995), pp. 199–218.
[13] E. Gordienko and O. Hernández-Lerma, *Average cost Markov control processes with weighted norms: Value iteration*, Appl. Math. (Warsaw), 23 (1995), pp. 219–237.
[14] P. Hall and C.C. Heyde, *Martingale Limit Theory and Its Application*, Academic Press, New York, 1980.
[15] O. Hernández-Lerma, *Adaptive Markov Control Processes*, Springer-Verlag, New York, 1989.
[16] O. Hernández-Lerma and J.B. Lasserre, *Discrete-Time Markov Control Processes: Basic Optimality Criteria*, Springer-Verlag, New York, 1996.
[17] O. Hernández-Lerma and J.B. Lasserre, *Policy iteration for average cost Markov control processes on Borel spaces*, Acta Appl. Math., 47 (1997), pp. 125–154.
[18] O. Hernández-Lerma and O. Vega-Amaya, *Infinite-horizon Markov control processes with undiscounted cost criteria: From average to overtaking optimality*, Appl. Math. (Warsaw), 25 (1998), pp. 153–178.
[19] A. Hordijk, *Dynamic Programming and Markov Potential Theory*, Mathematical Centre Tracts, 51, Mathematisch Centrum, Amsterdam, 1974.
[20] A. Hordijk and A.A. Yushkevich, *Blackwell optimality in the class of stationary policies in Markov decision chains with a Borel state space and unbounded rewards*, Math. Methods Oper. Res., 49 (1999), pp. 1–39.
[21] Y. Huang and L.C.M. Kallenberg, *On finding optimal policies for Markov decision chains: A unifying framework for mean-variance tradeoffs*, Math. Oper. Res., 19 (1994), pp. 434–448.
[22] N.V. Kartashov, *Strong Stable Markov Chains*, VSP, Utrecht, The Netherlands, 1996.
[23] P.R. Kumar and P. Varaiya, *Stochastic Systems: Estimation, Identification, and Adaptive Control*, Prentice–Hall, Englewood Cliffs, NJ, 1986.
[24] M. Kurano, *Markov decision processes with a minimum-variance criterion*, J. Math. Anal. Appl., 123 (1987), pp. 572–583.
[25] H.J. Kushner, *Stochastic Stability and Control*, Academic Press, New York, 1967.
[26] J.B. Lasserre, *Sample-path average optimality for Markov control processes*, IEEE Trans. Automat. Control, 44 (1999), pp. 1966–1971.
[27] M. Loève, *Probability Theory* II, Springer-Verlag, New York, 1978.
[28] P. Mandl, *On the variance in controlled Markov chains*, Kybernetika (Prague), 7 (1971), pp. 1–12.
[29] P. Mandl, *A connection between controlled Markov chains and martingales*, Kybernetika (Prague), 9 (1973), pp. 237–241.
[30] P. Mandl, *Estimation and control in Markov chains*, Adv. Appl. Probab., 6 (1974), pp. 40–60.
[31] S.P. Meyn and R.L. Tweedie, *Markov Chains and Stochastic Stability*, Springer-Verlag, London, 1993.
[32] M.L. Puterman, *Markov Decision Processes*, John Wiley, New York, 1994.
[33] L.I. Sennott, *Stochastic Dynamic Programming and the Control of Queueing Systems*, John Wiley, New York, 1999.
[34] H.C. Tijms, *Stochastic Modelling and Analysis: A Computational Approach*, John Wiley, Chichester, U.K., 1986.
[35] O. Vega-Amaya, *Sample-path average optimality of Markov control processes with strictly unbounded cost*, Appl. Math. (Warsaw), to appear.
[36] O. Vega-Amaya, *Markov Control Processes in Borel Spaces: Undiscounted Cost Criteria*, Doctoral Thesis, UAM-Iztapalapa, México, 1998 (in Spanish).
[37] A.A. Yushkevich, *On a class of strategies in general Markov decision models*, Theory Probab. Appl., 18 (1973), pp. 777–779.

SIAM J. CONTROL OPTIM.
Vol. 38, No. 1, pp. 94–123

ACTOR-CRITIC–TYPE LEARNING ALGORITHMS FOR MARKOV DECISION PROCESSES*

VIJAYMOHAN R. KONDA† AND VIVEK S. BORKAR‡

Abstract. Algorithms for learning the optimal policy of a Markov decision process (MDP) based on simulated transitions are formulated and analyzed. These are variants of the well-known "actor-critic" (or "adaptive critic") algorithm in the artificial intelligence literature. Distributed asynchronous implementations are considered. The analysis involves two time scale stochastic approximations.

Key words. reinforcement learning, Markov decision processes, actor-critic algorithms, stochastic approximation, asynchronous iterations

AMS subject classifications. 93E35, 68T05, 62L20

PII. S036301299731669X

1. Introduction.

Recently there has been a lot of interest in simulation-based algorithms for MDPs. In these, one takes as the starting point one of the classical iterative algorithms for computing the value function: the value iteration or policy iteration. The conditional expectation term in these is then replaced by its integrand evaluated at an actual simulated transition. The algorithm is expected to "see" the conditional expectation through an averaging effect achieved by using stochastic approximationlike step sizes. These schemes are very useful when the exact dynamics is unknown or complicated but simulation is "easy." An instance would be a large communication network where an exact analytic model is hard to derive but simulating a typical transition using local dynamics is easy.

The simulation-based scheme derived from value iteration is called Q-learning and has been extensively analyzed [26, 24, 1, 2]. The scheme derived from policy iteration is called adaptive critic [3] and has eluded a satisfactory convergence analysis (see [27]). This is because policy iteration involves two loops. The inner loop computes the "value function" for a stationary policy and the outer one updates the latter. If both are done recursively, ideally the outer update should wait for the convergence of the inner loop. The aim of this paper is to propose several variants of the adaptive critic method where the above difficulty is circumvented by a two time scale stochastic approximation. That is, we use different step size schedules for different components of the iteration. Recall that the traditional stochastic approximation algorithm asymptotically tracks an associated ODE [17]. The two time scale algorithm correspondingly tracks a singular ODE. As in the case of singular ODEs, the fast component sees the slow component as quasi-constant, while the slow component sees the fast one as essentially equilibrated. Thus operating the outer loop of our algorithm on a slower (virtual) time scale than the inner loop achieves the desired effect. These

*Received by the editors February 12, 1997; accepted for publication (in revised form) February 17, 1999; published electronically December 3, 1999.
http://www.siam.org/journals/sicon/38-1/31669.html

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

‡School of Technology and Computer Science, Tata Institute of Fundamental Research, Mumbai 400005, India (borkar@tifr.res.in).The research of this author was supported in part by the Homi Bhabha Fellowship and in part by Department of Science and Technology (Government of India) grant III 5(12)/96-ET.

qualitative remarks will be made precise later. Another important feature of our work is the distributed, asynchronous iterations typical in applications. (Unfortunately, we have to stick to the synchronous iterations for one of the time scales in the case of the average cost problems, due to some technical difficulties; see [11].)

The reader is referred to [6, 14] for extensive accounts of earlier work on simulation-based algorithms. Two time scale stochastic approximations are studied in [8]. Our approach to asynchronous algorithms follows [9].

The paper is organized as follows. The next section proposes and analyzes a prototypical algorithm without any reference to MDPs. All the algorithms we propose will be variants of this. Section 3 introduces the MDP framework and the associated algorithms. These are analyzed in sections 5 and 6 for discounted and average cost, respectively, following some preliminaries in section 4. The last section concludes with some relevant remarks. Algorithm 2 was also presented in [10]. It is included here for the sake of completeness.

2. General algorithm.

2.1. Notation. Let l, $m \geq 1$, $I = \{1, \ldots, l\}$, $I' = \{1, \ldots, m\}$. Let e_i denote a unit l-vector whose ith component is one. Let $I_1, I_2, \ldots, I_K$ be a partition of I, and let $G_1, G_2, \ldots, G_K$ be closed convex subsets of R^l such that $G_i \subset H_i \stackrel{\Delta}{=} \text{span}\{e_k | k \in I_i\}$ and $G = \{\sum_i x_i : x_j \in G_j \ \forall j\}$ is compact. Let $\pi_i : R^l \to H_i$ denote the projection map; let $P_i : H_i \to G_i$ and $P : R^l \to G$ denote the projections given by $P_i(x) = y$ subject to $y \in G_i$, $\|y - x\| = \inf_{z \in G_i} \|z - x\|$, $P(x) = y$ s.t. $\|y - x\| = \inf_{z \in G} \|z - x\|$. It is then easy to see the following:

(1) For each $x \in R^l$, $x = \sum_{k=1}^K \pi_k(x)$.

(2) For each $i = 1, 2, \ldots, K$, $G_i = \pi_i(G)$.

For each $x \in R^l$, $P(x) = \sum_{i=1}^K P_i(\pi_i(x))$. We further assume that $\{G_i\}$ are of the form

$$G_i = \{x \in H_i : q_{ij}(\pi_i(x)) \leq 0, \quad 1 \leq j \leq K_i$$

for some continuously differentiable functions $\{q_{ij}(\cdot)\}$.

Let $f : G \times R^m \to R^m$, $g : G \times R^m \to R^l$ be prescribed Lipschitz maps such that for each $x \in G$, $f(x, y) = 0$ has a unique solution $y = \lambda(x)$. Assume that $\lambda(\cdot)$ is Lipschitz. Define

$$\bar{P}_x(g(x, \lambda(x))) = \lim_{\Delta \downarrow 0} \frac{P(x + \Delta g(x, \lambda(x))) - x}{\Delta}.$$

The algorithm proposed below aims to find the solution of $\bar{P}_x(g(x, \lambda(x))) = 0$ based on noisy samples. We describe this next.

Let $\{\bar{X}_n, \bar{Y}_n\}$ be set-valued random processes taking values in the subsets of I, I', respectively, and set

$$\nu_1(i, n) = \sum_{k=0}^{n-1} I\{i \in \bar{X}_k\}, \quad \nu_1(i, 0) = 0, \quad 1 \leq i \leq l,$$

$$\nu_2(j, n) = \sum_{k=0}^{n-1} I\{j \in \bar{Y}_k\}, \quad \nu_2(j, 0) = 0, \quad 1 \leq j \leq m.$$

Let $\{a(n)\}$ be positive sequences satisfying

$$\sum_n a(n) = \sum_n b(n) = \infty, \quad \sum_n a(n)^2 < \infty, \quad \sum_n b(n)^2 < \infty. \tag{2.1}$$

Then the algorithm is as follows: For $1 \leq k \leq K$, $1 \leq i \leq m$,

(2.2)
$$\pi_k(X(n+1)) = P_k\left(\pi_k(X(n)) + \sum_{j \in I_k} a(\nu_1(j,n))(g_j(\tilde{X}^j(n), \tilde{Y}^j(n)) + M_j(n))I\{j \in \bar{X}_n\}e_j\right),$$

$$Y_i(n+1) = Y_i(n) + b(\nu_2(i,n))(f_i(\hat{X}^i(n), \hat{Y}^i(n)) + M'_i(n))I\{i \in \bar{Y}_n\}. \tag{2.3}$$

Here $\{M(n)\}$, $\{M'(n)\}$ are vector martingale difference sequences and

$$\begin{aligned}
\tilde{X}^j(n) &= [X_1(n - \tau_{j1}(n)), \ldots, X_l(n - \tau_{jl}(n))]^T, \\
\tilde{Y}^j(n) &= [Y_1(n - \bar{\tau}_{j1}(n)), \ldots, Y_m(n - \bar{\tau}_{jm}(n))]^T, \\
\hat{X}^i(n) &= [X_1(n - \rho_{i1}(n)), \ldots, X_l(n - \rho_{il}(n))]^T, \\
\hat{Y}^i(n) &= [Y_1(n - \bar{\rho}_{i1}(n)), \ldots, Y_m(n - \bar{\rho}_{im}(n))]^T,
\end{aligned}$$

for nonnegative random "delays" $\{\tau_{ij}(\cdot)\}$, $\{\bar{\tau}_{ij}(\cdot)\}$, $\{\rho_{ij}(\cdot)\}$, $\{\bar{\rho}_{ij}(\cdot)\}$.

The interpretation is as follows: $\bar{Y}_n$ is the set of indices i such that $Y_i(n)$ gets updated at time n. Similarly, $\bar{X}_n$ is the set of indices j such that the noisy sample $(g_j(\tilde{X}^j(n), \tilde{Y}^j(n)) + M_j(n))$ is used to update $X(n)$. Each component of $X(\cdot)$, $Y(\cdot)$ is assigned to a unique processor which receives the outputs of other processors with some random delays and updates all the components assigned to it. Each component of $Y(\cdot)$ is assigned to a separate processor. Two components X_i, X_j of $X(\cdot)$ are assigned to the same processor and experience the same communication delays if there exists $1 \leq k \leq K$ such that $i, j \in I_k$. $\{M(n)\}$, $\{M'(n)\}$ are martingale difference sequences (with respect to the "natural" σ-fields generated by processes under consideration) representing measurement noise. $\nu_1(i,n)$ (respectively, $\nu_2(i,n)$) is the number of times the ith component of $X(\cdot)$ (respectively, $Y(\cdot)$) was updated up to time n. These are known to the respective processors who need not know n (the global "clock"). See [9] for a further discussion of this paradigm.

2.2. Assumptions. We shall make the following assumptions:

(A1) *Ideal step size.* Letting $c(n)$ stand for either $a(n)$ or $b(n)$, we assume, in addition to (2.1), that $c(n)$ is eventually decreasing and the following hold:

1. $a(n) = o(b(n))$.
2. For $x \in (0,1)$,
$$\sup_n c([xn])/c(n) < \infty, \tag{2.4}$$
where $[\cdots]$ stands for the integer part of "...".
3. For $x \in (0,1)$ and $A(n) = \sum_{i=0}^{n} c(i)$,
$$A([yn])/A(n) \to 1 \tag{2.5}$$
uniformly in $y \in [x,1]$.

Examples are $c(n) = \frac{1}{n}$, $\frac{1}{n \log n}$ for $n \geq 2$ with suitable modifications for $n = 0, 1$.

(A2) *Boundedness of iterates.* $\sup_n \|Y(n)\| < \infty$.

(A3) *ODE stability.* For each $x \in G$, the ODE
$$\dot{y}(t) = f(x, y(t)), \qquad y(0) = y,$$

has a unique globally asymptotically stable equilibrium point $\lambda(x)$, where $\lambda(\cdot)$ is Lipschitz. (Note that this means that the ODE $\dot{x}(t) = 0$, $\dot{y}(t) = f(x(t), y(t))$ has $\text{graph}(\lambda) \stackrel{\Delta}{=} \{(x, \lambda(x)) : x \in G\}$ as the globally asymptotically stable set.) Also, the ODE

$$\dot{x}(t) = \bar{P}_{x(t)}(g(x(t), \lambda(x(t)))), \qquad x(0) = x,$$

has a strict Liapunov function, i.e., a twice continuously differentiable function $L(\cdot) : G \to R$ such that

$$\nabla L(x) \cdot \bar{P}_x(g(x, \lambda(x))) < 0$$

whenever $\bar{P}_x(g(x, \lambda(x))) \neq 0$.

(A4) *Ideal noise.* $\{M(n)\}$, $\{M'(n)\}$ are square integrable and satisfy

$$\sup_n \|M(n)\|, \sup_n \|M'(n)\| < \infty \text{ almost surely},$$

$$\mathrm{E}[M(n)|\mathcal{F}_n] = \mathrm{E}[M'(n)|\mathcal{F}_n] = 0 \text{ a.s.},$$

$$\sup_n \mathrm{E}[\|M(n)\|^2|\mathcal{F}_n], \sup_n \mathrm{E}[\|M'(n)\|^2|\mathcal{F}_n] < \infty \text{ a.s.},$$

where $\mathcal{F}_n = \sigma(X(k), Y(k), \{\tau_{ij}(k)\}, \{\bar{\tau}_{ij}(k)\}, \{\rho_{ij}(k)\}, \{\bar{\rho}_{ij}(k)\}, k \leq n, M(k), M'(k), k < n), n \geq 0$.

(A5) *Bounded delays.* All interprocessor delays are bounded by a common deterministic constant D.

(A6) *Frequent updates.* $\bar{X}_n$, $\bar{Y}_n \neq \emptyset$ $\forall n$ and there exists an $\bar{\eta} > 0$ such that

$$\liminf_{n\to\infty} \frac{\nu_1(i,n)}{n} \geq \bar{\eta},$$

$$\liminf_{n\to\infty} \frac{\nu_2(j,n)}{n} \geq \bar{\eta}$$

a.s. $\forall i, j$.

Furthermore, if, for $\bar{a}(\cdot), \bar{b}(\cdot)$ defined as in section 4.2 below,

$$N(n,x) = \min\left\{m > n : \sum_{i=n+1}^{m} \bar{a}(i) \geq x\right\},$$

$$N'(n,x) = \min\left\{m > n : \sum_{i=n+1}^{m} \bar{b}(i) \geq x\right\},$$

for $x > 0$, then the limits

$$\lim_{n\to\infty} \frac{\sum_{j=\nu_1(i,n)}^{\nu_1(i,N(n,x))} a(j)}{\sum_{j=\nu_1(k,n)}^{\nu_1(k,N(n,x))} a(j)}, \tag{2.6}$$

$$\lim_{n\to\infty} \frac{\sum_{j=\nu_2(i,n)}^{\nu_2(i,N'(n,x))} b(j)}{\sum_{j=\nu_2(k,n)}^{\nu_2(k,N'(n,x))} b(j)} \tag{2.7}$$

exist a.s. (Together, these conditions imply that the components are updated "comparably often" in an "evenly spread" manner.) In section 6 where we consider learning algorithms for the average cost problems, we strengthen this to $\nu_2(j,n) = n$, i.e., $\bar{Y}_n = I'$.

3. Algorithms for MDPs.

3.1. MDPs. We shall consider MDPs on a finite state space $S = \{1, 2, \ldots, s\}$ with a finite action space $A = \{a_0, \ldots, a_r\}$. $\mathcal{P}(A)$ will denote the set of probability vectors on A. Also given is a transition function $p : S \times S \times A \to [0, 1]$ satisfying

$$\sum_j p(i, j, a) = 1 \ \forall i, a.$$

The MDP is an S-valued process $\{X_n\}$ satisfying a.s.

$$P(X_{n+1} = j | X_k, Z_k, k \le n) = p(X_n, j, Z_n), \qquad n \ge 0,$$

where $\{Z_n\}$ is an A-valued "control" process. If $Z_n = v(X_n)\ \forall n$ for some $v : S \to A$, call $\{Z_n\}$ or, by abuse of terminology, the map v itself, a stationary policy. More generally, if for each n, Z_n is conditionally independent of $\{X_m, Z_m, m < n\}$, given X_n, and Z_n has the same conditional law given X_n for each n, then we call it a stationary randomized policy and identify it with the map $\varphi : S \to \mathcal{P}(A)$ which gives the conditional law of Z_n given X_n. For $i \in S$, $a \in A$, let $\pi(i, a)$ denote the ath component of $\varphi(i)$. Under a stationary policy v (respectively, a stationary randomized policy φ), $\{X_n\}$ is a time-homogeneous Markov chain with transition probabilities $[[p(i, j, v(i))]]$ (respectively, $[[q(i, j, \varphi(i))]]$ where $q(i, j, \varphi(i)) = \sum_a p(i, j, a)\pi(i, a)$). By a further abuse of terminology, we identify the stationary randomized policy φ with the vector $\pi = [\pi(i, a)]$, where the elements are ordered lexicographically. Let $k : S \times A \to R$ be a prescribed "running cost" function and $\alpha \in (0, 1)$ a "discount factor." The two control problems we consider follow in sections 3.2 and 3.3.

3.2. The infinite horizon discounted cost problem. Here the aim is to minimize over all admissible $\{Z_n\}$ the quantity $\mathrm{E}[\sum_{n=0}^{\infty} \alpha^n k(X_n, Z_n)]$. Define the value function $V_\alpha : S \to R$ by

$$V_\alpha(i) = \inf \mathrm{E}\left[\sum_{n=0}^{\infty} \alpha^n k(X_n, Z_n) | X_0 = i\right],$$

where the infimum is over all admissible $\{Z_n\}$. Then $V_\alpha(\cdot)$ is the unique solution to the equation

$$V_\alpha(i) = \min_a \left[k(i, a) + \alpha \sum_j p(i, j, a) V_\alpha(j)\right], \qquad i \in S, \tag{3.1}$$

and $\{Z_n\}$ is optimal if and only if $Z_n \in \mathrm{Arg\ min}(k(X_n, \cdot) + \alpha \sum_j p(X_n, j, \cdot) V_\alpha(j))$ a.s. $\forall n$. In particular, a stationary policy v (respectively, stationary randomized policy φ) is optimal if and only if $v(i) \in \mathrm{Arg\ min}(k(i, \cdot) + \alpha \sum_j p(i, j, \cdot) V_\alpha(j))$ (respectively, $\mathrm{support}(\varphi(i)) \subset \mathrm{Arg\ min}(k(i, \cdot) + \alpha \sum_j p(i, j, \cdot) V_\alpha(j)))\ \forall i$ that are visited with positive probability. Thus the existence of a stationary policy $v(\cdot)$ follows, as well as a recipe for finding it, viz., by minimizing the right-hand side (RHS) of (3.1) to find $v(i)$; see [21] for this and related results.

For future reference, we also associate with a stationary randomized policy $\varphi = [[\pi(i, a)]]$ a "stationary value function" $V_\pi : S \to R$ defined by

$$V_\pi(i) = \mathrm{E}\left[\sum_{n=0}^{\infty} \alpha^n k(X_n, Z_n) | X_0 = i\right], \qquad i \in S,$$

where $\{Z_n\} \approx \varphi$. This is easily seen to be the unique solution of the linear equation

$$V_\pi(i) = \sum_a \pi(i,a) \left[k(i,a) + \alpha \sum_j p(i,j,a) V_\pi(j) \right], \qquad i \in S. \tag{3.2}$$

3.3. Average expected cost problem. Here the aim is to minimize

$$\limsup_{N\to\infty} \frac{1}{N} \mathrm{E} \left[\sum_{n=0}^{N-1} k(X_n, Z_n) \right]$$

over all admissible $\{Z_n\}$. We assume that $\{X_n\}$ is an irreducible Markov chain under every stationary policy. Under this condition, it can be shown that there exists a pair $(h(\cdot), \lambda^*)$, $h : S \to R$, $\lambda^* \in R$, satisfying

$$\lambda^* + h(i) = \min_a \left[k(i,a) + \sum_j p(i,j,a) h(j) \right]. \tag{3.3}$$

Furthermore, λ^* is uniquely specified as the optimal cost, equal to the cost of an optimal stationary policy guaranteed to exist and characterized by $v(i) \in \mathrm{Arg}\ \min(k(i,\cdot) + \sum_j p(i,j,\cdot) h(j))$. A stationary randomized policy φ is optimal if and only if support of $\varphi(i)$ lies in the above Argmin. The vector h is unique up to an additive constant; i.e., if h, h' satisfy (3.3), then $h(i) - h'(i)$ is independent of i. We shall denote by h^* the unique solution satisfying $h^*(i_0) = \lambda^*$ for a prescribed $i_0 \in S$. For future reference, we also associate with a stationary randomized policy $\varphi = [[\pi(i,a)]]$ the corresponding cost λ_π and the function $h_\pi : S \to R$ uniquely specified by $h_\pi(i_0) = \lambda_\pi$ and

$$\lambda_\pi + h_\pi(i) = \sum_a \pi(i,a) \left[k(i,a) + \sum_j p(i,j,a) h_\pi(j) \right], \qquad i \in S.$$

3.4. Policy iteration. A popular iterative scheme for solving (3.1) or (3.3) is policy iteration, known to converge to the desired solution in finitely many steps [21]. Consider (3.1). One starts with an initial guess for optimal stationary policy, say, $v_0(\cdot)$. At nth iteration, one performs the following two steps: Given current candidate $v_n(\cdot)$ for the stationary optimal policy,

Step 1. Compute the stationary value function $V_n \triangleq V_{v_n}$.

Step 2. Pick $v_{n+1}(i) \in \mathrm{Argmin}(k(i,\cdot) + \alpha \sum_j p(i,j,\cdot) V_n(j))$.

Step 1 involves solution of a linear system. We may replace it by an "inner loop" that computes V_n iteratively via a "stationary value iteration": Step 1′: $V_n^{m+1}(i) = k(i, v_n(i)) + \alpha \sum_j p(i,j,v_n(i)) V_n^m(j)$ for $m = 0, 1, 2, \ldots$.

This entails that Step 2 wait till the inner loop iterations converge. Among other things, the algorithms we propose work around this difficulty by using two time scales.

The policy iteration for the average expected cost reads exactly the same except that the linear system in question now is

$$h_n(i) + h_n(i_0) = k(i, v_n(i)) + \sum_j p(i,j,v_n(i)) h_n(j), \qquad i \in S,$$

to be solved for $h_n(\cdot)$. An iterative scheme for this, replacing Step 1 above, is

$$h_n^{m+1}(i) = k(i, v_n(i)) + \sum_j p(i, j, v_n(i)) h_n^m(j) - h_n^m(i_0),$$

where $i_0 \in S$ is a prescribed state as before. Step 2 is as before with $\alpha = 1$. This is the "relative value iteration" scheme [21]. It is known to converge to h^*.

3.5. Learning algorithms for MDPs. In this section, we introduce simulation-based learning algorithms based on policy iteration. As already mentioned, the key point in simulation-based algorithms is that the transition probabilities $p(i, \cdot, a)$ are not available, but it is possible to simulate a transition according to any of these probabilities. Thus one replaces the conditional average of candidate V (or h, as the case may be) with respect to a given transition probability by an evaluation of the same function at the state reached by a simulated transition as per the said probability. Nevertheless, in order for the algorithm to mimic the classical iterative algorithms of the previous subsection, it must "see" the appropriate conditional average through an averaging effect. This is achieved by using a stochastic approximationlike step size schedule.

The "learning" variants of policy iteration were introduced by Barto, Sutton, and Anderson [3] and dubbed "adaptive critic" algorithms. Our formulations differ in that we explicitly use two time scales to simulate the two (i.e., inner and outer) loops of iterations. The outer loop operates on a slower scale and thus sees the inner loop as essentially equilibrated, while the inner (fast) loop sees the outer one as quasi-static. In addition, we allow for a distributed asynchronous implementation, as is usually the case in practice. All in all, our algorithms will fit the framework of the "general algorithm" discussed in the last section.

We shall denote by $V_n(\cdot)$, $\pi_n(\cdot, \cdot)$ the current estimate for the value function and optimal stationary randomized policy, respectively. Let I_1, I_2 be a collection of nonempty subsets of S, $S \times (A \backslash \{a_0\})$, respectively. Let $\{Y_n\}$, $\{Z_n\}$ be I_1, I_2-valued random processes, respectively, with the interpretation that $Y_n =$ the set of $i \in S$ for which $V_n(i)$ gets updated at time n, $Z_n =$ the set of $(i, a) \in S \times (A \backslash \{a_0\})$ for which $\pi_n(i, a)$ gets updated at time n. (Note that $\pi_n(i, a_0) = 1 - \sum_{a \neq a_0} \pi_n(i, a)$.) Define $\nu_1(i, n)$, $\nu_2(i, a, n)$ by

$$\nu_1(i, 0) = 0, \qquad \nu_1(i, n) = \sum_{m=0}^{n-1} I\{i \in Y_m\}, \qquad n > 0,$$

$$\nu(i, a, 0) = 0, \qquad \nu_2(i, a, n) = \sum_{m=0}^{n-1} I\{(i, a) \in Z_m\}, \qquad n > 0.$$

We assume the following throughout the counterpart of (A6). For a deterministic $\Delta > 0$,

$$\liminf_{n \to \infty} \frac{\nu_1(i, n)}{n} \geq \Delta, \qquad \text{a.s. } \forall i \in S,$$

$$\liminf_{n \to \infty} \frac{\nu_2(i, a, n)}{n} \geq \Delta, \qquad \text{a.s. } \forall i \in S, \quad a \in A.$$

Also, the appropriate analogues of (2.6), (2.7) are assumed to hold.

Introduce delays $\tau_n(i, j)$, $\bar{\tau}_n(i, j)$, $\hat{\tau}_n(i)$. The processor updating $V_n(i)$ receives at time n, $V_{n - \tau_n(i,j)}(j)$ instead of $V_n(j)$ and $\pi_{n - \hat{\tau}_n(i)}(i, a)$ instead of $\pi_n(i, a)$. Similarly

the processor updating $\pi_n(i,\cdot)$ at time n receives $V_{n-\bar\tau_n(i,j)}(j)$ instead of $V_n(j)$. These delays are assumed to be bounded a.s. by a deterministic $T > 0$. No messages are lost. Thus, for example, for $j \neq i$, each $\{V_m(j)\}$ is received by the processor that updated $V_m(i)$, albeit with a delay of at most T and not necessarily in the order sent. Let $\{\xi_n(i,a)\}$, $\{\eta_n(i,a)\}$, $i \in S$, $a \in A$, be independent families of independently and identically distributed (i.i.d.) random variables with law $p(i,\cdot,a)$. Let $\varphi_n(i)$ be an A-valued random variable whose conditional law given $\mathcal{F}_n \stackrel{\Delta}{=} \sigma(V_m(\cdot), \pi_m(\cdot), \xi_n(\cdot,\cdot), \eta_m(\cdot,\cdot), \tau_m(\cdot,\cdot), \bar\tau_m(\cdot,\cdot), \hat\tau_m(\cdot),\ m \le n)$ is $\pi_{n-\hat\tau_n(i)}(i,\cdot)$. (Note that this is available to the processor updating $V_n(i)$ at time n.)

Let $PS \subset R^r$ be the simplex defined by

$$PS = \left\{ (x_1,\dots,x_r) : x_i \ge 0; \forall i = 1,\dots,r \text{ and } \sum_{i=1}^r x_i \le 1 \right\}.$$

Let P be the projection from R^r to PS. A stationary randomized policy π can be identified with a $\hat\pi \in PS^s$ with $\hat\pi = [[\pi(i,a)]]$, $i \in S$, $a \in A\backslash\{a_0\}$ (because $\pi(i,a_0)$ gets automatically specified $\forall i$). In particular $\pi \approx \hat\pi$ is thus an element in $(PS)^s$. We shall use $\hat\pi(i)$ and $\hat\pi(i,\cdot)$ interchangeably. The algorithms update $\hat\pi$ rather than π. In what follows, $\{a(n)\}$, $\{b(n)\}$ are as in the preceding section. The updating scheme for V_n is the same in Algorithms 1–3 and therefore is mentioned only once.

3.6. Algorithm 1.

$$\begin{aligned} V_{n+1}(i) = V_n(i) &+ b(\nu_1(i,n))[k(i,\varphi_n(i)) + \alpha V_{n-\tau_n(i,\xi_n(i,\varphi_n(i)))}(\xi_n(i,\varphi_n(i))) \\ &- V_n(i)]I\{i \in Y_n\}, \end{aligned}$$

$$\begin{aligned} \hat\pi_{n+1}(i) = P\Bigg(\hat\pi_n(i) + \sum_{l=1}^r a(\nu_2(i,a_l,n))[k(i,a_0) - k(i,a_l) + \alpha(V_{n-\bar\tau(i,\eta_n(i,a_0))}(\eta_n(i,a_0)) \\ - V_{n-\bar\tau(i,\eta_n(i,a_l))}(\eta_n(i,a_l)))]I\{(i,a_l) \in Z_n\}e_l \Bigg), \end{aligned}$$

where e_l is the unit vector in lth coordinate direction.

3.7. Algorithm 2. Let $\{\psi_{il}(n)\}_{i\in S, a_l \in A}$ denote zero mean i.i.d. noise satisfying the following: For $M_n(i,l) \stackrel{\Delta}{=} \alpha\pi_n(i,a_l)(\sum_k p(i,k,a_l)V_n(k) - V_{n-\bar\tau(i,\eta_n(i,a_l))}(\eta_n(i,a_l)))$, the following holds:

$$\inf_{\theta\in R^r, \|\theta\|=1} \mathrm{E}\left[\max\left(0, \left\langle \theta, \sum_l (M_n(i,l) + \psi_{il}(n))e_l \right\rangle \right)\right] > 0.$$

This condition is difficult to verify. Crudely speaking, it requires that the noise intensity uniformly dominate the effect of the "error" term $M_n(\cdot,\cdot)$ in all directions. In simulations, simple ad hoc "noise" schemes seem to work well. We shall have more to say on this in the following section. The algorithm is as follows: $\{V_n\}$ is updated as above and

$$\begin{aligned} \hat\pi_{n+1}(i) = P\Bigg(\hat\pi_n(i) + \sum_{l=1}^r a(\nu_2(i,a_l,n))[(V_{n-\bar\tau_n(i,i)}(i) - k(i,a_l) \\ - \alpha V_{n-\bar\tau_n(i,\eta_n(i,a_l))}(\eta_n(i,a_l)))\hat\pi_n(i,a_l) + \psi_{il}(n)]I\{(i,a_l) \in Z_n\}e_l \Bigg). \end{aligned}$$

3.8. Algorithm 3. In this, we parametrize $\pi(\cdot,\cdot)$ by parameters $\beta(\cdot,\cdot)$ according to

$$\pi_\beta(i,a) = \frac{\exp(\beta(i,a))}{\sum_{a'} \exp(\beta(i,a'))}, \quad i \in S, \quad a \in A.$$

Each $\beta(i,a)$ is updated by a separate processor. The setup is as before except that the conditional distribution of $\varphi_n(i)$ now is

$$\frac{\exp(\beta_{n-\hat{\tau}_n(i)}(i,\cdot))}{\sum_a \exp(\beta_{n-\hat{\tau}_n(i)}(i,a))}$$

and Z_n is an $S \times A$-valued process.

The algorithm is as follows: $\{V_n\}$ is updated as above and

$$\begin{aligned}\beta_{n+1}(i,a) = P_{\beta_0}(\beta_n(i,a) + a(\nu_2(i,a,n))[V_{n-\bar{\tau}(i,i)}(i) - k(i,a) \\ - \alpha V_{n-\bar{\tau}(i,\eta_n(i,a))}(\eta_n(i,a))]I\{(i,a) \in Z_n\}),\end{aligned}$$

where $P_{\beta_0} : R \to R$, $\beta_0 > 0$ is defined by

$$P_{\beta_0}(x) = \begin{cases} -\beta_0 & \text{for} \quad x \le -\beta_0, \\ x & \text{for} \quad -\beta_0 < x < \beta_0, \\ \beta_0 & \text{otherwise.} \end{cases} \tag{3.4}$$

The next three algorithms deal with the average expected cost problem and are exact counterparts of the above. The delays are defined as before, with h_n replacing V_n. Because of some technical difficulties that will become apparent later in section 5, we require all components of $\{h_n\}$ to sit on the same processor and get updated synchronously. The algorithms are as follows.

3.9. Algorithm 4.

$$h_{n+1}(i) = h_n(i) + b(n)[k(i,\varphi_n(i)) + h_n(\xi_n(i,\varphi_n(i))) - h_n(i) - h_n(i_0)], \tag{3.5}$$

$$\begin{aligned}\hat{\pi}_{n+1}(i) = P\Bigg(\hat{\pi}_n(i) + \sum_{l=1}^{r} a(\nu_2(i,a_l,n))[k(i,a_0) - k(i,a_l) + (h_{n-\bar{\tau}(i,\eta_n(i,a_0))}(\eta_n(i,a_0)) \\ - h_{n-\bar{\tau}(i,\eta_n(i,a_l))}(\eta_n(i,a_l)))]I\{(i,a_l) \in Z_n\}e_l\Bigg).\end{aligned}$$

3.10. Algorithm 5. $\{h_n\}$ is updated as above and

$$\begin{aligned}\hat{\pi}_{n+1}(i) = P\Bigg(\hat{\pi}_n(i) + \sum_{l=1}^{r} a(\nu_2(i,a_l,n))[(h_{n-\bar{\tau}_n(i,i_0)}(i_0) + h_{n-\bar{\tau}(i,i)}(i) - k(i,a_l) \\ - h_{n-\bar{\tau}(i,\eta_n(i,a_l))}(\eta_n(i,a_l)))\hat{\pi}_n(i,a_l) + \psi_{il}(n)]I\{(i,a_l) \in Z_n\}e_l\Bigg).\end{aligned}$$

3.11. Algorithm 6. Here π is parameterized by $\beta(\cdot,\cdot)$ as before. $\{h_n\}$ is updated as above and

$$\beta_{n+1}(i,a) = P_{\beta_0}(\beta_n(i,a) + a(\nu_2(i,a,n))[h_n(i_0) + h_{n-\bar{\tau}(i,i)}(i) - k(i,a) - h_{n-\bar{\tau}(i,\eta_n(i,a))}(\eta_n(i,a))]I\{(i,a) \in Z_n\}).$$

We now briefly describe the intuition behind the three schemes for updating $\hat{\pi}(\cdot)$. Note that ideally $\hat{\pi}$ should concentrate on Argmin of the quantity in square brackets in (3.1), which we call the "Q-value." This suggests that $\hat{\pi}$ should be incrementally adjusted in the direction of lowering the average Q-value with respect to $\hat{\pi}$. This is achieved by comparing Q-values of, say, (i,a) with that of (i,a_0) and using the difference as the "reinforcement signal" for adjusting $\hat{\pi}(i,a)$ in the appropriate direction. (Note that $\hat{\pi}(i,a_0)$ gets automatically adjusted.) This is the philosophy behind Algorithms 1 and 4.

In Algorithms 2 and 5, on the other hand, we take as reinforcement signal the offset between value function and the Q-value, keeping in mind that the former should ideally be the minimum of the latter over a. The additional multiplication by $\hat{\pi}(i,a)$ on the RHS ensures that the iterates "sort of" remain within the desired simplex. (They do exactly so in an asymptotic sense.)

Algorithms 3 and 6 are different in a crucial way. They parametrize $\hat{\pi}$ by hypothesizing a Gibbsian form parametrized by a parameter vector and update the latter rather than $\hat{\pi}$ itself. This search is on a restricted subdomain of the simplex; thus only near optimality can be hoped for. But this parametric form leads to simple algorithms with the possibility of alleviating the "curse of dimensionality" by interfacing with low-dimensional parametrizations.

4. Convergence analysis for the general algorithm.

4.1. Preliminary lemmas. This section sets forth some key technical results underlying the convergence analysis to follow. To start with, consider the ODE

$$\dot{x} = h(x(t),t), \qquad t \geq 0, \tag{4.1}$$

with $h : R^d \times [0,\infty) \to R^d$ measurable and the set $J \stackrel{\Delta}{=} \{x : h(x,t) = 0\}$ compact and independent of t. A continuously differentiable function $V : R^d \to R$ is said to be a strict Liapunov function for (4.1) if

1. $V(x) \to \infty$ as $\|x\| \to \infty$, uniformly in $\|x\|$;

2. $H(x,t) \stackrel{\Delta}{=} \nabla V(x) \cdot h(x,t) \leq 0 \ \forall x,t$ with equality holding only when $x \in J$.

Define $J^\epsilon = \{x : \|x - y\| < \epsilon \text{ for some } y \in J\}$, $\epsilon > 0$. We further assume that $\inf_{t\geq 0, x\notin B} |H(x,t)|$ is nonzero for every open set B containing J.

DEFINITION 4.1. *Given T, $\delta > 0$, a (T,δ)-perturbation of ODE* (4.1) *is a bounded measurable function $y : [0,\infty) \to R^d$ such that there exist $0 = T_0 < T_1 < \cdots T_n \uparrow \infty$ and solutions $x^j(t)$, $t \in [T'_j, T'_{j+1}]$, $j \geq 0$, of* (4.1) *such that $T'_{j+1} - T'_j = T_{j+1} - T_j \geq T$ for $j \geq 0$ and $\|y(t) - x^j(T'_j + t - T_j)\| < \delta$, $T_j \leq t \leq T_{j+1}$, $j \geq 0$.*

The following is a small extension of [13, Theorem 1, p. 339].

LEMMA 4.2. *For any T, $\epsilon > 0$, there exists a $\delta_0 = \delta_0(T,\epsilon) > 0$ such that for any $\delta \in (0,\delta_0)$, any (T,δ)-perturbation of* (4.1) *will converge to J^ϵ.*

Proof. Let $y(\cdot)$ be a (T,δ)-perturbation of (4.1). For $\eta > 0$, define

$$B(\eta) = \{x : |V(x) - V(y)| < \eta \text{ for some } y \in J\}.$$

Let B be a ball large enough so that it contains both $\{y(t), t \geq 0\}$ and $B(\eta)$ for some η chosen so that $B(\eta) \subset J^\epsilon$. Define K, Δ by

$$K = \sup_{x \in \bar{B}} \|\nabla V(x)\|, \qquad \Delta = \inf_{t \geq 0} \inf_{x \in \bar{B} \setminus B(\eta)} |H(x,t)|.$$

Call the restriction of $y(\cdot)$ to $[T_j, T_{j+1}]$ the jth patch of $y(\cdot)$. Let $\delta < \frac{\Delta T}{4K}$. It is easily seen that $V(y(T_j)) - V(y(T_{j+1})) + 2\delta K \geq V(x^j(T'_j)) - V(x^j(T'_{j+1})) \geq \Delta T$ whenever $x^j(t)$, $t \in [T'_j, T'_{j+1}]$, does not intersect $B(\eta)$. But then

$$V(y(T_j)) - V(y(T_{j+1})) \geq \Delta T/2. \tag{4.2}$$

This can occur for at most finitely many consecutive j's because $V(\cdot)$ is bounded from below. Also, if the jth patch of $y(\cdot)$ does not intersect $B(\eta + \delta K)$, $x^j(\cdot)$ cannot intersect $B(\eta)$. We conclude that eventually $x^j(\cdot)$ and the jth patch of $y(\cdot)$ must intersect $B(\eta)$, $B(\eta + \delta K)$, respectively. Now, for all j,

$$V(x^j(t)) \leq V(x^j(s)) \ \forall s, \quad t \in [T'_j, T'_{j+1}], \quad t \geq s.$$

So

$$V(y(t)) \leq V(y(s)) + 2\delta K \ \forall s, \quad t \in [T_j, T_{j+1}], \quad t \geq s.$$

Therefore, the patch of $y(\cdot)$ that intersects $B(\eta + \delta K)$ remains in $B(\eta + \delta K + 2\delta K)$ after hitting $B(\eta + \delta K)$. Since $2\delta K < \Delta T/2$, (4.2) implies that the subsequent patch must hit $B(\eta + \delta K)$. It follows that $y(\cdot)$ remains in $B(\eta + \delta K + 2\delta K)$ once it hits $B(\eta + \delta K)$. Pick η, δ sufficiently small so that $B(\eta + \delta K + 2\delta K) \subset J^\epsilon$. This completes the proof. □

COROLLARY 4.3. *For $X(\cdot) \in C([0,\infty); R^d)$, $0 \leq t(n) \uparrow \infty$, if $\{X(t(n) + \cdot)\}$ is conditionally compact in $C([0,\infty); R^d)$ and all the limit points of $\{X(t(n) + \cdot)\}$ in $C([0,\infty); R^d)$ as $n \to \infty$ satisfy (4.1), then $X(t)$ converges to J as $t \to \infty$.*

Proof. Let $T > 0$. Let $\delta > 0$ be such that $X(t(n) + \cdot)$ is not a (T, δ)-perturbation of (4.1) for any n. Then there exists sequence $n_1, n_2, \ldots, n_i \uparrow \infty$ such that

$$\sup_{t \in [0,t]} \|X(t(n_i) + t) - z(t)\| \geq \delta \ \forall i > 0,$$

for all $z(\cdot)$ satisfying ODE (4.1). This means that there cannot exist a subsequence of $\{X^{n_i}(\cdot)\}$ that converges to a solution of ODE (4.1). Thus for all $\delta > 0$, there exists n sufficiently large such that $X(t(n) + \cdot)$ is a (T, δ)-perturbation of (4.1). This, by Lemma 4.2, implies that $X(t)$ converges to J as $t \to \infty$. □

Next define U^d as the space of $[0,1]^d$-valued trajectories $\bar{\mu} = \{\mu_t, t \geq 0\}$ with the coarsest topology that renders continuous the maps $\bar{\mu} \mapsto \int_0^T h(t)\mu_t(i)dt$ for $T > 0$, $1 \leq i \leq d$, $h \in L_2[0,T]$. Using the Banach–Alaoglu theorem, it is easily verified that U^d is compact metrizable and, hence, Polish. For $\bar{\mu} \in U^d$ let $M^{\bar{\mu}}(t)$ denote the diagonal matrix $\text{diag}(\mu_t(1), \ldots, \mu_t(d))$, $t \geq 0$. For $\mu \in [0,1]^d$ define $M^\mu = \text{diag}(\mu(1), \ldots, \mu(d))$.

LEMMA 4.4. *Let $\{X^n(\cdot)\} \subset C([0,\infty); R^d)$ be uniformly bounded with $X^n(\cdot) \to X^\infty(\cdot)$ and $\{\bar{\mu}^n\} \subset U^d$ such that $\bar{\mu}^n \to \bar{\mu}^\infty$. Then for every $h \in C(R^d; R^d)$,*

$$\lim_{n \to \infty} \int_0^t M^{\bar{\mu}^n}(s) h(X^n(s)) ds = \int_0^t M^{\bar{\mu}^\infty}(s) h(X^\infty(s)) ds.$$

Proof.

$$\begin{aligned}
&\left\| \int_0^t M^{\bar{\mu}^\infty}(s) h(X^\infty(s)) ds - \int_0^t M^{\bar{\mu}^n}(s) h(X^n(s)) ds \right\| \\
&\quad \leq \left\| \int_0^t M^{\bar{\mu}^n}(s) (h(X^\infty(s)) - h(X^n(s))) ds \right\| \\
&\qquad + \left\| \int_0^t M^{\bar{\mu}^\infty}(s) h(X^\infty(s)) ds - \int_0^t M^{\bar{\mu}^n}(s) h(X^\infty(s)) ds \right\| \\
&\quad \leq \int_0^t \| h(X^\infty(s)) - h(X^n(s)) \| ds \\
&\qquad + \left\| \int_0^t M^{\bar{\mu}^\infty}(s) h(X^\infty(s)) ds - \int_0^t M^{\bar{\mu}^n}(s) h(X^\infty(s)) ds \right\|.
\end{aligned}$$

Given our hypotheses, it is easy to see that both terms on the right go to zero. ∎

The last lemma of this subsection concerns the noise sequence.

LEMMA 4.5. $\sum_n a(\nu_1(i,n)) M_i(n) I\{i \in \bar{X}_n\}$, $\sum_n b(\nu_2(j,n)) M'_j(n) I\{j \in \bar{Y}_n\}$ *converge a.s.*

Proof. Note that in each case the partial sums form a square integrable $\{\mathcal{F}_n\}$-martingale. By (A4) and the fact that $\forall i, j$,

$$\sum_n a(\nu_1(i,n))^2 I\{i \in \bar{X} n\}, \qquad \sum_n b(\nu_2(j,n))^2 I\{j \in \bar{Y}_n\} < \infty \text{ a.s.},$$

it follows that the quadratic variation process for these martingales remain bounded a.s. The claim now follows from [18, Proposition VII-3(c), pp. 149–150]. ∎

4.2. Convergence analysis. We shall proceed through a sequence of lemmas. Define

$$\begin{aligned}
\bar{b}(n) &= \max_{i \in \bar{Y}_n} b(\nu_2(i,n)), \\
\bar{a}(n) &= \max_{j \in \bar{X}_n} a(\nu_1(j,n)).
\end{aligned}$$

Let $t(0) = 0$, $t(n) = \sum_{k=0}^{n-1} \bar{b}(k)$, $n > 0$ and define sequences $\{\mu'_n\}$, $\{\mu_n\}$ in $[0,1]^m$, $[0,1]^l$ by the following: For $n \geq 0$,

$$\begin{aligned}
\mu'_n(i) &= b(\nu_2(i,n)) I\{i \in \bar{Y}_n\} / \bar{b}(n), \qquad 1 \leq i \leq m, \\
\mu_n(j) &= a(\nu_1(j,n)) I\{j \in \bar{X}_n\} / \bar{a}(n)
\end{aligned}$$

for $1 \leq j \leq l$. Then rewrite the algorithm (2.2)–(2.3) as

$$X(n+1) = P(X(n) + \bar{a}(n) M^{\mu_n}(W(n) + M(n))), \tag{4.3}$$

$$Y(n+1) = Y(n) + \bar{b}(n) M^{\mu'_n}(W'(n) + M'(n)), \tag{4.4}$$

where $W_i(n) = g_i(\tilde{X}^i(n), \tilde{Y}^i(n))$, $W'_j(n) = f_j(\hat{X}^j(n), \hat{Y}^j(n))$ for $1 \leq i \leq m$, $1 \leq j \leq l$.

LEMMA 4.6. $\bar{b}(n), \bar{a}(n) \to 0$ *and* $\bar{a}(n) = o(\bar{b}(n))$ *a.s.*

Proof. By assumption (A1) $\hat{a} = \sup_n \sup_{y \in [x,1]} a([yn])/a(n)$ is finite $\forall x \in (0,1]$. Let $\bar{\eta} > \epsilon > 0$, where $\bar{\eta}$ is as in (A6). By (A6), a.s., $\nu_1(i,n)/n > \bar{\eta} - \epsilon$ eventually.

Setting $x = \bar{\eta} - \epsilon$ in the above, $a(\nu_1(i,n))/a(n) \leq \hat{a} + \delta$ eventually a.s. for any $\delta > 0$. Since $a(n) \to 0$ it follows that $\bar{a}(n) \to 0$; $\bar{b}(n) \to 0$ is proved similarly. Now, for large n,

$$\frac{\bar{a}(n)}{\bar{b}(n)} \leq \frac{a([n(\bar{\eta} - \epsilon)])}{b(n)} \leq \hat{a}\frac{a(n)}{b(n)}$$

by (A1). □

LEMMA 4.7. $\|g(X(n), Y(n)) - W(n)\|, \|f(X(n), Y(n)) - W'(n)\| \to 0$ *a.s.*

Proof. Let C be a bound on the $\{g_i(X^i(n), Y^i(n))\}$ (cf. (A2)). For D as in (A5) and $1 \leq i \leq K$, we have, for $n \geq D$ and suitable $C_1 > 0$ depending on C,

$$\begin{aligned}\|X(n) - \tilde{X}^i(n)\| &\leq C_1 \sum_{k=n-D}^{n} \bar{a}(k) + \sum_{k=n-D}^{n} \|\bar{a}(k)M(k)\| \\ &\leq (D+1)\left(C_1 \sup_{k \geq n-D} \bar{a}(k) + \sup_{k \geq n-D} \|\bar{a}(k)M(k)\|\right).\end{aligned}$$

Using Lemma 4.5 and (A4) we conclude that the r.h.s. above tends to zero a.s. Similarly, one proves that a.s.

$$\|\hat{X}^i(n) - X(n)\|, \quad \|\hat{Y}^i(n) - Y(n)\|, \quad \|\tilde{Y}^j(n) - Y(n)\| \to 0.$$

The claim follows from (A2) and continuity of $f(\cdot,\cdot)$, $g(\cdot,\cdot)$. □

Define $\bar{\mu} \in U^m$ by the following: $\mu_t = \mu'_n$, $t \in [t(n), t(n+1))$, $n \geq 0$. Let $\bar{\mu}^n = \{\mu_{t+t(n)}, t \geq 0\}$. In the following lemma, it is worth keeping in mind that the convergence of $\bar{\mu}^n$'s along a subsequence to μ^* (say) is with respect to the topology on U^m defined above. That is, for each $T > 0$, $f \in L_2[0,T], i \in S$,

$$\int_0^T f(t)\mu_t^n(i)dt \to \int_0^T f(t)\mu_t^*(i)dt.$$

LEMMA 4.8. *Almost surely, every limit point $\bar{\mu}^*$ of $\{\bar{\mu}^n\}$ in U^m satisfies the following: $\mu_t^*(i) = c(t)/m$ almost every t (a.e.t) for some bounded measurable function $c(\cdot) : [0,\infty) \to [1,\infty)$.*

Proof. Let $h(\cdot)$ be an eventually decreasing smooth function $[0,\infty) \to [0,\infty)$ such that $h(t(n)) = b(n)$ $\forall n$. Note that

$$\begin{aligned}\lim_{t\to\infty} \frac{\int_z^t \mu_s(j)ds}{\int_z^t \mu_s(i)ds} &= \lim_{n\to\infty} \frac{\sum_{k=0}^{\nu_1(j,n)} b(k)}{\sum_{k=0}^{\nu_1(i,n)} b(k)} = \lim_{n\to\infty} \frac{\int_0^{\nu_1(j,n)} h(s)ds}{\int_0^{\nu_1(i,n)} h(s)ds} \\ &= \lim_{n\to\infty} \frac{\int_0^{\nu_1(j,n)} h(s)ds}{\int_0^n h(s)ds} \frac{\int_0^n h(s)ds}{\int_0^{\nu_1(i,n)} h(s)ds} \\ &= 1 \ a.s.,\end{aligned}$$

uniformly in z for $z \leq C$ (say), by virtue of (A1)(3) and (A6). Then for $x > 0$,

$$\lim_{t\to\infty} \frac{\int_0^x \int_0^t \mu_{s+y}(j)dsdy}{\int_0^x \int_0^t \mu_{s+y}(i)dsdy} = \lim_{t\to\infty} \frac{\int_0^t \int_0^x \mu_{s+y}(j)dyds}{\int_0^t \int_0^x \mu_{s+y}(i)dyds} = 1 \ a.s.$$

By l'Hôpital's rule,

$$\lim_{t\to\infty} \frac{\int_0^x \mu_{t+y}(j)dy}{\int_0^x \mu_{t+y}(i)dy} = 1 \text{ a.s.}$$

(That this limit is known to exist a priori follows from the second half of assumption (A6).) Thus, a.s., any limit point $\bar{\mu}^*$ of $\{\bar{\mu}^n\}$ must satisfy

$$\frac{\int_0^x \mu_t^*(j)dt}{\int_0^x \mu_t^*(i)dt} = 1 \text{ a.s.}$$

Since x was arbitrary, we have $\mu_t^*(j)/\mu_t^*(i) = 1$ for a.e.t. This along with the fact that $m \geq \sum_i \mu_t(i) \geq 1$ for all t implies the above claim. □

The proofs of the next two lemmas are given in outline, as the details are lengthy but routine.

LEMMA 4.9. $\|\lambda(X(n)) - Y(n)\| \to 0$ a.s.

Proof (sketch). Fix a sample point outside the zero probability set where the foregoing lemmas and assumptions fail. Rewrite the algorithm as

$$X(n+1) = X(n) + \bar{b}(n)\left(\frac{\bar{a}(n)}{\bar{b}(n)}\right) M^{\mu_n}(W(n) + M(n))$$

$$+ \bar{b}(n)\left(\frac{P(X(n) + \bar{a}(n)M^{\mu_n}(W(n) + M(n))) - X(n) - \bar{a}(n)M^{\mu_n}(W(n) + M(n))}{\bar{b}(n)}\right),$$

$$Y(n+1) = Y(n) + \bar{b}(n)M^{\mu_n'}(f(X(n), Y(n)) + ((W'(n) - f(X(n), Y(n))) + M'(n))).$$

Using our definition of $P(\cdot)$, it follows that

$$\left\| \frac{P(X(n) + \bar{a}(n)M^{\mu_n}(W(n) + M(n))) - X(n) - \bar{a}(n)M^{\mu_n}(W(n) + M(n))}{\bar{b}(n)} \right\|$$

$$\leq \left(\frac{\bar{a}(n)}{\bar{b}(n)}\right) \|W(n) + M(n)\|.$$

Since $\|W(n) + M(n)\|$ remains bounded, we have, by Lemma 4.6,

$$X(n+1) = X(n) + \bar{b}(n)Z(n),$$

where $Z(n)$ is $o(1)$. Define $\tilde{X}(\cdot) : [0, \infty) \to R^l$, $\tilde{Y}(\cdot) : [0, \infty) \to R^m$ by $\tilde{X}(t(n)) = X(n)$, $\tilde{Y}(t(n)) = Y(n)$, with linear interpolation on $[t(n), t(n+1)]$, $n \geq 0$. In view of Lemmas 4.4–4.8, a standard argument using the discrete Gronwall inequality shows that for any T, $\delta > 0$, $\tilde{X}(t(n) + \cdot)$, $\tilde{Y}(t(n) + \cdot)$ is, for sufficiently large n, (see, e.g., [9, p. 847]) a (T, δ)-perturbation of the ODE (for a suitable $c : R^+ \to [1, \infty)$)

$$\dot{x}(t) = 0, \qquad \dot{y}(t) = \frac{c(t)}{m} f(x(t), y(t)).$$

By Lemma 4.2, $(X(n), Y(n)) \to \text{graph}(\lambda)$. The claim follows from the uniform continuity of $\lambda(x)$. □

Now define the sequence $s(n) \uparrow \infty$ by $s(0) = 0$, $s(n) = \sum_{k=0}^{n-1} \bar{a}(k)$, $n > 0$. Redefine $\bar{\mu} \in U^l$ by $\mu_t = \mu_n$, $t \in [s(n), s(n+1))$, $n \geq 0$.

LEMMA 4.10. $(X(n), Y(n)) \to \{(x, y) : \bar{P}_x(g(x, \lambda(x))) = 0 \text{ and } y = \lambda(x)\}$ *a.s.*

Proof (sketch). Fix ω outside a zero probability set where any of the foregoing assumptions and lemmas fail. Note that

$$\begin{aligned} X(n+1) = X(n) &+ \bar{a}(n)M^{\mu_n}(g(X(n), \lambda(X(n))) + M(n) \\ &+ ([g(X(n), Y(n)) - g(X(n), \lambda(X(n)))] \\ &+ [W(n) - g(X(n), Y(n))])) + r(n), \end{aligned} \tag{4.5}$$

where

$$r(n) = P(X(n) + \bar{a}(n)M^{\mu_n}(W(n) + M(n))) - (X(n) + \bar{a}(n)M^{\mu_n}(W(n) + M(n))).$$

Let $\tilde{\mu}^n = \{\mu_{s(n)+t}, t \geq 0\}$, $n \geq 0$. It is easily verified that Lemma 4.8 applies to $\{\tilde{\mu}^n\}$ as well. Define $\tilde{X}(\cdot) : [0, \infty) \to R^l$ by $\tilde{X}(s(n)) = X(n)$, with linear interpolation on $[s(n), s(n+1)], n \geq 0$. A standard argument along the lines of [16, 17] using Lemmas 4.4–4.9 and Corollary 4.3 shows that for any T, $\delta > 0$, $\tilde{X}(s(n) + \cdot)$ is a (T, δ)-perturbation of the ODE

$$\dot{x}(t) = \frac{c(t)}{l} \bar{P}_{x(t)}(g(x(t), \lambda(x(t))))$$

for sufficiently large n, where $c(\cdot)$ is a scalar measurable function on $[0, \infty)$ satisfying $c(t) \geq 1$ $\forall t$. (The only differences from [9] are the following: (i) the nonautonomous term $c(t)$ arising from the asymptotics of $\{\mu_{s(n)+\cdot}\}$, which is handled as in the preceding lemma (see [9] for details of this limiting argument); (ii) the additional "error" terms in the square brackets in (4.5), whose contribution is asymptotically negligible in view of the foregoing lemmas; and (iii) the term $r(n)$ due to projection operator which can be handled along the lines of [17, section 5.3]—see [16] for details.) Now invoke Lemma 4.2 to conclude. □

5. Convergence analysis for discounted cost.

5.1. Preliminaries. The convergence analysis of this and the next section rests on verifying that the algorithms in question fit the general model of section 2. For this purpose, we start with the following minor variant of a result from [24]. Consider the following iterations in R^d for computing $X(n)$, $n \geq 0$, $X(\cdot) = [X_1(\cdot), \ldots X_d(\cdot)]^T$: For $1 \leq i \leq d$,

$$X_i(n+1) = X_i(n) + a_i(n)(F_i^n(\tilde{X}^i(n)) - X_i(n) + w_i(n)),$$

where $\tilde{X}^i(n)$ is a d-vector whose jth component is $X_j(n - \tau_{ij}(n))$ for some random delays $\{\tau_{ij}(\cdot)\}$ and $\{w_i(n)\}$ is a stochastic process. We make the following assumptions.

1. For $1 \leq i \leq d$, $\{a_i(n)\} \subset [0, \infty)$ are random with $\sum_n a_i(n) = \infty$ a.s., $\sum_n a_i(n)^2 \leq \bar{C}$ a.s. for a finite constant $\bar{C} > 0$.
2. $\{a_i(n)\}$, $\{w_i(n-1)\}$, $\{\tau_{ij}(n)\}$, $\{X_i(n)\}$ are $\{\mathcal{F}_n\}$-adapted for an increasing family of σ-fields $\{\mathcal{F}_n\}$ with

$$\mathrm{E}[w_i(n)|\mathcal{F}_n] = 0, \qquad \mathrm{E}[w_i^2(n)|\mathcal{F}_n] \leq A + B \max_j \max_{k \leq n} |X_j(k)|^2 \ \forall i,$$

   a.s. for suitable A, $B > 0$.
3. $\|F^n(x)\|_\infty \leq \beta \|x\|_\infty + D$ $\forall x$, for suitable $\beta \in (0, 1)$, $D > 0$ and

$$\|F^n(x) - F^n(y)\|_\infty \leq \bar{\beta} \|x - y\|_\infty \forall x, y$$

   for some $\bar{\beta} \in (0, 1)$.

LEMMA 5.1. *Under the above hypotheses,* $\sup_n \|X(n)\| < \infty$ *a.s.*

This is precisely [24, Theorem 2.1] with one small modification that does not affect its proof: [24] has a fixed $F(\cdot)$ in place of $\{F^n(\cdot)\}$.

COROLLARY 5.2. *In Algorithms* 1–3, $\{V_n\}$ *is bounded a.s.*

Proof. Make the following correspondence with the above lemma:

$$a_i(n) = b(\nu_1(i,n))I\{i \in Y_n\},$$
$$\mathcal{F}_n = \sigma(\xi_m(i,a), \eta_m(i,a), \varphi_m(i), m < n, i \in S, a \in A,$$
$$Y_m, Z_m, V_m, \pi_m, \tau_m(i,j), \bar{\tau}_m(i,j), \hat{\tau}_m(i), m \le n, i,j \in S, a \in A),$$
$$F_i^n(V) = \sum_a \pi_{n-\hat{\tau}_n(i)}(i,a)\left[k(i,a) + \alpha \sum_j p(i,j,a)V(j)\right],$$
$$w_i(n) = [k(i,\varphi_n(i)) + \alpha V_{n-\tau_n(i,\xi_n(i,\varphi_n(i)))}(\xi_n(i,\varphi_n(i))) - V_n(i)]$$
$$- \mathrm{E}[[k(i,\varphi_n(i)) + \alpha V_{n-\tau_n(i,\xi_n(i,\varphi_n(i)))}(\xi_n(i,\varphi_n(i))) - V_n(i)]|\mathcal{F}_n].$$

Note that $w(n)$ corresponds to "$M(n)$" in (A4). The claim follows easily. □

This verifies (A2). Assumptions (A1), (A5), and (A6) are already built in. Assumption (A4) follows easily from the a.s. boundedness of $\{V_n(\cdot)\}$. We now verify (A3). Consider the ODE

$$\dot{V}(t) = F(\pi, V(t)) - V(t), \tag{5.1}$$

where $F(\cdot,\cdot) = [F_1(\cdot,\cdot), \ldots, F_s(\cdot,\cdot)]^T$ is defined by

$$F_i(\pi, x) = \sum_a \pi(i,a)k(i,a) + \alpha \sum_{aj} \pi(i,a)p(i,j,a)x_j$$

for $\pi = [[\pi(i,a)]]$, $i \in S$, $a \in A$ with $\pi(i,\cdot) \in \mathcal{P}(A)$ $\forall i$, $x = [x_1, \ldots, x_s] \in R^s$. Note that π is a fixed parameter in (5.1).

LEMMA 5.3. *V_π is the unique asymptotically stable equilibrium point for* (5.1).

Proof. Direct verification shows that the ODE in question is an asymptotically stable linear system of the form

$$\dot{x}(t) = Ax(t) + B$$

with the desired equilibrium point as its unique globally asymptotically stable equilibrium. □

Note that V_π for given $\pi = [[\pi(i,a)]]$ is obtained by solving a linear system. Writing it explicitly using Cramer's rule, one verifies that it is a C^1 function of $[[\pi(i,a)]]$, in particular Lipschitz. This verifies the first part of (A3). The second part is specific to each algorithm and will be verified separately.

In what follows, we shall say that $\hat{G}(\cdot)$ is a vector field on $(PS)^s$ when it is so with the latter viewed as a compact manifold with boundary (i.e., when the corresponding integral curves do not leave $(PS)^s$). Such a $\hat{G}(\cdot)$ will also define a unique vector field $G(\cdot)$ on $(\mathcal{P}(A))^s$ given by

$$G_{ia}(\pi) = \hat{G}_{ia}(\hat{\pi}) \ \forall i \in S, a \in A \backslash \{a_0\},$$

$$G_{ia_0}(\pi) = -\sum_{a \neq a_0} \hat{G}_{ia}(\pi).$$

Define vectors $G_i = [G_i.], \hat{G}_i = [\hat{G}_i.]$ in R^{r+1}, R^r, respectively. Define $K : (\mathcal{P}(A))^s \to R^{s \times (r+1)}$ by $K(\pi) = [[K_{ia}(\pi)]]$ for $i \in S$, $a \in A$, where

$$K_{ia}(\pi) = k(i,a) + \alpha \sum_j p(i,j,a)V_\pi(j) - V_\pi(i).$$

Note that

$$\sum_a K_{ia}(\pi)\pi(i,a) = 0 \ \forall i \in S. \tag{5.2}$$

Define $\hat{K} : (PS)^s \to R^{s\times r}$ as

$$\hat{K}_{ia}(\hat{\pi}) = K_{ia}(\pi) - K_{ia_0}(\pi), \quad i \in S, \quad a \in A\backslash\{a_0\}$$

and the vectors corresponding to K: $K_i = [K_i.]$, $\hat{K}_i = [\hat{K}_i.]$.

LEMMA 5.4. *If G satisfies*

$$G_i(\pi) \cdot K_i(\pi) \leq 0 \qquad \forall i \in S, \tag{5.3}$$

then

$$D_G V_\pi(i) \leq G_i(\pi) \cdot K_i(\pi) \leq 0 \ \forall i \in S,$$

where D_G represents the directional derivative along G.

Proof. Define the operator $T_\pi : R^s \to R^s$ by $T_\pi(\cdot) = F(\pi, \cdot)$. From the geometry of $\mathcal{P}(A)^s$, it is clear that if G is a vector field on $\mathcal{P}(A)^s$, then for every $\pi \in \text{int}(\mathcal{P}(A)^s)$ there exists $\delta_0(\pi)$ such that for all $0 \leq \delta \leq \delta_0(\pi)$, $\pi + \delta G \in \mathcal{P}(A)^s$. Pick $\delta > 0$ so that $\pi + \delta G \in (\mathcal{P}(A))^s$. Simple algebra, using the facts that $T_\pi V_\pi = V_\pi$ and $\sum_a G_{ia}(\pi) = 0$, leads to

$$T_{\pi+\delta G} V_\pi(i) - V_\pi(i) = \delta \sum_a K_{ia}(\pi) G_{ia}(\pi) \leq 0.$$

Iterating, $T^n_{\pi+\delta G} V_\pi \leq T^{n-1}_{\pi+\delta G} V_\pi \leq \cdots \leq V_\pi$. Also

$$\frac{(T^n_{\pi+\delta G} V_\pi)(i) - V_\pi(i)}{\delta} \leq \sum_a K_{ia}(\pi) G_{ia}(\pi) \text{ for } n > 1.$$

Since $T^n_{\pi'} V_0 \to V_{\pi'}, \forall \pi', V_0$, by the contraction mapping principle, we may let $n \to \infty$ in the above inequality to obtain

$$\frac{V_{\pi+\delta G}(i) - V_\pi(i)}{\delta} \leq \sum_a K_{ia}(\pi) G_{ia}(\pi).$$

For π on the boundary of $\mathcal{P}(A)^s$, the claim follows by continuity. To conclude, we let $\delta \downarrow 0$. □

The lemma can be equivalently stated as follows: if $\hat{G}$ satisfies

$$\hat{G}_i(\hat{\pi}) \cdot \hat{K}_i(\hat{\pi}) \leq 0 \ \forall i \in S,$$

then

$$\nabla V_\pi(i) \cdot \hat{G}(\hat{\pi}) \leq \hat{G}_i(\hat{\pi}) \cdot \hat{K}_i(\hat{\pi}) \leq 0 \ \forall i \in S.$$

COROLLARY 5.5. *If in addition, $\hat{G}_i(\hat{\pi}) \cdot \hat{K}_i(\hat{\pi}) = 0$ for all $i \in S$ only if $\hat{G}(\hat{\pi}) = 0$ (i.e., only on equilibrium points of the vector field), then $\sum_{i\in S} V_\pi(i)$ serves as a strict Liapunov function for ODE $\dot{\hat{\pi}}(t) = G(\hat{\pi}(t))$.*

5.2. Convergence of Algorithm 1. We start with Algorithm 1. The key step is to verify that it fits the format of the "general algorithm" with $(\hat{\pi}_n, V_n)$ replacing $(X(n), Y(n))$. Thus corresponding g is $g = [[g_{ia}]]$ with

$$g_{ia}(\hat{\pi}, V) = k(i, a_0) - k(i, a) + \alpha \sum_j (p(i, j, a_0) - p(i, j, a))V(j), \quad i \in S, \quad a \in A\backslash\{a_0\},$$

with $M(n), M'(n)$ appropriately defined. (Note that this does not depend on π.) To verify the rest of (A3), consider the ODE

$$\dot{\hat{\pi}}(i) = \bar{P}_{\hat{\pi}}(g_i(\hat{\pi}, V_\pi)), \qquad i \in S, \tag{5.4}$$

with

$$g_i(\hat{\pi}, V_\pi) = [g_i.(\hat{\pi}, V_\pi n)],$$

$$\bar{P}_{\hat{\pi}}(g_i(\hat{\pi}_i, V_\pi)) = \lim_{\Delta \downarrow 0} \frac{P(\hat{\pi}(i) + \Delta g_i(\hat{\pi}, V_\pi)) - \hat{\pi}(i)}{\Delta},$$

where P is the projection into PS. Write K, $\hat{K}$ for $K(\pi)$, $\hat{K}(\hat{\pi})$, respectively, for the sake of simplicity.

LEMMA 5.6. *$\sum_{i \in S} V_\pi(i)$ is a strict Liapunov functions for ODE* (5.4).

Proof. Make the correspondence $\hat{G}_i = \bar{P}_{\hat{\pi}}(g_i(\hat{\pi}, V_\pi)) = g_i(\hat{\pi}, V_\pi) + r_i(\hat{\pi})$ $\forall i$, where $r_i(\cdot)$ is the correction term at boundary due to projection. Note that $\|g_i(\hat{\pi}, V_\pi)\| \geq \|r_i(\hat{\pi})\|$ $\forall i$ with equality holding only when $g_i(\hat{\pi}, V_\pi) = -r_i(\hat{\pi})$. Also $g_i(\hat{\pi}, V_\pi) = -\hat{K}_i(\hat{\pi})$ by definition. Thus $\forall i$,

$$\begin{aligned} \hat{G}_i \cdot \hat{K}_i &= g_i \cdot \hat{K}_i + r_i \cdot \hat{K}_i \\ &= -\|\hat{K}_i\|^2 + r_i \cdot \hat{K}_i \\ &\leq -\|\hat{K}_i\|^2 + \|r_i\| \, \|\hat{K}_i\| \\ &\leq 0 \end{aligned}$$

with equality if and only if $g_i = -r_i$. This means that $D_G V_\pi(i) = 0$ for all $i \in S$ only on equilibrium points of (5.4). The claim follows (cf. Corollary 5.5). □

LEMMA 5.7. *If $\hat{\pi}$ is an equilibrium point of* (5.4)*, then $V_\pi = V_\alpha$.*

Proof. Fix $i \in S$. Let $KT(i)$ denote $\{\hat{\pi} \in PS^s : g_i(\hat{\pi}, V_\pi) = -r_i(\hat{\pi})\}$ (the "Kuhn–Tucker points") and $e_a \in R^r$ the unit vector whose elements are indexed by elements of $A\backslash\{a_0\}$ and are zero, except for the one with index a. Since on the boundary of PS, r_i is normal to the active constraints among constraints defining PS and directed toward the interior, we have

$$r_i = \sum_{a \neq a_0} \lambda_a e_a - \lambda_{a_0} \sum_{a \neq a_0} e_a,$$

for some $\lambda_a \geq 0$, such that λ_a is zero when $\pi(i, a) > 0$. Thus, a $\hat{\pi} \in KT(i)$ will satisfy

$$\hat{K}_i = \sum_{a \neq a_0} \lambda_a e_a - \lambda_{a_0} \sum_{a \neq a_0} e_a, \qquad \lambda_a \geq 0,$$

with $\lambda_a = 0$ when $\pi(i, a) > 0$. To start with, let $\lambda_{a_0} = 0$; then $\hat{K}_i = \sum_{a \neq a_0} \lambda_a e_a$. Thus for $a \neq a_0$ and $\pi(i, a) > 0$, $\hat{K}_{ia} = 0$, i.e.,

$$k(i, a_0) + \alpha \sum_j p(i, j, a_0) V_\pi(j) = k(i, a) + \alpha \sum_j p(i, j, a) V_\pi(j).$$

Also, for $a \neq a_0$ and $\pi(i,a) = 0$, $\hat{K}_{ia} = \lambda_a \geq 0$, i.e.,

$$k(i,a_0) + \alpha \sum_j p(i,j,a_0) V_\pi(j) \leq k(i,a) + \alpha \sum_j p(i,j,a) V_\pi(j).$$

Therefore,

$$\begin{aligned} V_\pi(i) &= \sum_a \pi(i,a) \left[k(i,a) + \alpha \sum_j p(i,j,a) V_\pi(j) \right] \\ &= \left[k(i,a_0) + \alpha \sum_j p(i,j,a_0) V_\pi(j) \right] \\ &= \min_a \left[k(i,a) + \alpha \sum_j p(i,j,a) V_\pi(j) \right]. \end{aligned}$$

Now suppose $\lambda_{a_0} \neq 0$. Then $\pi(i,a_0) = 0$. Note that

$$V_\pi(i) = \min_a \left[k(i,a) + \alpha \sum_j p(i,j,a) V_\pi(j) \right] \tag{5.5}$$

would also follow if we show that

$$\begin{aligned} \pi(i,a) = 0 &\Rightarrow K_{ia} \geq 0, \\ \pi(i,a) > 0 &\Rightarrow K_{ia} = 0. \end{aligned}$$

For $a \neq a_0$ and $\pi(i,a) > 0$, $\hat{K}_{ia} = -\lambda_{a_0}$, i.e., $K_{ia} = K_{ia_0} - \lambda_{a_0}$. For $a \neq a_0$, $\pi(i,a) = 0$, we have $K_{ia} = K_{ia_0} - \lambda_{a_0} + \lambda_a$. Using $\pi(i,a_0) = 0$ and (5.2), we have

$$\sum_{a \neq a_0} K_{ia} \pi(i,a) = K_{ia_0} - \lambda_{a_0} = 0,$$

implying $K_{ia_0} \geq 0$. For $a \neq a_0$, $\pi(i,a) = 0$, we then have

$$K_{ia} = K_{ia_0} - \lambda_{a_0} + \lambda_a = \lambda_a \geq 0$$

and for $a \neq a_0$ and $\pi(i,a) > 0$ we have

$$K_{ia} = K_{ia_0} - \lambda_{a_0} = 0.$$

Thus (5.5) holds. The set of equilibria is precisely $KT = \cap_{i \in S} KT(i)$, which therefore contains only optimal $\hat{\pi}$. □

The following then is a consequence of the results of section 2.

THEOREM 5.8. $(\hat{\pi}_n, V_n) \to \{(\hat{\pi}, V_\pi) : V_\pi = V_\alpha\}$ *a.s.*

5.3. Convergence of Algorithm 2. For Algorithm 2, consider (5.4) with g redefined by

$$g_{ia}(\hat{\pi}, V) = \pi(i,a) \left(V(i) - k(i,a) - \alpha \sum_j p(i,j,a) V(j) \right), \quad i \in S, \quad a \in A \backslash \{a_0\}.$$

Note that $g_{ia}(\hat{\pi}, V_\pi) = -\pi(i,a)K_{ia}$ and $\bar{P}_{\hat{\pi}}(g_i(\hat{\pi}, V_\pi)) = g_i(\hat{\pi}, V_\pi)$, so $\bar{P}_{\hat{\pi}}(\cdot)$ can be dropped in (5.4). The equilibrium points of (5.4) are now characterized by

$$\pi(i,a)\left(V(i) - k(i,a) - \alpha\sum_j p(i,j,a)V(j)\right) = 0 \ \forall i,a.$$

(Note that this does depend on π.) It is easy to see that the stable equilibria are precisely those for which $V_\pi = V_\alpha$.

THEOREM 5.9. *$(\hat{\pi}_n, V_n)$ converges to $\{(\hat{\pi}, V_\pi) : V_\pi = V_\alpha\}$ a.s.*

Proof. Setting

$$G_{ia}(\pi) = g_{ia}(\pi, V_\pi) = -\pi(i,a)K_{ia},$$

we have $G_i \cdot K_i = -\sum_a \pi(i,a)K_{ia}^2 \le 0$ with equality only at equilibria of (5.4). Thus $\sum_i V_\pi(i)$ serves as a strict Liapunov function as before, implying asymptotic stability of $\{\pi : G(\pi) = 0\}$. To claim that the convergence is in fact to the stable equilibrium a.s., argue as in [19], which is possible due to our assumptions on the external noise $\{\psi_{ij}(n)\}$. Unlike the isolated unstable equilibria of [19], we now have unstable faces, but essentially the same argument as [19] goes through. □

Remark. The condition on $\{\psi(n)\}$ is rather awkward, since it seems to imply that a "large" noise input is required in order to avoid unstable equilibria. In simulation studies, however, simple ad hoc schemes such as adding a small noise only at the boundary to push the process into the interior of $(PS)^s$ seemed to do quite well.

5.4. Convergence of Algorithm 3. Here we work with $\{\beta_n\}$ instead of $\{\pi_n\}$. Redefine g as

$$g_{ia}(\beta, V) = V(i) - k(i,a) - \alpha\sum_j p(i,j,a)V(j), \quad i \in S, \quad a \in A\backslash\{a_0\}.$$

(Note that this does not depend on π.) It is easy to see that $g_{ia}(\beta, V_{\pi_\beta}) = -K_{ia}(\pi_\beta)$. (5.4) gets replaced by

$$\dot{\beta}(i,a) = \bar{P}_{\beta_0}(g_{ia}(\beta, V_{\pi_\beta})), \quad i \in S, \quad a \in A, \tag{5.6}$$

with

$$\bar{P}_{\beta_0}(g_{ia}(\beta, V_{\pi_\beta})) = \lim_{\Delta\downarrow 0}\frac{P_{\beta_0}(\beta(i,a) + \Delta g_{ia}(\beta, V_{\pi_\beta})) - \beta(i,a)}{\Delta}$$

and P_{β_0} is as before.

LEMMA 5.10. *$\sum_{i\in S} V_{\pi_\beta}(i)$ are strict Liapunov functions for* (5.6).

Proof. By explicit differentiation, we get

$$\begin{aligned} G_{ia}(\pi_\beta) &\triangleq \dot{\pi}_\beta(i,a) \\ &= \frac{\exp(\beta(i,a))\dot{\beta}(i,a)}{\sum_{a'}\exp(\beta(i,a'))} - \frac{\exp(\beta(i,a))\sum_{a'}\dot{\beta}(i,a')\exp(\beta(i,a'))}{(\sum_{a'}\exp(\beta(i,a')))^2} \\ &= \pi_\beta(i,a)\bar{P}_{\beta_0}(g_{ia}(\beta, V_{\pi_\beta})) - \pi_\beta(i,a)\sum_{a'}\pi_\beta(i,a')\bar{P}_{\beta_0}(g_{ia'}(\beta, V_{\pi_\beta})). \end{aligned}$$

Let

$$\gamma_{ia}(\beta) = \begin{cases} 0 & \text{if} \quad \beta(i,a) = \beta_0 \quad \text{and} \quad K_{ia}(\pi_\beta) \le 0, \\ & \text{or} \quad \beta(i,a) = -\beta_0 \quad \text{and} \quad K_{ia}(\pi_\beta) \ge 0, \\ 1 & \text{otherwise.} \end{cases}$$

Then $\bar{P}_{\beta_0}(g_{ia}(\beta, V_{\pi_\beta})) = -K_{ia}(\pi_\beta)\gamma_{ia}(\beta)$ and a simple computation using (5.2) leads to

$$\sum_a K_{ia}(\pi_\beta) G_{ia}(\pi_\beta) = -\sum_a \pi_\beta(i,a) K_{ia}(\pi_\beta)^2 \gamma_{ia}(\beta) \le 0$$

with equality only at equilibrium points of (5.6). The claim follows. □

Since π_β always remains in a proper subset of $\mathcal{P}(A)^s$, the algorithm may not converge to optimal policies. Nevertheless, the following results show that we can get an ϵ-optimal solution for any $\epsilon > 0$ by choosing β_0 sufficiently large. Define the operator $T_\alpha : R^s \to R^s$ by

$$[T_\alpha(x)]_i = \min_a \left[k(i,a) + \alpha \sum_j p(i,j,a) x_j \right], \quad x = [x_1, \ldots, x_s] \in R^s, \quad 1 \le i \le s.$$

LEMMA 5.11. $\|V - V_\alpha\|_\infty \le (1-\alpha)^{-1} \|T_\alpha V - V\|_\infty$.

Proof Consider

$$\begin{aligned} \|V - V_\alpha\|_\infty &\le \|V - T_\alpha V\|_\infty + \|T_\alpha V - V_\alpha\|_\infty \\ &= \|V - T_\alpha V\|_\infty + \|T_\alpha V - T_\alpha V_\alpha\|_\infty \\ &\le \|V - T_\alpha V\|_\infty + \alpha \|V - V_\alpha\|_\infty. \end{aligned}$$

The claim follows. □

LEMMA 5.12. *For any $\epsilon > 0$ there exists sufficiently large $\bar{\beta}$ such that for all $\beta_0 \ge \bar{\beta}$ the following holds true: If β is an equilibrium point of* (5.6)*, then $V_\alpha(i) \le V_{\pi_\beta}(i) \le V_\alpha(i) + \epsilon \ \forall i \in S$*

Proof. Let $K^-_{ia}(\pi) = \max(0, -K_{ia}(\pi)), K^+_{ia}(\pi) = \max(0, K_{ia}(\pi))$. Note that

$$\|T_\alpha V_\pi - V_\pi\|_\infty = \max_{i,a} K^-_{ia}(\pi). \tag{5.7}$$

If β is an equilibrium point of (5.6), one has

$$\begin{aligned} -\beta_0 < \beta(i,a) < \beta_0 &\Rightarrow K_{ia}(\pi_\beta) = 0 \\ \beta(i,a) = -\beta_0 &\Rightarrow K_{ia}(\pi_\beta) \ge 0, \\ \beta(i,a) = \beta_0 &\Rightarrow K_{ia}(\pi_\beta) \le 0. \end{aligned}$$

Thus

$$\pi_\beta(i,a) K_{ia}(\pi\beta) = \frac{\exp(-\beta_0) K^+_{ia}(\pi\beta) - \exp(\beta_0) K^-_{ia}(\pi\beta)}{\sum_a \exp \beta(i,a)},$$

which, with (5.2), leads to

$$\sum_a K^-_{ia}(\pi_\beta) = \exp(-2\beta_0) \sum_a K^+_{ia}(\pi_\beta). \tag{5.8}$$

Let $K = \|k(\cdot,\cdot)\|_\infty$. Then $\|V_\pi\|_\infty \le K(1-\alpha)^{-1}\ \forall \pi$. Hence

$$\begin{aligned}|K_{ia}(\pi)| &\le |k(i,a)| + \alpha \sum_j p(i,j,a)|V_\pi(j)| + |V_\pi(i)| \\ &\le K + \frac{\alpha K}{1-\alpha} + \frac{K}{1-\alpha} \\ &= K\left(\frac{2+\alpha}{1-\alpha}\right).\end{aligned}$$

With (5.7), (5.8), this leads to

$$\|T_\alpha V_{\pi_\beta} - V_{\pi_\beta}\|_\infty \le (r+1)K\exp(-2\beta_0)\left(\frac{2+\alpha}{1-\alpha}\right).$$

Choose β_0 such that

$$(r+1)K\exp(-2\beta_0)\left(\frac{1+\alpha}{(1-\alpha)^2}\right) \le \epsilon.$$

The claim follows. □

THEOREM 5.13. *For any $\epsilon > 0$ there exists $\bar\beta = \bar\beta(\epsilon)$ such that for all $\beta_0 \ge \bar\beta$, (β_n, V_n) converges to the set $\{(\beta, V_{\pi_\beta}) : V_\alpha(i) \le V_{\pi_\beta}(i) \le V_\alpha(i) + \epsilon\ \forall i \in S\}$ a.s.*

6. Convergence analysis for average cost.

6.1. Preliminaries. Recall that in addition to all the assumptions underlying the preceding section, we also assume now that S is irreducible under every stationary randomized policy. As in section 4, we begin with verification of (A2), (A3). First we verify the first half of (A2) and (A3), as they are common for all the algorithms. Because the rest are algorithm specific, we deal with them individually. Our approach to verification of boundedness of h_n follows [7]. Define span seminorm $\|\cdot\|_s$ on R^s by

$$\|x\|_s = \max_i x_i - \min_i x_i.$$

Then $\|x\|_s = 0$ if and only if x is a multiple of $[1,\dots,1] \in R^s$. Rewrite (3.5) as

$$h_{n+1}(i) = h_n(i)(1-b(n)) + b(n)[k(i,\varphi_n(i)) + h_n(\xi_n(i,\varphi_n(i))) - h_n(i_0)]. \tag{6.1}$$

Let $\{h'_n\}$, $\{h''_n\}$ denote iterates of (6.1) with identical $\{(\psi_n,\varphi_n,\xi_n)\}$ but with different initial conditions.

LEMMA 6.1. $\sup_n \|h'_n - h''_n\|_s < \infty$.

Proof. Let $A(n) = \max_{m\le n}\|h'_m - h''_m\|_s$. From (6.1), one has $A(n+1) \le A(n)\ \forall n$. The claim follows. □

COROLLARY 6.2. *If the iterates of* (6.1) *remain bounded for one initial condition, they do so for all initial conditions (sample pathwise).*

Proof. From Lemma 6.1, it follows that $\|h_n\|_s$ remains bounded for all initial conditions if $\|h_n\|_\infty$ (and hence $\|h_n\|_s$) does so for one initial condition. Then it suffices to verify that one component of $\{h_n\}$ remains bounded. Letting $K_1 = \sup_n \|h_n\|_s$, $K_2 = \max_{i,a}|k(i,a)|$, we have

$$h_{n+1}(i_0) = h_n(i_0)(1-b(n)) + b(n)[k(i_0,\varphi_n(i_0)) + h_n(\xi_n(i_0,\varphi_n(i_0))) - h_n(i_0)].$$

Thus

$$|h_{n+1}(i_0)| \leq |h_n(i_0)|(1 - b(n)) + b(n)(K_1 + K_2).$$

It follows by induction that $\sup_n |h_n(i_0)| < \infty$. □

Next, consider the ODE in R^s:

$$\dot{x}(t) = F^\pi(x(t)) - x(t), \tag{6.2}$$

where $F^\pi(\cdot) = [F_1^\pi(\cdot), \ldots, F_s^\pi(\cdot)]^T : R^s \to R^s$ is defined by

$$F_i^\pi(x) = \sum_a \pi(i,a)k(i,a) + \sum_{a,j} \pi(i,a)p(i,j,a)x_j - x_{i_0}, \qquad i \in S$$

for $x = [x_1, \ldots, x_s]^T$.

LEMMA 6.3. *h_π is the unique globally asymptotically stable equilibrium point of* (6.2).

Proof. Direct verification shows that the ODE in question is an asymptotically stable linear system of the form

$$\dot{x}(t) = Ax(t) + B$$

with the desired equilibrium points as its unique globally asymptotically stable equilibrium. □

Remark. Note that asymptotic stability implies the existence of a smooth strict Liapunov function [28]. See [12] for related results.

This verifies the first half of (A3). We still need to complete our verification of (A2). For this purpose, note that h_π, being the unique solution of a well-posed linear system whose coefficients depend smoothly on π, varies continuously with π, which varies over a compact set. Thus the set Q of h_π as π varies over stationary randomized policies is compact. Take a ball $B \subset R^s$ large enough so that it contains the ϵ-neighborhood of Q (for a prescribed $\epsilon > 0$) in its interior. Let $\{h_n\}$ denote our original iterates and $\{h'_n\}$ modified iterates obtained by resetting to h_{π_n} every time it exits from B, otherwise it is identical, i.e., driven by the same random sequences.

LEMMA 6.4. *$\{h_n\}$ remains bounded a.s.*

Proof (sketch). Note that $\{h'_n\}$ is bounded a.s. by construction. Writing its iteration between two consecutive resets in the standard form

$$h'_{n+1} = h'_n + b(n)(G(h'_n) + M_n + e(n))$$

for an appropriately defined $G(\cdot)$, an "error" term $e(n)$, and a martingale difference term M_n, one verifies that $\{e(n)\}$, $\{M_n\}$ are a.s. bounded because $\{h'_n\}$ is. Thus the appropriate piecewise linear interpolation of $\{h'_n\}$ between two consecutive resets becomes a better and better approximation of trajectories of (6.2) as $n \to \infty$. Now argue as in the proof of Lemma 4.2 to conclude that with probability one, eventually the reset trajectory of $\{h'_n\}$ will not exit the ϵ-neighborhood of Q, and therefore the number of resets is a.s. finite. The claim now follows from Corollary 6.2. □

The remaining assumptions are straightforward, except for the second half of (A3), which is to be verified separately for each algorithm as before.

Next, we set up the counterpart of the results of the preceding section for the average cost problem. For this purpose, redefine $K : \mathcal{P}(\mathcal{A})^s \to R^{s \times (r+1)}$ as $K(\pi) = [[K_{ia}(\pi)\|$, where

$$K_{ia}(\pi) = k(i,a) + \sum_j p(i,j,a)h_\pi(j),$$

and $\hat{K} : (PS)^s \to R^{s\times r}$ by

$$\hat{K}_{ia}(\hat{\pi}) = K_{ia_0} \ \forall a \neq a_0, i.$$

Define vectors $K_i = [K_i.], \hat{K}_i = [\hat{K}_i.]$ correspondingly. Let G denote a vector field on $\mathcal{P}(A)^s$.

LEMMA 6.5. *If G satisfies*

$$G_i(\pi) \cdot K_i(\pi) \leq 0 \ \forall i \in S, \tag{6.3}$$

then

$$D_G \lambda_\pi \leq 0.$$

Furthermore, equality holds if and only if it does so in the former inequality for all i.

Proof. Consider $\delta_0 > 0$ small enough so that $\pi' = \pi + \delta G \in \mathcal{P}(A)^s \ \forall \delta \in (0, \delta_0)$. Let $\{X_n\}$, $\{X'_n\}$ be the chains controlled by stationary randomized policies π, π', respectively, and $\{Z'_n\}$ the control process corresponding to the latter. Then

$$\begin{aligned}
\mathrm{E}[h_\pi(X'_n)|X'_{n-1}] &= \sum_a \sum_j \pi'(X'_{n-1}, a) p(X'_{n-1}, j, a) h_\pi(j) \\
&= \delta \sum_a \sum_j G_{X'_{n-1}a}(\pi) p(X'_{n-1}, j, a) h_\pi(j) \\
&\quad + \sum_a \sum_j \pi(X'_{n-1}, a) p(X'_{n-1}, j, a) h_\pi(j) \\
&= \delta \sum_a \sum_j G_{X'_{n-1}a}(\pi) p(X'_{n-1}, j, a) h_\pi(j) \\
&\quad + \lambda_\pi + h_\pi(X'_{n-1}) - \sum_a \pi(X'_{n-1}, a) k(X'_{n-1}, a) \\
&= \delta \sum_a G_{X'_{n-1}a}(\pi) \left[k(X'_{n-1}, a) + \sum_j p(X'_{n-1}, j, a) h_\pi(j) \right] \\
&\quad + \lambda_\pi + h_\pi(X'_{n-1}) - \sum_a \pi'(X'_{n-1}, a) k(X'_{n-1}, a).
\end{aligned}$$

Therefore,

$$\begin{aligned}
&\frac{1}{N} \sum_{n=1}^{N} \mathrm{E}[(h(X'_n) - h(X'_{n-1}))|X'_{n-1}] - \lambda_\pi + \frac{1}{N} \sum_{n=1}^{N} \mathrm{E}[k(X'_{n-1}, Z'_{n-1})|X'_{n-1}] \\
&= \delta \frac{1}{N} \sum_{n=0}^{N-1} \sum_a G_{X'_{n-1}a}(\pi) \left[k(X'_{n-1}, a) + \sum_j p(X'_{n-1}, j, a) h_\pi(X'_{n-1}) \right].
\end{aligned}$$

Take expectations on both sides and let $N \to \infty$ to obtain

$$\begin{aligned}
\frac{\lambda_{\pi'} - \lambda_\pi}{\delta} &= \lim_{N\to\infty} \frac{1}{N} \sum_{n=0}^{N-1} \mathrm{E}[K_{X'_n}(\pi) \cdot G_{X'_n}(\pi)] \\
&= \mathrm{E}[K_{\bar{X}_n}(\pi) \cdot G_{\bar{X}_n}(\pi)],
\end{aligned}$$

where $\{\tilde{X}_n\}$ is the stationary process under π'. Now let $\delta \to 0$ and use continuous dependence of stationary distribution on π to obtain

$$D_G\lambda_\pi = \mathrm{E}[K_{\hat{X}_n}(\pi) \cdot G_{\hat{X}_n}(\pi)],$$

where $\{\hat{X}_n\}$ is the stationary process under π. Recall that stationary distribution under π assigns strictly positive probability to each state in S. The rest is easy. □

In terms of $\hat{K}$, $\hat{G}$, the above can be recast as follows: If $\hat{G}$ satisfies

$$\hat{G}_i(\hat{\pi}) \cdot \hat{K}_i(\hat{\pi}) \leq 0 \ \forall i \in S, \tag{6.4}$$

then

$$\nabla\lambda_\pi \cdot \hat{G}(\hat{\pi}) \leq 0$$

with equality in the latter if and only if equality holds in the former inequality for all i.

COROLLARY 6.6. *If* (6.3) (*respectively,* (6.4)) *holds with equality if and only if* $G(\pi) = 0$ (*respectively,* $\hat{G}(\hat{\pi}) = 0$), *then* λ_π *is a strict Liapunov function for the ODE*

$$\dot{\pi}(t) = G(\pi(t)),$$
$$(\textit{respectively}, \ \dot{\hat{\pi}}(t) = \hat{G}(\hat{\pi}(t))).$$

The treatment of Algorithms 4–6 closely parallels that of algorithms 1–3, so we shall only outline the essentials.

6.2. Convergence analysis of Algorithm 4. For this, define the corresponding "g" by

$$g_{ia}(\hat{\pi}, h) = k(i, a_0) - k(i, a) + \sum_j (p(i, j, a_0) - p(i, j, a))h(j)$$

for $i \in S$, $a \neq a_0$. Consider, for $\hat{\pi}_i(\cdot) \stackrel{\Delta}{=} [\hat{\pi}_{ia}(\cdot)]_{a \neq a_0}$,

$$\dot{\hat{\pi}}_i(t) = \bar{P}_{\hat{\pi}(t)}(g_i.(\hat{\pi}(t), h_{\pi(t)})), \qquad i \in S, \tag{6.5}$$

$$g_i(\hat{\pi}, h_\pi) = [g_i(\hat{\pi}, h_\pi)],$$

$$\bar{P}(g_i(\hat{\pi}, h_\pi)) = \lim_{\Delta \downarrow 0} \frac{P(\hat{\pi}(i) + \Delta g_i(\hat{\pi}, h_\pi)) - \hat{\pi}(i)}{\Delta},$$

and P is the projection into PS. Note that $g_i(\hat{\pi}, h_\pi) = -\hat{K}_i(\hat{\pi})$. Now argue as for Algorithm 1 to conclude.

THEOREM 6.7. λ_π *is a strict Liapunov function for* (6.5), *whose equilibrium points correspond to optimal* π *(i.e., such that* $\lambda_\pi = \lambda^*$*).*

COROLLARY 6.8. $(\hat{\pi}_n, h_n) \to \{(\hat{\pi}, h_\pi) : \lambda_\pi = \lambda^*\}$ *a.s.*

6.3. Convergence analysis of Algorithm 5. Change g to

$$g_{ia}(\hat{\pi}, h) = \pi(i,a)\left(h(i_0) + h(i) - k(i,a) - \sum_j p(i,j,a)h(j)\right)$$

for $i \in S$, $a \neq a_0$. The corresponding ODE is

$$\begin{aligned}\dot{\hat{\pi}}_i &= \bar{P}_{\hat{\pi}}(g_i(\hat{\pi}, h_\pi)) \\ &= g_i(\hat{\pi}(t), h_\pi(t)).\end{aligned}$$

Then it is easily checked that

$$G_i(\pi) \cdot K_i(\pi) = -\sum_a \pi(i,a)(\lambda_\pi + h_\pi(i) - K_{ia}(\pi))^2.$$

Thus we have the following analogues of the results for Algorithm 2.

THEOREM 6.9. *λ_π is a strict Liapunov function for* (6.6), *whose stable equilibrium points satisfy $\lambda_\pi = \lambda^*$.*

COROLLARY 6.10. *$(\hat{\pi}_n, h_n) \to \{(\hat{\pi}, h_\pi) : \lambda_\pi = \lambda^*\}$ a.s.*

6.4. Convergence analysis of Algorithm 6. Define π_β as before and

$$g_{ia}(\beta, h) = h(i_0) + h(i) - k(i,a) - \sum_j p(i,j,a)h(j)$$

for $i \in S$, $a \in A$. Consider the ODE

$$\dot{\beta}(i,a) = \bar{P}_{\beta_0}(g_{ia}(\beta, h_{\pi_\beta})), \tag{6.6}$$

where

$$\bar{P}_{\beta_0}(g_{ia}(\beta, h_{\pi_\beta})) = \lim_{\Delta \downarrow 0} \frac{P_{\beta_0}(\beta(i,a) + \Delta g_{ia}(\beta, h_{\pi_\beta})) - \beta(i,a)}{\Delta}$$

and P_{β_0} is as before.

LEMMA 6.11. *λ_{π_β} is a strict Liapunov function for* (6.6).

Proof Define $G(\cdot) = [[G_{ia}(\cdot)]]$ by

$$\begin{aligned}G_{ia}(\pi_\beta) &= \dot{\pi}_\beta(i,a) \\ &= \pi_\beta(i,a)\bar{P}_{\beta_0}(g_{ia}(\beta, h_{\pi_\beta})) - \pi_\beta(i,a)\sum_{a'} \pi_\beta(i,a')\bar{P}_{\beta_0}(g_{ia'}(\beta, h_{\pi_\beta})).\end{aligned}$$

Note that $\bar{P}_{\beta_0}(g_{ia}(\beta, h_{\pi_\beta})) = (\lambda_{\pi_\beta} + h_{\pi_\beta}(i) - K_{ia}(\pi_\beta))\gamma_{ia}(\beta)$, where

$$\gamma_{ia}(\beta) = \begin{cases} 0 & \text{if} \quad \beta(i,a) = \beta_0 \quad \text{and} \quad K_{ia}(\pi_\beta) \leq \lambda_{\pi_\beta} + h_{\pi_\beta}(i), \\ & \text{or} \quad \beta(i,a) = -\beta_0 \quad \text{and} \quad K_{ia}(\pi_\beta) \geq \lambda_{\pi_\beta} + h_{\pi_\beta}(i), \\ 1 & \text{otherwise.} \end{cases}$$

One verifies that

$$\begin{aligned}G_i(\pi_\beta) \cdot K_i(\pi_\beta) &= \sum_a K_{ia}(\pi_\beta)G_{ia}(\pi_\beta) \\ &= -\sum_a \pi_\beta(i,a)\gamma_{ia}(\beta)(\lambda_{\pi_\beta} + h_{\pi_\beta}(i) - K_{ia}(\pi_\beta))^2 \\ &\leq 0.\end{aligned}$$

Since $\pi_\beta(i,a) > 0$ always, equality holds for all $i \in S$ only at equilibrium points of (6.6). The rest is routine. □

The following results characterize the equilibrium points of (6.6). As before, let $e = [1, \ldots, 1]^T$ and define the operator $T : R^s \to R^s$ by

$$(Tu)(i) = \min_a \left[k(i,a) + \sum_j p(i,j,a)u(j) \right].$$

LEMMA 6.12. *If a stationary randomized policy π satisfies $\|Th_\pi - h_\pi - \lambda_\pi e\|_\infty \leq \epsilon$ for some $\epsilon > 0$, then $\lambda^* \leq \lambda_\pi \leq \lambda^* + \epsilon$.*

Proof. Let $\{X_n\}$, $\{Z_n\}$ denote, respectively, the state and control processes under some stationary randomized policy μ. Define the vector $r \in R^s$ by

$$r(i) = \lambda_\pi + h_\pi(i) - \min_a \left[k(i,a) + \sum_j p(i,j,a)h_\pi(j) \right]$$

. By hypothesis $\epsilon \geq r(i) \geq 0 \ \forall i$. Also,

$$\begin{aligned}
\mathrm{E}[h_\pi(X_{n+1})|X_n, Z_n] &= \sum_j p(X_n, j, Z_n)h_\pi(j) \\
&= \left[k(X_n, Z_n) + \sum_j p(X_n, j, Z_n)h_\pi(j) \right] - k(X_n, Z_n) \\
&\geq \min_a \left[k(X_n, a) + \sum_j p(X_n, j, a)h_\pi(j) \right] - k(X_n, Z_n) \\
&= \lambda_\pi + h_\pi(X_n) - r(X_n) - k(X_n, Z_n).
\end{aligned}$$

Hence

$$\frac{1}{N}\sum_{n=0}^{N-1} \mathrm{E}[(h_\pi(X_{n+1}) - h_\pi(X_n))|X_n, Z_n] \geq \lambda_\pi - \frac{1}{N}\sum_{n=0}^{N-1} r(X_n) - \frac{1}{N}\sum_{n=0}^{N-1} k(X_n, Z_n).$$

Take expectations on both sides and let $N \to \infty$ to obtain

$$\lambda_\pi \leq \lambda_\mu + \lim_{N\to\infty} \frac{1}{N}\sum_{n=0}^{N-1} \mathrm{E}[r(X_n)].$$

Since μ was arbitrary and $\|r\|_\infty \leq \epsilon$, the claim follows. □

LEMMA 6.13. *For any $\epsilon > 0$ there exists $\bar\beta = \bar\beta(\epsilon)$ such that for all $\beta_0 \geq \bar\beta$, every equilibrium point β of (6.6) satisfies $\lambda^* \leq \lambda_{\pi_\beta} \leq \lambda^* + \epsilon$.*

Proof. From continuous dependence of solutions of well-posed linear equations on data, we have $\pi \mapsto h_\pi$ continuous. Thus $\{h_\pi\}$ remains bounded. The rest of the proof closely mimics the proof of Lemma 5.12. □

THEOREM 6.14. *For any $\epsilon > 0$ there exists $\bar\beta > 0$ such that for $\beta_0 \geq \bar\beta$, $(\beta_n, h_n) \to \{(\beta, h_{\pi_\beta}) : \lambda^* \leq \lambda_{\pi_\beta} \leq \lambda^* + \epsilon\}$ a.s.*

7. Conclusions. The techniques used in this paper are quite general and can be used to derive the corresponding learning algorithm, given a policy iteration algorithm. In particular, learning algorithms for the stochastic shortest path (SSP) problem based on its policy iteration (see [5]) can be easily derived. The connection shown by Bertsekas [4] between SSP and average cost problem can be exploited to change our update of h_n in the algorithms for average cost problems, so that our algorithms become three time scale asynchronous algorithms, which allow general delays as in the discounted case.

There are many other interesting directions that need to be explored.

Average cost problems. In our analysis of these, we were obliged to make a part of the iterations synchronous, because Lemma 5.1 fails otherwise. Our analysis does, however, show that in the asynchronous case, one would have the desired convergence a.s. on the set of sample paths for which the iterates remain bounded. Thus the missing link is precisely a proof of a.s. boundedness for the general asynchronous case. It is our pleasure to note that while this paper was under review, this issue was settled by a novel stability test developed by Borkar and Meyn [11].

Approximation issues. Tsitsiklis and Van Roy [25] consider several approximation schemes based on compact representations in Q-learning. The idea is to directly approximate $V_n(\cdot)$ by a function belonging to a parametrized family and update the parameter in question rather than updating $V_n(\cdot)$ directly. In the actor-critic scheme, however, there is an additional iteration for $\pi_n(\cdot,\cdot)$. One can conceivably write a function approximation scheme for $\pi(\cdot,\cdot)$ using a parametrized family (such as a neural network, e.g., see [22]) and update the probabilities recursively. These possibilities need to be explored.

Feedback implementations. The algorithms we consider in this work are "off-line"; i.e., they are based on a simulation run rather than an actual system being controlled in real time. One can convert these into on-line (or feedback) adaptive control algorithms for a controlled Markov chain X_n, $n \geq 0$, by setting $Y_n = \{X_n\}$, $Z_n = \{(X_n, \varphi_n(X_n))\}$ $\forall n$ and letting $\varphi_n(X_n)$ be the actual randomized control law being implemented.

A natural question then is whether the scheme is asymptotically optimal. (For the appropriate concept of "asymptotically optimal" in the discounted framework, see [23].) Recall that the convergence of the algorithm to desired limits requires that all state-action pairs be tried sufficiently often. For states, this may happen automatically if suitable irreducibility conditions are met, even in the feedback case. But if the $\pi_n(\cdot,\cdot)$ converge rapidly (to the desired limit or otherwise) all the state-action pairs may not get updated frequently enough. A simple way out of this conflict is to modify the feedback law to a convex combination of $\varphi_n(\cdot)$ and the uniform distribution on A so as to ensure a minimum probability $\epsilon > 0$ of each $a \in A$ being picked. For $\epsilon > 0$ sufficiently small, the scheme will be nearly optimal within a prescribed tolerance. However, too small an ϵ may slow down convergence. Thus there is a trade-off involved. A potentially promising scheme is to start with a large $\epsilon \in (0, 1/r]$ (to ensure all state-action pairs being tried frequently) and then reduce it "slowly" enough to ensure optimality. (Recall the simulated annealing algorithm for global optimization.) It is, however, a nontrivial task to capture the optimal rate of decrease of ϵ in a precise manner. These issues need further study. One should also add that presence of interprocessor communication delays causes nontrivial complications in the feedback case.

Rate of convergence. We have not provided any theoretical analysis of convergence

rate. Since a stochastic approximation algorithm eventually tracks the associated ODE, the convergence of its interpolated version to a given neighborhood of the asymptotically stable limit of the ODE (assuming one exists) will closely mimic that of the ODE itself. The rate of the latter could be gauged from the Liapunov function approach. One must invert the time scaling $n \to t(n)$ to get the convergence behavior of the original algorithm.

Even this may be worthless if "eventually" is in too distant a future. There are problems, too: The ODE captures the averaging affect of the algorithm akin to the law of large numbers. But there can be fluctuations around the average behavior of the "central limit theorem" variety. For a two time scale algorithm, the time scales should be separated enough so that the slow one does not get swamped by the fluctuations of the fast one.

A related issue is the generally high variance of the stochastic approximation algorithms. An additional averaging can reduce this; see, e.g., [20]. This and other issues pertaining to improving the performance of the algorithms need careful study.

Constant ratio step sizes. We have $a(n)/b(n) \to 0$ in order to simulate the "singular ODE" effect. However Theorem 3.4, p. 516, in [15] suggests that there exists $\epsilon^* > 0$ such that for all $\epsilon \in (0, \epsilon^*) a(n) = \epsilon b(n)$ suffices for convergence of our algorithms. It would be interesting to obtain lower bounds on ϵ^* and see if there are any situations in which $\epsilon^* > 1$. The latter situation may arise when, among the limiting coupled ODEs, the rate of convergence to equilibrium for one is naturally much faster than the other without the crutches of separated time scales.

Acknowledgments. The authors are grateful to an anonymous referee for an outstanding job of refereeing, which uncovered subtle errors that had escaped our notice.

REFERENCES

[1] J. Abounadi, D. P. Bertsekas, and V. S. Borkar, *Stochastic Approximation for Non-expansive Maps: Applications to Q-Learning Algorithms*, preprint LIDS-P-2433, Lab for Info. and Decision Sciences, MIT, Cambridge, MA, 1988.

[2] J. Abounadi, D. P. Bertsekas, and V. S. Borkar, *Learning Algorithms for Markov Decision Processes with Average Cost*, preprint LIDS-P-2434, Lab for Info. and Decision Sciences, MIT, Cambridge, MA, 1988.

[3] A. Barto, R. Sutton, and C. Anderson, *Neuron-like elements that can solve difficult learning control problems*, IEEE Trans. Systems, Man and Cybernetics, 13 (1983), pp. 835–846.

[4] D. P. Bertsekas, *A new value iteration method for average cost dynamic programming problem*, SIAM J. Control Optim., 36 (1998), pp. 742–759.

[5] D. P. Bertsekas and J. N. Tsitsiklis, *An analysis of stochastic shortest path problem*, Math. Oper. Res., 16 (1991), pp. 580–595.

[6] D. P. Bertsekas and J. N. Tsitsiklis, *Neurodynamic Programming*, Athena Scientific, Belmont, MA, 1996.

[7] V. S. Borkar, *Recursive self-tuning control of finite Markov chains*, Applicationes Mathematicae, 24 (1996), pp. 169–188.

[8] V. S. Borkar, *Stochastic approximation with two time scales*, Systems Control Lett., 29 (1996), pp. 291–294.

[9] V. S. Borkar, *Asynchronous stochastic approximations*, SIAM J. Control Optim., 36 (1998), pp. 840–851.

[10] V. S. Borkar and V. R. Konda, *Actor-critic algorithm as multi-time scale stochastic approximation*, Sādhanā, 22 (1997), pp. 525–543.

[11] V. S. Borkar and S. P. Meyn, *Stability and convergence of stochastic approximation using the ODE method*, SIAM J. Control Optim., to appear.

[12] V. S. Borkar and K. Soumyanath, *A new analog parallel scheme for fixed point computation part* I: *Theory*, IEEE Trans. Circuits Systems I Fund. Theory Appl., 44 (1997), pp. 351–355.

[13] M. W. Hirsch, *Convergent activation dynamics in continuous time networks*, Neural Networks, 2 (1989), pp. 331–349.
[14] S. S. Keerthi and B. Ravindran, *A tutorial survey of reinforcement learning*, Sādhanā, 19 (1994), pp. 851–889.
[15] P. V. Kokotovic, *Applications of singular perturbation techniques to control problems*, SIAM Rev., 26 (1984), pp. 501–550.
[16] V. R. Konda, *Learning Algorithms for Markov Decision Processes*, M.S. thesis, Department of Computer Science and Automation, Indian Institute of Science, Bangalore, India, 1997.
[17] H. J. Kushner and D. S. Clark, *Stochastic Approximation for Constrained and Unconstrained Systems*, Springer-Verlag, New York, 1978.
[18] J. Neveu, *Discrete Parameter Martingales*, North–Holland, Amsterdam, 1975.
[19] R. Pemantle, *Non-convergence to unstable points in urn models and stochastic approximations*, Ann. Probab., 18 (1990), pp. 698–712.
[20] B. T. Polyak, *New method of stochastic approximation type*, Automat. Remote Control, 51 (1990), pp. 937–946.
[21] M. Puterman, *Markov Decision Processes*, John Wiley, New York, 1994.
[22] G. Santharam and P. S. Sastry, *A reinforcement learning neural network for adaptive control of Markov chains*, IEEE Trans. Systems, Man and Cybernetics, 27 (1997), pp. 588–600.
[23] M. Schäl, *Estimation and control in discounted dynamic programming*, Stochastics, 20 (1987), pp. 51–71.
[24] J. N. Tsitsiklis, *Asynchronous stochastic approximation and Q-learning*, Mach. Learning, 16 (1994), pp. 185–202.
[25] J. N. Tsitsiklis and B. Van Roy, *Feature-based methods for large scale dynamic programming*, Mach. Learning, 22 (1996), pp. 59–94.
[26] C. Watkins and P. Dayan, *Q-learning*, Mach. Learning, 8 (1992), pp. 279–292.
[27] R. Williams and L. Baird, *A mathematical analysis of actor-critic architectures for learning optimal controls through incremental dynamic programming*, in Sixth Yale Workshop on Adaptive and Learning Systems, New Haven, CT, 1990, pp. 96–101.
[28] F. W. Wilson, *Smoothing derivatives of functions and applications*, Trans. Amer. Math. Soc., 139 (1969), pp. 413–428.

SIAM J. CONTROL OPTIM.
Vol. 38, No. 1, pp. 124–137

LIPSCHITZ-TYPE STABILITY IN NONSMOOTH CONVEX PROGRAMS*

ROBERT JANIN† AND JACQUES GAUVIN‡

Abstract. This paper deals with upper Lipschitzian continuity of the optimal solution to parametrized convex programs with linear equality and inequality constraints and with a convex nondifferentiable objective function. Under quadratic growth conditions for the objective function, some accurate bound for the rate of the upper Lipschitzian continuity is provided.

Key words. convex minimization, stability, sensitivity analysis

AMS subject classifications. 90C31, 90C34

PII. S0363012996298990

1. Introduction.

Consider some elementary convex mathematical program of the following form:

$$\min f(x), x \in \mathbb{R}^n,$$
$$\text{subject to (s.t.)} Ax \in y + K,$$

where the function f is convex and continuous but possibly nondifferentiable, the operator A is linear from $\mathbb{R}^n$ into $\mathbb{R}^m$, and the set K is some polyhedral convex cone of $\mathbb{R}^m$. The parameter is denoted by y (as in [14]) and will also be called perturbation. Does the strong convexity of the objective function f imply the Lipschitz-type stability of the optimal solution?

The local Lipschitz dependence, with respect to parameters, of the optimal solution to minimization problems has been studied much. When the function f is twice differentiable, the answer for the question is positive; when the function f is nondifferentiable, the answer for the question may be negative. It was shown in [3] that a necessary condition for upper Lipschitzian continuity of the optimal solution is that the linearized problem has some optimal solution. By upper Lipschitzian continuity we mean there is some constant k such that $\| x(y) - x(y_o) \| \leq k \| y - y_o \|$ for any perturbation y. Here we denote by $\| x \|$ the euclidean norm of the vector x, the norm for the linear operator A is defined by $\| A \| = \sup_{\|x\|=1} \| Ax \|$. If in addition to the above condition some condition denoted by (A) in [18] is assumed, then the upper Lipschitzian stability of the optimal solution holds.

The purpose of the present paper is to show how the nonsmoothness of the objective function steps into the loss of Lipschitzian stability by means of subgradients and Lagrange–Kuhn and Tucker multipliers and to give some accurate bound for the rate of upper Lipschitzian continuity. For the sake of a clear result, the minimization program has been taken as simple as possible.

In the present paper the elements of $\mathbb{R}^p$ are represented by column vectors, the linear mappings are represented by their associate matrix in the canonical basis, the

*Received by the editors February 16, 1996; accepted for publication (in revised form) July 16, 1998; published electronically December 7, 1999.
http://www.siam.org/journals/sicon/38-1/29899.html
†Département de Mathematiques et Informatiques, U.F.R. Sciences Exactes et Naturelles, Université des Antilles et de la Guyane, BP592, 97167 Pointe a Pitre Cedex, French West Indies (janin@univ-ag.fr).
‡Ecóle Polytechnique, Montreal, Quebec H3C 3A7, Canada (Jacques.Gauvin@mail.polymte.ca).

superscript T means transposition, and the subspace $E^{\perp}$ is the orthogonal complement of the subspace E. The polyhedral convex cone K we consider here has the form

$$K = \{y \in \mathbb{R}^m : y_i = 0 (1 \leq i \leq r), y_i \geq 0 (r+1 \leq i \leq m)\},$$

and we assume that the so-called Mangasarian–Fromovitz regularity condition (MFC) holds at the optimal solution x_o for the parameter $y_o = 0$:

$$\text{(MFC)} \quad \begin{cases} \text{(i) the rows } A^i \ (1 \leq i \leq r) \text{ of the matrix } A \text{ are linearly independent;} \\ \text{(ii) there exists some vector } z \in \mathbb{R}^n \text{ such that} \\ A^i z = 0 \ (1 \leq i \leq r),\ A^i z > 0 \ (i \in I(x_o)), \end{cases}$$

where $I(x) = \{i : r+1 \leq i \leq m \ , \ A^i x = 0\}$. The regularity condition MFC is a particular case of Robinson's condition [15]. On the other hand, assume that the independence of the family $A^i (i = 1, 2, \ldots, m,\ A^i x_o = 0)$ is a particular case of the MFC. Assuming that x_o is some optimal solution and the MFC holds at the point x_o, then there exists some multiplier vector $\lambda \in K^+ = \{\lambda \in \mathbb{R}^m : \lambda^T u \geq 0,\ u \in K\}$ such that

$$0 \in \partial f(x_o) \ - A^T \lambda \text{ and } \lambda^T A x_o = 0; \tag{1}$$

that is,

$$\lambda_i \geq 0, i \in I(x_o),\ \sum_{i=1}^{m} \lambda_i A^{iT} \in \partial f(x_o), \text{ and } \lambda_i A^i x_o = 0 \ (1 \leq i \leq m). \tag{2}$$

Here we denote by $\partial f(x_o)$ the subdifferential of the convex function f at the point x_o, the elements of the subdifferential of the function f at the point x_o are called subgradients of f. The directional derivative and the subdifferential of the function f at the point x_o are related by the relationship

$$Df(x; y) = \lim_{t \searrow 0} \frac{1}{t}[f(x + ty) - f(x)] \geq \xi^T . y \ \forall \xi \in \partial f(x),$$

and if we assume the function f is continuous at the point x then we have

$$Df(x; y) = \max_{\xi \in \partial f(x)} \xi^T y. \tag{3}$$

In the study of asymptotic properties of the optimal solution, the first approach, to our knowledge, has been based upon implicit function theorems applied to the system of $m + n$ equations with $m + n$ unknowns

$$\begin{gathered} \sum_{i=1}^{m} \lambda_i A^i - \nabla f(x_o) \ = 0, \\ A^i x = y^i \qquad (1 \leq i \leq r), \\ \lambda_i (A^i x - y^i) = 0 \qquad (r+1 \leq i \leq m). \end{gathered}$$

For smooth functions, assuming the independence of the gradients of binding constraints and nonzero Lagrange–Kuhn and Tucker multipliers corresponding to binding inequality constraints ($\lambda_i > 0$ for $i \in I(x_o)$, strict complementarity condition) Pallu de la Barrière [13, Thm. 6, p. 296] and Fiacco and McCormick [6] presented the

differentiability of the optimal solution with respect to the parameter y; Jittorntrum [12] obtained directional derivatives without strict complementarity condition.

For a more general constraint set, Robinson [16] presented some generalized implicit function theorem and got the Lipschitzian stability of the solution of the generalized equation:

$$0 \in g(\xi, y) + \partial\Omega_F(\xi), \tag{4}$$

where $\Omega_F(.)$ is the indicator function of the closed constraints set F and g is differentiable with respect to the variable ξ and Lipschitz continuous with respect to the variable y; similar results are presented in [1], [17], [5], [11]. The problem we consider here may be reformulated in the framework of generalized equations; define, for $\xi = (x, \lambda)$, the Lagrangian function of the program:

$$L(\xi, y) = L(x, \lambda, y) = f(x) - \sum_{i=1}^{m} \lambda_i (A^i x - y^i),\ \lambda_i \geq 0\ (r+1 \leq i \leq m). \tag{5}$$

Then, assuming the MFC is satisfied, the above conditions, necessary for optimality, are formulated by the following generalized equations:

$$\begin{cases} 0 \in \partial_x L(\xi, y) = \partial f(x) - \sum \lambda_i A^i, \\ 0 \in \partial_\lambda L(\xi, y) = -Ax + y \ + \ K_\lambda, \end{cases} \tag{6}$$

where $K_\lambda = \{z \in K^+ : \lambda^T.z = 0\}$. Since the program is convex (6) are also sufficient for the point x to be optimal; the relationship with (4) is clear with

$$g(\xi, y) = g(x, \lambda, y) = \begin{pmatrix} \partial f(x) - \sum \lambda_i A^i \\ -Ax + y \end{pmatrix}$$

and with the feasible set

$$F = \{\xi = (x, \lambda) \in \mathbb{R}^n \times R^m :\ \lambda_i \geq 0\ i = r+1, r+2, \ldots, m\}.$$

Because in the present paper the function f is not smooth, then the function g is possibly multivalued and cannot be differentiable with respect to the variable $\xi = (x, \lambda)$. Therefore, the above method does not apply.

In recent years a different method has given important improvements in sensitivity analysis and in the study of the asymptotic expansion of the optimal solution. The method is very common in the study of partial differential equations to obtain a priori bounds for the solution u of the variational equation

$$a(u, v) = \langle h, v \rangle,\ v \in V,$$

where V is some Hilbert space, a is some real functional on $V \times V$, and h is some element of the dual V'; for $v \in V$, the number $h(v)$ is denoted by $\langle h, v \rangle$ and satisfies the inequality $| \langle h, v \rangle | \leq \| h \|_{V'} . \| v \|_V\ (\forall v \in V, \forall h \in V')$.

Assume that the functional a is coercive, i.e., there is some positive number c such that

$$a(v, v) \geq c \| v \|_V^2\ (\forall v \in V),$$

and assume $u \in V$ is some solution of the problem. Then

$$c \parallel u \parallel_V^2 \leq \parallel h \parallel_{V'} \parallel u \parallel_V \quad \text{and} \quad \parallel u \parallel_V \leq \frac{1}{c} \parallel h \parallel_{V'} . \tag{7}$$

The inequality (7) then gives some upper bound for the rate of Lipschitzian continuity for the solution u with respect to the parameter h.

The above method has been introduced into sensitivity analysis independently by Shapiro [19] and by Gauvin and Janin [7], [8]: the coercivity condition is related to second-order sufficient optimality condition, and the inequality of pairing is related to upper estimates on the optimal value function. This idea is then developed particularly in [2], [20], [18], [3], and for nonisolated minima in [4]. We follow this method in our paper.

Authors [19], [1], [7], [2], [20], [3], [4] have generally related the Lipschitzian continuity of the optimal solution to the only second-order derivative (in classical or generalized sense) of the Lagrangian function (5) of the program . In reference [10], we have considered the most simple case where constraints are equalities $Ax = y$; we have shown that, for nonsmooth function f, the rate of upper Lipschitzian continuity, when it holds, is also related to some part of the subdifferential of the objective function at the optimal point, namely the set

$$\sum = \{\sigma \in \partial f(x_o) : \ \sigma^T A^{\#} y > \lambda^T y \text{ for any } \lambda \text{ satisfying (2)}\}.$$

For any full row ranked matrix A we denote by $A^{\#} = A^T(AA^T)^{-1}$ the least square sense inverse of A. If the MFC is satisfied, then the matrix A is not full row ranked but convenient full row ranked submatrices will be characterized.

2. Illustrative example. For $\varepsilon \geq 0$ consider the program

$$\begin{gathered} \min f(x_1, x_2) = \min \mid x_2 - \varepsilon x_1 \mid + \tfrac{1}{2}(x_1^2 + x_2^2) \\ \text{s.t. } x_2 = y. \end{gathered}$$

Here $n = 2, m = 1$, $K = \{0\}$ with $K^+ = \mathbb{R}$, $A = [0\ 1]$ with $A^{\#} = \begin{bmatrix} 0 \\ 1 \end{bmatrix}$.

For any ε and for $y = 0$, the optimal solution is

$$x_o = \begin{bmatrix} 0 \\ 0 \end{bmatrix}.$$

For $\varepsilon > 0$, the first-order optimality condition (2) is satisfied with the unique multiplier $\lambda = 0$:

$$\begin{bmatrix} 0 \\ 1 \end{bmatrix} [0] = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \in \partial f(0,0) = \left\{ (1-2t) \begin{bmatrix} \varepsilon \\ -1 \end{bmatrix} \ : 0 \leq t \leq 1 \right\}.$$

Then, for $y > 0$, we have

$$\sum = \{\sigma \in \partial f(0,0) : \sigma^T A^{\#} y > 0\} = \left\{ (1-2t) \begin{bmatrix} \varepsilon \\ -1 \end{bmatrix} : t \in \left]\frac{1}{2}, 1\right[\right\}.$$

The result obtained in [10] is the following:

$$\parallel x(y) - x_o \parallel \leq \sqrt{\parallel A^{\#} \parallel^2 + \sup_{\sigma \in \Sigma} \frac{\parallel (\sigma - \sigma_o)^T A^{\#} \parallel^2}{\parallel (\mathbf{1} - A^{\#}A)(\sigma - \sigma_o) \parallel^2}}\, y = \sqrt{1 + \frac{1}{\varepsilon^2}}\, y.$$

For $-\varepsilon^2 \leq y \leq \varepsilon^2$, we have the optimal solution: $x(y) = \begin{bmatrix} y/\varepsilon \\ y \end{bmatrix}$, for which the rate of Lipschitz continuity is exactly the number given by the above formula.

On the other hand, for $\varepsilon = 0$, the first-order optimality condition is satisfied with any $\lambda \in [-1, +1]$:

$$\begin{bmatrix} 0 \\ 1 \end{bmatrix} [\lambda] = \begin{bmatrix} 0 \\ \lambda \end{bmatrix} \in \partial f(0,0) = \{0\} \times [-1, +1] \ .$$

Here we have $\Sigma = \emptyset$; the result, given in [7], becomes

$$\| \, x(y) - x_o \, \| \leq \| \, A^{\#} \, \| \quad | \, y \, | = | \, y \, | \ .$$

For this case, the optimal solution is $\begin{bmatrix} 0 \\ y \end{bmatrix}$, and the rate of Lipschitz continuity is again the number given in the above formula.

Our aim is now to extend the results of [10] for the more general case of equality and inequality constraints.

3. Assumptions. The objective function $f : \mathbb{R}^n \to \mathbb{R}$ is assumed to be convex and finite, and we denote by x_o the optimal solution corresponding to the parameter $y = 0$ and we assume the MFC holds at the point x_o. We assume moreover that the following local conditions are satisfied.

1. *Superquadratic growth condition*: there is $\rho > 0$ such that
   $f(x) \geq f(x_o) + Df(x_o; x - x_o) + \frac{\rho}{2} \, \| \, x - x_o \, \|^2$, $x \in x_o + B$.
2. *Subquadratic growth condition*: there is $\Gamma \geq \rho$ such that
   $f(x) \leq f(x_o) + Df(x_o; x - x_o) + \frac{\Gamma}{2} \, \| \, x - x_o \, \|^2$, $x \in x_o + B$,
   where the set B is some compact convex neighborhood of 0.

REMARK 1. *Let γ be some univariate convex increasing function and consider the program*

$$\begin{array}{c} \min \gamma[f(x)], x \in \mathbb{R}^n, \\ \text{s.t. } Ax \in y + K. \end{array}$$

The value function $v(y) = \inf_{Ax \in y+K} f(x)$ becomes $\gamma[v(y)]$ but the optimal solution remains the same. Therefore the above conditions 1 *and* 2 *may not be satisfied for a given f but satisfied by $\gamma[f]$ for a convenient function γ. By example the function $f(x_1, x_2) = (x_1^2 + x_2^2)^{\frac{3}{4}}$ does not satisfy condition* 2*, but for $\gamma(t) = t^{\frac{4}{3}}$ it is clear that both conditions hold for the function*

$$\gamma[f](x_1, x_2) = f(x_1, x_2)^{\frac{4}{3}} = x_1^2 + x_2^2.$$

REMARK 2. *As remarked by one of the referees of this article, the subquadratic growth condition does correspond to the upper estimate of the value function and it has to be compared with conditions which deal with upper estimates of the value function. The parametric optimization problem considered here can be formulated in the following form:*

$$\min_{(x,\alpha) \in \mathbb{R}^n \times \mathbb{R}} \alpha \ \text{ s.t. } G((x, \alpha), y) \in C, \tag{8}$$

where $G((x, \alpha), y) = ((x, \alpha), Ax - y)$, $C = (epi \, f) \times K$, and epi $f = \{(x, \alpha) \in \mathbb{R}^n \times R : \alpha \geq f(x)\}$. In such a formulation the nonsmoothness of the objective function has

been removed and it is hidden now in the convex set C. Consider the "linearized" problem, here the problem

$$\min_h Df(x_o;h) \text{ s.t. } Ah - \delta y \in T_K(Ax_o)$$

and in the other formulation

$$\min_{(h,\delta\alpha)\in\mathbb{R}^n\times\mathbb{R}} \delta\alpha$$

$$\text{s.t. } G((h,\delta\alpha),\delta y) \in T_C((x_o,f(x_o)),0) = T_{epif}(x_o,f(x_o)) \times T_K(Ax_o),$$

where $T_C(z)$ is the contingent cone to the set C at the point z_o. The assumption (A) in [18] *which gives the upper estimate consists mainly in the following: there exists some optimal solution (ξ,β) of the linearized problem such that*

$$dist\{G((x_o+t\xi, f(x_o)+t\beta),\ t\delta y), C\} = O(t^2).$$

It is clear that $\beta = Df(x_o,\xi)$ and that the assumption (A) of [18] *becomes*

$$f(x_o+t\xi) = f(x_o) + tDf(x_o,\xi) + O(t^2).$$

That is precisely, for only one direction, our subquadratic growth condition. Some subquadratic-growth-type condition can be also identified in [3] *each time the condition $T_C^2(\xi,\beta) \neq \emptyset$ holds, as in Proposition* 2.2 *of* [3]. *Here (ξ,β) are as above, $\beta = Df(x_o;\xi)$, and*

$$T_C^2(z,d) = \left\{k :\ dist\left\{z+td+\frac{t^2}{2}k, C\right\} = o(t^2)\right\}.$$

The condition $T_C^2(\xi,\beta)\neq\emptyset$ is slightly stronger than the assumption (A) in [18] *because of $o(t^2)$ instead of $O(t^2)$.*

REMARK 3. *On the other hand the superquadratic growth condition corresponds to the lower estimate obtained by means of second-order sufficient optimality conditions (assumption (B) of* [18]*). The formulation* (8) *is not convenient for assumption (B) in* [18] *since the second-order derivatives of the objective function cancel.*

4. Second-order marginal analysis result. The optimal value function $v(.)$ is convex and the subgradients at the point zero are optimal solutions of the dual program; thus the subgradients of the value function v are precisely the Lagrange–Kuhn and Tucker multipliers of the program and the following first-order marginal result holds under the MFC [14, Thm. 29.1, p. 298]:

$$\partial v(0) = \left\{\lambda\in\mathbb{R}^m : \lambda_i \geq 0\ (i\in I(x_o)), \sum_{i=1}^m \lambda_i A^{iT} \in \partial f(x_o), \lambda^T Ax_o = 0\right\}. \tag{9}$$

We are interested here with the second-order term of the power series of the value function $v(.)$ near the point $y=0$. Denote by $X(y)$ the set of optimal solutions for the value y of the perturbation; the following lemma states precisely the continuity of $X(.)$ at $y=0$.

LEMMA 4.1. *Assume that the superquadratic growth condition and the MFC are satisfied. Then the multivalued mapping*

$$y \longrightarrow X(y)$$

is continuous at the point $y = 0$.

Proof. If the MFC is satisfied, then there exists some vector $\xi \in \mathbb{R}^n$ such that

$$\begin{aligned} &A^{iT}\xi = 0 \quad (1 \le i \le r), \\ &A^{iT}\xi > 0 \quad (r+1 \le i \le m, A^{iT}x_o = 0). \end{aligned}$$

The family of linear forms $A^i (1 \le i \le r)$ is linearly independent; there is then some linear mapping $A^\natural$ from $\mathbb{R}^r$ into $\mathbb{R}^n$ such that $A^{iT}A^\natural z = z^i (1 \le i \le r)$.

Denote by y_r the projection of the vector $y \in \mathbb{R}^m$ onto the subspace $\mathbb{R}^r \times \{0\}_{m-r}$, and define the continuous mapping $\xi(y) = x_o + A^\natural y_r + k \|y\| \xi$, where k is some constant to be determined in order for the point $\xi(y)$ to be a feasible solution for any perturbation y. We prove now that such a number k exists.

Since $A^{iT}A^\natural y_r = y^i$ and $A^{iT}\xi = 0$ $(1 \le i \le r)$,

$$A^{iT}\xi(y) = y^i (1 \le i \le r).$$

On the other hand, for $r+1 \le i \le m$ such that $A^{iT}x_o = 0$, we have

$$A^{iT}\xi(y) - y^i = A^{iT}A^\natural y_r - y^i + k \|y\| A^{iT}\xi;$$

since there exists some positive number l_i such that $A^{iT}A^\natural y_r - y^i \ge -l_i \|y\|$ and $A^{iT}\xi > 0$, there exists a number k_i such that

$$A^{iT}\xi(y) - y^i \ge -l_i \|y\| + k_i \|y\| A^{iT}\xi \ge 0 \text{ for any perturbation } y,$$

taking k as the maximum of the numbers $k_i (\, r+1 \le i \le m \,, A^{iT}x_o = 0)$, the point $\xi(y)$ is some feasible solution for any perturbation y close to zero. Since $\lim_{y \to O} \xi(y) = x_o$, there exists then some neighborhood V of the null perturbation such that

$$\forall y \in V \Longrightarrow \ \xi(y) \in B,$$

which shows that for any $y \in V$ the program has feasible solutions in the set B.

Consider the extended value function

$$\begin{aligned} &f_B(x) = f(x) \text{ for } x \in B, \\ &f_B(x) = +\infty \text{ for } x \notin B. \end{aligned}$$

Denote by $X_B(y)$ the set of optimal solutions of the program where f has been replaced by f_B; for any $y \in V$, and for any $x \in X_B(y)$, we have

$$f(x) = f_B(x) \le f_B(\xi(y)) = f(\xi(y)) < +\infty.$$

Denote by $x(y)$ some selection of $X_B(y)$. Consider now some cluster point $\widetilde{x}$ of $x(.)$ as the perturbation y tends to zero, taking some subsequence $x_i = x(y_i)$ tending to $\widetilde{x}$ we have the following:

$$f(\widetilde{x}) = \lim_i f(x(y_i)) \ \le \lim_{y_i \to 0} f(\xi(y_i)) = .f(x_o).$$

Since $\widetilde{x} \in B$ and is some feasible solution for the null perturbation, $f(\,\widetilde{x}) = f(x_o)$, and because of superquadratic growth condition, the point x_o is the only optimal solution for the null perturbation. Then we have $\widetilde{x} = x_o$.

Since x_o is the only cluster point of $x(y)$ as y tends to zero, we have

$$\lim_{y\to 0} X_B(y) = x_o.$$

That is to say, for any open subset Ω containing x_o there exists some neighborhood W of the perturbation 0 such that

$$y \in W \Longrightarrow X_B(y) \subset \Omega;$$

taking for Ω the interior of B denoted by int(B), there exists some neighborhood W of zero such that

$$X_B(y) \subset \text{int}(B)\ (\forall y \in W).$$

We show now that, for any $y \in W$, the convex compact set $X_B(y)$ is the set $X(y)$ of optimal solutions for the original program with the objective function f. Consider some feasible solution x such that $x \notin X_B(y)$ and

$$f(x) \leq f(x') \text{ for any } x' \in X_B(y).$$

Consider the point $x_t = (1-t)x' + tx$ for $0 \leq t \leq 1$; there exist two positive numbers t_1 and t_2 such that $x_t \in B$ and $x_t \notin X_B(y)$ for $t_1 \leq t \leq t_2$, but by the inequality of convexity $f_B(x_t) \leq f_B(x')$ which cannot hold because $x' \in X_B(y)$ and $x_t \notin X_B(y)$. Such a feasible solution x cannot exist and $X(y) = X_B(y)$; then $\lim_{y\to 0} X(y) = x_o$.

LEMMA 4.2. *Suppose the assumption of subquadratic growth condition holds and the MFC is satisfied; then there is some $\varepsilon > 0$ such that, for $\| y \| \leq \varepsilon$, the optimal value $v(y)$ satisfies*

$$v(y) \leq v(0) + \min_{A\xi \in y + K_o} Df(x_o; \xi) + \frac{\Gamma}{2} \| \xi \|^2,$$

where $K_o = \{y \in \mathbb{R}^m : y_i = 0 (1 \leq i \leq r), y_i \geq 0\ (i \in I(x_o))\}$.

Proof. Let $\delta > 0$ be some positive number to be chosen later but such that the ball B_δ with radius δ is contained in the neighborhood where subquadratic growth condition holds; because of the continuous dependence of the optimal solutions at the point 0, there exists $\varepsilon_1 > 0$ such that for $\|y\| < \varepsilon_1$, any optimal solution $x(y)$ is contained in $x_o + B_\delta$ and $v(y)$ is attained by some point in $x_o + B_\delta$.

Because of the subquadratic growth condition, for any $\xi \in B_\delta$ such that $A(x_o + \xi) \in y + K$, we have the following:

$$v(y) \leq f(x_o + \xi) \leq f(x_o) + Df(x_o; \xi) + \frac{\Gamma}{2}\|\xi\|^2.$$

Minimizing now with respect to ξ we get

$$v(y) - v(0) \leq \min_{\xi \in B_\delta\ ,\ A(x_o+\xi)\in y+K} Df(x_o; \xi) + \frac{\Gamma}{2}\|\xi\|^2.$$

Now we choose the radius δ such that

$$\{x : Ax \in y + K_o, x \in x_o + B_\delta\} = \{x : Ax \in y + K, x \in x_o + B_\delta\}.$$

The second set is a priori contained in the first set; we show now the converse. The number $\alpha = \min_{i\notin I(x_o)} A^{iT} x_o$ is positive. Take $\varepsilon_2 = \frac{\alpha}{3}$ and r such that $| A^{iT}(x - x_o) | < \frac{\alpha}{3}$ for any $i \notin I(x_o)$ and for any x such that $\| x - x_o \| < \delta$.

Consider some perturbation y with $\| y \| < \varepsilon_2$ and some x with $\| x - x_o \| < \delta$ and $Ax \in y + K_o$; then for $i \notin I(x_o)$

$$A^{iT}x - y^i = A^{iT}x_o - y^i + A^{iT}(x - x_o) \geq \alpha - \frac{\alpha}{3} - \frac{\alpha}{3} > 0.$$

We then have

$$Ax \in y + K.$$

Take $\varepsilon_3 = \min\{\varepsilon_1, \varepsilon_2\}$; then, for $\|y\| \leq \varepsilon_3$, we have

$$v(y) - v(0) \leq \min_{\xi \in B_\delta, A(x_o+\xi) \in y+K_o} Df(x_o; \xi) + \frac{\Gamma}{2}\|\xi\|^2.$$

Since $-Ax_o \in K_o$, $-Ax_o + K_o = K_o$, and the following holds:

$$v(y) - v(0) \leq \min_{\xi \in B_\delta, A\xi \in y+K_o} Df(x_o; \xi) + \frac{\Gamma}{2}\|\xi\|^2.$$

Since the optimal $\xi(y)$ in the minimization of the right-hand side is continuous with respect to y at the point $y = 0$, and since $\lim_{y\to 0} \xi(y) = 0$, there exists $\varepsilon_4 > 0$ such that $\|\xi(y)\| < \delta$ for $\|y\| < \varepsilon_4$. The result follows.

The next lemma refines the result of the above lemma.

LEMMA 4.3. *Suppose the assumptions of superquadratic and subquadratic growth conditions hold and the MFC is satisfied; then there exists $\varepsilon > 0$ such that for all perturbation $y \in \mathbb{R}^m$ with $\| y \| \leq \varepsilon$, the optimal value $v(y)$ satisfies the following:*

$$v(y) \leq v(0) + \max_{\sigma \in \partial f(x_o)} \min_{A\xi \in y+K_o} \frac{\Gamma}{2} \left(\left\| \xi + \frac{\sigma}{\Gamma} \right\|^2 - \left\| \frac{\sigma}{\Gamma} \right\|^2 \right). \tag{10}$$

Proof. The function f is continuous at the point x_o; then by (3)

$$v(y) \leq v(0) + \min_{A\xi \in y+K_o} \max_{\sigma \in \partial f(x_o)} \left(\sigma^T \xi + \frac{\Gamma}{2} \| \xi \|^2 \right).$$

The function $\xi \to \sigma^T \xi + \frac{\Gamma}{2} \| \xi \|^2$ is convex and tends to $+\infty$ as $\| \xi \|$ tends to $+\infty$; on the other hand the function $\sigma \to \sigma^T \xi + \frac{\Gamma}{2} \| \xi \|^2$ is concave on the compact set $\partial f(x_o)$. By the saddle point theorem, Theorem VII 4.3.1, in [9] we have

$$v(y) \leq v(0) + \max_{\sigma \in \partial f(x_o)} \min_{A\xi \in y+K_o} \left(\sigma^T \xi + \frac{\Gamma}{2} \| \xi \|^2 \right);$$

since we have

$$\sigma^T \xi + \frac{\Gamma}{2} \| \xi \|^2 = \frac{\Gamma}{2} \left(\left\| \xi + \frac{\sigma}{\Gamma} \right\|^2 - \left\| \frac{\sigma}{\Gamma} \right\|^2 \right),$$

the assertion holds.

The following result gives some precision about the second-order term of the power series of the marginal function $v(y)$.

THEOREM 4.4. *Suppose the assumptions of superquadratic and subquadratic growth conditions hold and the MFC is satisfied, then there exists $\varepsilon > 0$ such that for all perturbation $y \in \mathbb{R}^m$ with $\| y \| \leq \varepsilon$, there exist $\sigma_o \in \partial f(x_o)$ and $\lambda \in \mathbb{R}^s$, $\lambda_i \geq 0$ $(i \in I(x_o))$, and some full row ranked submatrix A_L such that*

1. $\operatorname{rank}(A_L) = \operatorname{rank}(A)$, $A_L^T\lambda = \sigma_o$, $\lambda^T A_L x_o = 0$, *and* $Dv(0;y) = \lambda^T y = \sigma_o^T A_L^{\#} y$,
2. *if* $\Sigma = \{\sigma \in \partial f(x_o) : \sigma T A_L^{\#} y > Dv(0;y)\}$ *then the following alternative holds: either* $\Sigma = \emptyset$, *in which case*

$$v(y) - v(0) \le Dv(0;y) + \frac{\Gamma}{2} \parallel A_L^{\#} y \parallel^2,$$

or $\Sigma \neq \emptyset$, *in which case*

$$v(y) - v(0) \le Dv(0;y) + \frac{\Gamma}{2}\left(\parallel A_L^{\#} y \parallel^2 + \sup_{\sigma\in\Sigma} \frac{\mid (\sigma - \sigma_o)^T A_L^{\#} y \mid^2}{\parallel (\mathbf{1} - A_L^{\#} A_L)(\sigma - \sigma_o) \parallel^2} \right).$$

Proof. Following [9] we have

$$\begin{aligned} Dv(0;y) &= \max\{\lambda^T y : \lambda \text{ satisfies (2)}\} \\ &= \max\{\lambda^T y : A^T\lambda = \sigma, \sigma \in \partial f(x_o), \lambda^i \ge 0 \ (i \in I(x_o))\}. \end{aligned}$$

There exists $\sigma_o \in \partial f(x_o)$ such that

$$Dv(0;y) = \max\{\lambda^T y : A^T\lambda = \sigma_o, \lambda^i \ge 0 \ (i \in I(x_o))\}.$$

The right-hand-side maximization is a linear program which is feasible and bounded above. The dual program

$$\begin{aligned} &\text{Min } \sigma_o^T \xi \\ &\text{s.t.} \begin{cases} A^i\xi = y^i & (1 \le i \le r), \\ A^i\xi \ge y^i & (i \in I(x_o)) \end{cases} \end{aligned}$$

also has optimal solutions. The set $\mathcal{S}$ of feasible solutions is polyhedral convex and closed. It may be written

$$\mathcal{S} = \mathcal{S} \cap F + \ker A,$$

where F is some complement of $\ker A$; take, for example, $F = \operatorname{range}(A^T) = \ker(A)^{\perp}$. The function to be minimized is linear, then there is at least one minimizer, denoted by ξ_o, in $\mathcal{S}\cap F$, which is an extremal point of $\mathcal{S}\cap F$. The point ξ_o satisfies $s = \operatorname{rank}(A^T) = \operatorname{rank}(A)$ independent constraints. Denote by A_L the corresponding submatrix of the matrix A; then

$$\xi_o = A_L^{\#} y \text{ and } A_L^{\#} y + \ker A \subset \{\xi : A\xi \in y + K_o\}.$$

Following (10) we have

$$v(y) \le v(0) + \max_{\sigma\in\partial f(x_o)} \min_{\xi \in A_L^{\#} y + \ker A} \frac{\Gamma}{2}\left(\left\| \xi + \frac{\sigma}{\Gamma} \right\|^2 - \left\| \frac{\sigma}{\Gamma} \right\|^2 \right).$$

The optimal ξ in the right-hand side is the projection of $-\frac{\sigma}{\Gamma}$ on $A_L^{\#} y + \ker A$,

$$\xi = -\frac{\sigma}{\Gamma} + A_L^{\#} A_L \frac{\sigma}{\Gamma} + A_L^{\#} y,$$

and by Pythagoras's theorem

$$v(y) - v(0) \leq \max_{\sigma \in \partial f(x_o)} \frac{\Gamma}{2} \parallel A_L^{\#} y \parallel^2 + \sigma^T A_L^{\#} y - \frac{1}{2\Gamma} \parallel (\mathbf{1} - A_L^{\#} A_L)\sigma \parallel^2 . \tag{11}$$

The subgradient σ_o which has been characterized above is such that

$$Dv(0; y) = \sigma_o^T A_L^{\#} y,$$

and since $\sigma_o \in \text{range}(A_L^T) = \ker(A_L)^{\perp}$ we have $(\mathbf{1} - A_L^{\#} A_L)\sigma_o = 0$

$$A_L^{\#} A_L \sigma_o = \sigma_o.$$

The inequality (11) then gives

$$v(y) - v(0) \leq \sigma_o^T A_L^{\#} y \; + \frac{\Gamma}{2} \parallel A_L^{\#} y \parallel^2$$

$$+ \max_{\sigma \in \partial f(x_o)} (\sigma^T - \sigma_o^T) A_L^{\#} y - \frac{1}{2\Gamma} \parallel (\mathbf{1} - A_L^{\#} A_L)(\sigma - \sigma_o) \parallel^2 . \tag{12}$$

Consider the expression

$$E = \max_{\sigma \in \partial f(x_o)} (\sigma^T - \sigma_o^T) A_L^{\#} y - \frac{1}{2\Gamma} \parallel (\mathbf{1} - A_L^{\#} A_L)(\sigma - \sigma_o) \parallel^2 .$$

The maximum with respect to $\sigma \in \partial f(x_o)$ is certainly lower than the sup extended to the cone $\sigma_o + \mathbb{R}^+(\partial f(x_o) - \sigma_o)$:

$$E \leq \sup_{\sigma \in \partial f(x_o)} \sup_{s \geq 0} \{ s(\sigma - \sigma_o)^T A_L^{\#} y - \tfrac{s^2}{2\Gamma} \parallel (\mathbf{1} - A_L^{\#} A_L)(\sigma - \sigma_o) \parallel^2 \}$$

$$\leq \begin{cases} 0 \text{ if } (\sigma - \sigma_o)^T A_L^{\#} y \leq 0, \\ \frac{\Gamma}{2} \frac{|(\sigma - \sigma_o)^T A_L^{\#} y|^2}{\|(\mathbf{1} - A_L^{\#} A_L)(\sigma - \sigma_o)\|^2} \text{ otherwise.} \end{cases}$$

This, in (12), completes the proof.

REMARK 4. *The operator* $\mathbf{1} - A_L^{\#} A_L$ *is the orthogonal projection on* $\ker A_L$*; then we have* $(\mathbf{1} - A_L^{\#} A_L)\sigma = 0$ *if and only if* $\sigma \in \text{range}(A_L^{\#}) = \text{range}(A_L^T)$*. On the other hand, following optimality conditions* (2)*, any* $\sigma \in \Sigma$ *which is an element of* $\partial f(x_o)$ *cannot be an element of range*(A_L^T)*. Then* $(\mathbf{1} - A_L^{\#} A_L)(\sigma - \sigma_o) = (\mathbf{1} - A_L^{\#} A_L)\sigma \neq 0$ *for any* $\sigma \in \Sigma$*.*

5. Lipschitz-type stability of the optimal solution. We are now ready for the main result.

THEOREM 5.1. *Suppose the assumptions of superquadratic and subquadratic growth conditions hold and the MFC is satisfied. Then there exists* $\varepsilon > 0$ *such that for all perturbations* $y \in \mathbb{R}^m, \parallel y \parallel \leq \varepsilon$*, there exist* $\sigma_o \in \partial f(x_o)$ *and some full row ranked submatrix* A_L *such that the following holds:*

1. $\text{rank}(A_L) = \text{rank}(A)$*,* $\sigma_o \in \text{range}(A_L)$*, and* $Dv(0; y) = \sigma_o^T A_L^{\#} y$*,*
2. *if* $\Sigma = \{\sigma \in \partial f(x_o) : \sigma^T A_L^{\#} y > Dv(0; y)\}$*, then for any optimal solution* $x(y)$*, the following alternative holds: either* $\Sigma = \emptyset$*, in which case*

$$\parallel x(y) - x(0) \parallel \leq \sqrt{\frac{\Gamma}{\rho}} \parallel A_L^{\#} \parallel \parallel y \parallel,$$

or $\Sigma \neq \emptyset$, *in which case*

$$\| x(y) - x(0) \| \leq \sqrt{\frac{\Gamma}{\rho}\left(\| A_L^{\#} \|^2 + \sup_{\sigma \in \Sigma} \frac{\| (\sigma - \sigma_o)^T A_L^{\#} \|^2}{\| (\mathbf{1} - A_L^{\#} A_L)(\sigma - \sigma_o) \|^2} \right)} \ \| y \| .$$

Proof. For any $\sigma \in \partial f(x_o)$, by the superquadratic growth condition

$$\frac{\rho}{2} \| x(y) - x_o \|^2 \leq f(x(y)) - f(x_o) - \sigma^T (x(y) - x_o).$$

By (9), σ may be chosen such that $\sigma = A^T \lambda$, where $\lambda \in \partial v(0)$; taking into account that $f(x(y)) = v(y)$ and $f(x_o) = v(0)$

$$\frac{\rho}{2} \| x(y) - x_o \|^2 \leq v(y) - v(0) - \lambda^T A(x(y) - x_o).$$

The multiplier λ and some submatrix A_L may be chosen as in Theorem 4.4; then

$$\frac{\rho}{2} \| x(y) - x_o \|^2 \leq v(y) - v(0) - Dv(0; y)$$

$$\leq \begin{cases} \frac{\Gamma}{2} \| A_L^{\#} \|^2 \| y \|^2 & \text{if } \Sigma = \emptyset, \\ \frac{\Gamma}{2} (\| A_L^{\#} \|^2 + \sup_{\sigma \in \Sigma} \frac{\| (\sigma - \sigma_o)^T A_L^{\#} \|^2}{\| (\mathbf{1} - A_L^{\#} A_L)(\sigma - \sigma_o) \|^2}) \| y \|^2 & \text{otherwise.} \end{cases}$$

This completes the proof.

REMARK 5. *The submatrix* A_L *is not known a priori, but the index set* L *includes* $\{1, 2, \ldots, r\}$ *and indices of binding constraints such that* $\lambda_i > 0$ $(r + 1 \leq i \leq m)$, *for some* λ *such that* $\lambda^T y = Dv(0; y)$. *In the following classical smooth case, the matrix* A *is known. Suppose that the assumptions of superquadratic and subquadratic growth conditions hold, assume the function* f *is differentiable at the point* x_o, *and assume that the so-called independence condition holds. The family* $A^i (1 \leq i \leq m,\ A^i x_o = 0)$ *is linearly independent; then any optimal solution* $x(y)$ *is upper Lipschitz continuous with respect to the parameter* y *at the point* $y_o = 0$, *with the rate*

$$\| x(y) - x_o \| \ \leq \sqrt{\frac{\Gamma}{\rho}} \| A^{\#} \| \| y \|,$$

where A *is the matrix corresponding to active constraints. If the function* f *is twice continuously differentiable, then* Γ *and* ρ *are related to the highest and the lowest eigenvalues of the Hessian matrix of the function* f.

COROLLARY 5.2. *Suppose the assumptions of superquadratic and subquadratic growth conditions hold and the MFC is satisfied; assume* $f = \max_{1 \leq j \leq p} f_j$, *where the functions* f_j *are differentiable, and assume that the independence condition holds; then the optimal solution is upper Lipschitz continuous with respect to the parameter* y *at the point* 0 *with the rate, in case* $\Sigma = \emptyset$:

$$\| x(y) - x_o \| \ \leq \sqrt{\frac{\Gamma}{\rho}} \| A^{\#} \| \| y \|$$

and with the rate, in case $\Sigma \neq \emptyset$:

$$\| x(y) - x_o \| \ \leq \sqrt{\frac{\Gamma}{\rho}\left(\| A^{\#} \|^2 + \max_{\sigma \in \Sigma} \frac{\| (\sigma - \sigma_o)^T A^{\#} \|^2}{\| (\mathbf{1} - A^{\#} A)\sigma \|^2} \right)} \ || y || .$$

Here A *denotes the matrix corresponding to active constraints.*

5.1. Examples. Now we will answer two questions by examples.

1. In the case $\Sigma \neq \emptyset$ the upper bound may be $+\infty$; does there then exist any program with no upper Lipschitz continuity at the point 0?
2. Is it possible to relax the subquadratic growth assumption in the case where the function f is assumed to be differentiable?

We will answer both questions with the help of the following example in $\mathbb{R}^2$. We denote

$$w(x_1, x_2) = x_1 + \sqrt{x_1^2 + x_2^2}.$$

The subdifferential of w at that point $(0,0)$ is the closed disc C with center at $(0,1)$ and with radius 1.

For $\alpha \geq 1$, consider the program

$$\min f(x_1, x_2) = w(x_1, x_2)^\alpha + \frac{1}{2}(x_1^2 + x_2^2) \text{ s.t. } x_2 = y.$$

The objective function f is differentiable for $\alpha > 1$ and $\partial f(0,0) = C$ for $\alpha = 1$. On the other hand the quadratic growth conditions hold for $\alpha = 1$ but the subquadratic growth condition vanishes for $2 > \alpha > 1$. For $y = 0$, the optimal solution is $(0,0)$ and the unique subgradient which takes part in the optimality conditions is

$$\sigma_o = \begin{bmatrix} 0 \\ 0 \end{bmatrix}.$$

1. For the case where $\alpha = 1$, $y > 0$ we have

$$\sup_{\sigma \in \Sigma} \frac{\| \sigma - \sigma_o \|}{\| (\mathbf{1} - A^{\#}A)\sigma \|} = +\infty$$

because the sup is taken on the upper half disc. Optimal solutions are easily built by geometrical argument. Optimality holds at $(x_1, x_2) \neq (0,0)$ as the gradient of f

$$\nabla f(x_1, x_2) = \begin{bmatrix} 1 + x_1 + \frac{x_1}{\sqrt{x_1^2 + x_2^2}} \\ x_2 + \frac{x_2}{\sqrt{x_1^2 + x_2^2}} \end{bmatrix}$$

is normal to the line $\{x_2 = y\}$; therefore

$$1 + x_1 + \frac{x_1}{\sqrt{x_1^2 + x_2^2}} = 0$$

or else, with polar coordinates by taking $x_1 = \rho \cos\theta$ and $x_2 = \rho \sin\theta$, we have the following:

$$\rho = -1 - \frac{1}{\cos\theta}.$$

The unique optimal solution for the perturbation y stands at the intersection with the line $\{x_2 = y\}$. Since the curve is tangent to the half line $\{x_2 = 0, x_1 < 0\}$ for $\theta = \pi$, it is clear that the optimal solution cannot be upper Lipschitz continuous at $y = 0$. This example answers the first question in showing that the bound $+\infty$ is reached.

We can also remark that the linearized program has no optimal solution and by [3] we know that upper Lipschitzian continuity cannot hold.

2. For the case where $\alpha = \frac{3}{2}$ the optimal solutions stand at the crossing of the line $\{x_2 = y\}$ and the curve

$$\rho = \frac{9}{4}\frac{(1+\cos\theta)^2}{\cos^2\theta} \qquad \left(-\frac{\pi}{2} < \theta < \frac{3\pi}{2}\right).$$

Once again the optimal solution stands on the curve which is tangent at $(0,0)$ to the first axis, there is then no upper Lipschitz continuity at $\{y = 0\}$. This shows that the subquadratic growth condition cannot be relaxed, which was the second question.

Acknowledgments. We are thankful to both referees for their helpful comments and particularly to the one who helped us for organizing our results with respect to related works by [3] and [18].

REFERENCES

[1] J.P. Aubin, *Lipschitz behavior of the solution to convex optimization problems*, Math. Oper. Res., 9 (1984), pp. 87–111.
[2] A. Auslender and R. Cominetti, *First and second order sensitivity analysis of linear programs under directional constraints qualification condition*, Optimization, 21 (1990), pp. 351–363.
[3] J.F. Bonnans and R. Cominetti, *Perturbed optimization in Banach spaces* I: *A general theory based on a weak directionnal constraints qualification*, SIAM J. Control Optim., 34 (1996), pp. 1151–1171.
[4] J.F. Bonnans and A.D. Ioffe, *Quadratic growth and stability in convex programming problems with multiple solutions*, J. Convex Anal., 2 (1995), pp. 41–57.
[5] A.L. Dontchev and W.W. Hager, *Implicit functions, Lipschitz maps and stability in optimization*, Math. Oper. Res., 19 (1994), pp. 753–768.
[6] A.V. Fiacco and G.P. McCormick, *Nonlinear Programming: Sequential Unconstrained Minimization Techniques*, Wiley, New York, 1968.
[7] J. Gauvin and R. Janin, *Directional behavior of the optimal solutions in nonlinear mathematical programming*, Math. Oper. Res., 13 (1988), pp. 629–649.
[8] J. Gauvin and R. Janin, *Directional Lipschitzian optimal solutions and directional derivative for the optimal value function in nonlinear mathematical programming*, Ann. Inst. H. Poincaré Anal. Non Linéaire, 6 (1989), Supp., pp. 305–324.
[9] J.B. Hiriart-Urruty and C. Lemarechal, *Convex Analysis and Minimization Algorithms*, Springer-Verlag, Berlin, 1993.
[10] R. Janin and J. Gauvin, *Lipschitz dependence of the optimal solution to elementary convex programs*, in Proceedings of the 2nd Catalan Days on Applied Mathematics, Font-Romeu, Odeillo, France, M. Sofonea, ed., Presses Universitaires de Perpignan, Perpignan, France, 1995, pp. 149–161.
[11] L. Jiming, *Linear stability of generalized equations parts* I *and* II, Math. Oper. Res., 19 (1994), pp. 706–742.
[12] K. Jittorntrum, *Solution point differentiability without strict complementarity in nonlinear programming*, Math. Programming, 21 (1984), pp. 127–138.
[13] R. Pallu de la Barrière, *Cours d'Automatique Theorique*, Dunod, Paris, 1965.
[14] R.T. Rockafellar, *Convex Analysis*, Princeton University Press, Princeton, NJ, 1970.
[15] S.M. Robinson, *Stability theory for systems of inequalities, part* II*: Differentiable nonlinear systems*, SIAM J. Numer. Anal., 13 (1976), pp. 497–513.
[16] S.M. Robinson, *Strongly regular generalized equations*, Math. Oper. Res., 5 (1980), pp. 43–62.
[17] S.M. Robinson, *An implicit function theorem for a class of nonsmooth functions*, Math. Oper. Res., 16 (1991), pp. 292–309.
[18] A. Shapiro, *On Lipschitzian stability of optimal solutions of parametrized semi-infinite programs*, Math. Oper. Res., 19 (1994), pp. 743–752.
[19] A. Shapiro, *Second order sensitivity analysis and asymptotic theory of parametrized, nonlinear programs,* Math. Programming, 33 (1985), pp. 280–299.
[20] A. Shapiro, *Sensitivity analysis of nonlinear programs and differentiability properties of metric projections*, SIAM J. Control Optim., 26 (1988), pp. 628–645.

SIAM J. CONTROL OPTIM.
Vol. 38, No. 1, pp. 138–158

INDIRECT OBSTACLE CONTROL PROBLEM FOR SEMILINEAR ELLIPTIC VARIATIONAL INEQUALITIES*

QIHONG CHEN†

Abstract. An optimal control problem for a coupled system of a semilinear elliptic equation and an obstacle variational inequality is considered. Existence and optimality conditions of optimal triple are established.

Key words. optimal control, existence, optimality condition, elliptic obstacle variational inequality

AMS subject classifications. 49J20, 49K20

PII. S0363012998343379

1. Introduction.

In this paper, we investigate an optimal control problem in which the state y is governed by a controlled semilinear elliptic obstacle variational inequality,

$$\begin{cases} y \in H^2(\Omega) \cap H_0^1(\Omega), & \\ Ay \geq f(x,y,u) & \text{in} \quad \Omega, \\ y \geq \varphi & \text{in} \quad \Omega, \\ (Ay - f)(y - \varphi) = 0 & \text{in} \quad \Omega, \end{cases} \tag{1.1}$$

where the obstacle φ solves another semilinear elliptic equation with distributed control,

$$\begin{cases} A\varphi = g(x,\varphi,u) & \text{in} \quad \Omega, \\ \varphi|_{\partial\Omega} = 0, \end{cases} \tag{1.2}$$

and the cost functional is given by

$$J(y,\varphi,u) = \int_\Omega L(x,y(x),\varphi(x),u(x))dx, \tag{1.3}$$

where (y,φ,u) is a triple of state and control satisfying (1.1) and (1.2). Hence, the control problem is

$$\text{minimize} \quad J(y,\varphi,u) \quad \text{subject to}\,(1.1),(1.2) \qquad \text{and} \quad u(x) \in U \,\text{a.e. in}\,\Omega. \tag{1.4}$$

Optimal control problems for variational inequalities have been discussed by many authors in different aspects. See [2], [3], [4], [8], [9], [13], [15], [17], and [19] for examples. Some standard results for variational inequalities can be found in [7], [12], [16], and [18].

*Received by the editors August 10, 1998; accepted for publication (in revised form) March 16, 1999; published electronically December 7, 1999. This research is partly supported by the National Key Project of China, the Education Ministry Science Foundation of China, and the Natural Science Foundation of China. The author was also supported by the Science and Technology Foundation of Shanghai Higher Education.

http://www.siam.org/journals/sicon/38-1/34337.html

†Institute of Mathematics, Fudan University, Shanghai 200433, China and Shanghai Normal University, Shanghai 200023, China (chenqih@online.sh.ch).

Recently, the obstacle control problem has been studied in [1], where the obstacle is taken to be the control and the solution to the obstacle problem is taken to be the state; existence, uniqueness, and some characterizations of the optimal pairs were established based on the properties of the homogeneous state equation.

In our problem, if f does not explicitly depend on u, the action of control is *indirect* in the following sense: the control u acts upon the state y by means of the obstacle φ, the solution of controlled equation (1.2). This amounts to a design of the shape of the string by choosing a suitable "curvature" of the obstacle, if the one-dimensional obstacle problem of the string is considered. Although our problem setting is mathematically presented, our state system (1.1)–(1.2) is meaningful in practice. For instance, some free boundary problems such as the problem of the filtration of water through a homogeneous porous dam and the stationary water cone problem arising in the production of oil from a reservoir with bottom water, etc., after a Baiocchi-type transformation, can be reformulated in terms of the elliptic variational unilateral problem (cf. [16] and [20]). Also, a parabolic counterpart of (1.1)–(1.2) may appear in the studying of mathematical finance (such as the pricing of interest rate derivatives, etc.; cf. [11]).

The plan of this paper is as follows. In section 2, we state the main hypotheses and results, establish a $W^{2,p}$-estimate of state, and prove the continuity of the state with respect to the control variable. Section 3 is devoted to the existence of optimal triples. We study the approximate problems in section 4 and derive the optimality conditions (in the form of Pontryagin's principle) in section 5. Finally, section 6 is concerned with the case in which the control domain is convex.

2. State equation.

2.1. Main hypotheses and results. First, we introduce the following assumptions.

(H_1) $\Omega \subset \mathbb{R}^n$ is a bounded region with $C^{1,1}$ boundary $\partial\Omega$, U is a Polish space (a separable complete metric space), and

$$\mathcal{U} = \{u\colon \Omega \to U \,|\, u(\cdot) \text{ is measurable}\}. \tag{2.1}$$

(H_2) Operator A is defined by

$$Ay(x) = -\sum_{i,j=1}^{n} D_j(a_{ij}(x)D_i y(x)) \tag{2.2}$$

with $a_{ij} \in C^1(\overline{\Omega})$, $a_{ij} = a_{ji}$, $1 \le i, j \le n$, and for some $\lambda > 0$,

$$\sum_{i,j=1}^{n} a_{ij}(x)\xi_i\xi_j \ge \lambda \sum_{i=1}^{n} |\xi_i|^2 \quad \forall x \in \Omega,\ (\xi_1, \xi_2, \dots, \xi_n) \in \mathbb{R}^n. \tag{2.3}$$

(H_3) The functions $f, g\colon \Omega \times \mathbb{R} \times U \to \mathbb{R}$ have the following properties: $f(\cdot, y, u)$, $g(\cdot, \varphi, u)$ are measurable on Ω, and $f(x, \cdot, u), g(x, \cdot, u)$ are in $C^1(\mathbb{R})$ with $f(x, \cdot, \cdot)$, $f_y(x, \cdot, \cdot)$, $g(x, \cdot, \cdot)$, and $g_\varphi(x, \cdot, \cdot)$ continuous on $\mathbb{R} \times U$. Moreover, there exists a constant $K > 0$, such that

$$-K \le f_y,\ g_\varphi \le 0 \quad \text{on} \quad \Omega \times \mathbb{R} \times U \tag{2.4}$$

and

$$|f(x, 0, u)| + |g(x, 0, u)| \le K \quad \text{on} \quad \Omega \times U. \tag{2.5}$$

(H_4) The function $L: \Omega \times \mathbb{R} \times \mathbb{R} \times U \to \mathbb{R}$ satisfies the following: $L(\cdot, y, \varphi, u)$ is measurable on Ω, $L(x, \cdot, \cdot, u)$ is in $C^1(\mathbb{R} \times \mathbb{R})$ with $L(x, \cdot, \cdot, \cdot), L_y(x, \cdot, \cdot, \cdot)$, and $L_\varphi(x, \cdot, \cdot, \cdot)$ continuous on $\mathbb{R} \times \mathbb{R} \times U$, and for any $R > 0$, there exists a constant $K_R > 0$, such that

$$|L| + |L_y| + |L_\varphi| \le K_R \quad \text{on} \quad \Omega \times [-R, R] \times [-R, R] \times U. \tag{2.6}$$

Under (H_2), the operator A is associated with a positive symmetric bilinear form $a(\cdot, \cdot)$: $H_0^1(\Omega) \times H_0^1(\Omega) \to \mathbb{R}$

$$a(y, z) = \sum_{i,j=1}^{n} \int_\Omega a_{ij}(x) D_i y(x) D_j z(x) dx. \tag{2.7}$$

The weak solution of obstacle problem (1.1) can be defined as follows.

DEFINITION 2.1. *Given $u \in \mathcal{U}$ and $\varphi \in H_0^1(\Omega)$, a function $y \in H_0^1(\Omega)$ is called a weak solution of obstacle problem* (1.1)*, if*

$$\begin{cases} y \in K(\varphi), \\ a(y, z - y) \ge \int_\Omega f(x, y(x), u(x))(z - y) dx \quad \forall z \in K(\varphi), \end{cases} \tag{2.8}$$

where

$$K(\varphi) = \{ z \in H_0^1(\Omega) |\, z \ge \varphi \quad \text{a.e. in } \Omega \} \tag{2.9}$$

is a convex and closed subset of $H_0^1(\Omega)$.

Any element $u \in \mathcal{U}$ is referred to as a control. Any triple $(y, \varphi, u) \in H_0^1(\Omega) \times H_0^1(\Omega) \times \mathcal{U}$ satisfying (2.8) and (1.2) is called a feasible triple, and the corresponding (y, φ) and u will be referred to as a feasible state and a feasible control, respectively. The set of all feasible triples is denoted by $\mathcal{A}$. Clearly, under (H_1)–(H_4), $\mathcal{U}$ coincides with the set of all feasible controls, and for each $u \in \mathcal{U}$, there corresponds a unique feasible state (y, φ), and the cost functional (1.3) is well defined. Hereafter, we always assume (H_1)–(H_4). Thus, we can write $J(y, \varphi, u)$ as $J(u)$ without any ambiguity, and we will use $J(u)$ for convenience. Also we restate the control problem (1.4) as follows.

Problem (C). Find a feasible control $\overline{u} \in \mathcal{U}$, such that

$$J(\overline{u}) = \inf_{u \in \mathcal{U}} J(u). \tag{2.10}$$

If such a $\overline{u}$ exists, we call it an optimal control. Accordingly, the corresponding state $(\overline{y}, \overline{\varphi})$ and the feasible triple $(\overline{y}, \overline{\varphi}, \overline{u}) \in \mathcal{A}$ will be called an optimal state and triple, respectively.

Our main results on Problem (C) are the following two (see also Theorem 3.4 and Theorem 5.2).

EXISTENCE THEOREM. *In addition to* (H_1)–(H_4)*, we let the so-called Cesari property (see section* 3 *for details) hold. Then Problem* (C) *admits an optimal control $\overline{u} \in \mathcal{U}$.*

Pontryagin's principle. Let (H_1)–(H_4) hold and $(\overline{y}, \overline{\varphi}, \overline{u}) \in \mathcal{A}$ be an optimal triple for Problem (C). Then there exist $\overline{p}, \overline{q} \in H_0^1(\Omega)$ and $\overline{\mu} \in H^{-1}(\Omega) \cap \mathcal{M}(\overline{\Omega})$, such that

$$\operatorname{supp} \overline{\mu} \subset \{ x \in \Omega |\, \overline{y}(x) = \overline{\varphi}(x) \},$$

$$\begin{cases} A\overline{p} - f_y(x,\overline{y},\overline{u})\overline{p} = L_y(x,\overline{y},\overline{\varphi},\overline{u}) - \overline{\mu} & \text{in} \quad \Omega, \\ A\overline{q} - g_\varphi(x,\overline{\varphi},\overline{u})\overline{q} = L_\varphi(x,\overline{y},\overline{\varphi},\overline{u}) + \overline{\mu} & \text{in} \quad \Omega, \\ \overline{p}|_{\partial\Omega} = 0, \qquad \overline{q}|_{\partial\Omega} = 0, \end{cases}$$

and

$$H(x,\overline{y}(x),\overline{\varphi}(x),\overline{u}(x),\overline{p}(x),\overline{q}(x)) = \min_{u\in U} H(x,\overline{y}(x),\overline{\varphi}(x),u,\overline{p}(x),\overline{q}(x)) \quad \text{a.e. } x\in\Omega,$$

where $\mathcal{M}(\overline{\Omega})$ is the set of all regular signed measures on $\overline{\Omega}$, and

$$H(x,y,\varphi,u,p,q) = pf(x,y,u) + qg(x,\varphi,u) + L(x,y,\varphi,u)$$

for any $(x,y,\varphi,u,p,q)\in\Omega\times\mathbb{R}\times\mathbb{R}\times U\times\mathbb{R}\times\mathbb{R}$.

2.2. $W^{2,p}$-estimate of state. Let us start with a $W^{2,p}$-estimate of state, which is useful in what follows.

PROPOSITION 2.2. *Let $(y,\varphi,u)\in\mathcal{A}$. Then for any $p\geq 2$,*

$$\|y\|_{W^{2,p}(\Omega)} \leq C_p, \tag{2.11}$$

$$\|\varphi\|_{W^{2,p}(\Omega)} \leq C_p, \tag{2.12}$$

where C_p is a constant independent of the control variable u.

Proof. Under (H_1)–(H_3), (2.12) is easily obtained from the standard L^p-estimate of elliptic equations (cf. [10]).

To prove (2.11), we define

$$\beta(r) = \begin{cases} 0, & 0\leq r<+\infty, \\ -r^2, & -\frac{1}{2}\leq r<0, \\ r+\frac{1}{4}, & -\infty<r<-\frac{1}{2}, \end{cases} \tag{2.13}$$

and we introduce a family of approximation of state equation (2.8):

$$\begin{cases} Ay_\varepsilon + \frac{1}{\varepsilon}\beta(y_\varepsilon-\varphi) = f(x,y_\varepsilon,u) & \text{in} \quad \Omega, \\ y_\varepsilon|_{\partial\Omega} = 0, \end{cases} \tag{2.14}$$

where φ solves (1.2).

It is seen that, for any given $u\in\mathcal{U}$, $\varphi\in W_0^{1,p}(\Omega)$, and $\varepsilon>0$, $(2.14)_\varepsilon$ is uniquely solvable in $W^{2,p}(\Omega)\cap W_0^{1,p}(\Omega)$. The set of all triples $(y_\varepsilon,\varphi,u)\in H_0^1(\Omega)\times H_0^1(\Omega)\times\mathcal{U}$ satisfying $(2.14)_\varepsilon$ and (1.2) will be denoted by $\mathcal{A}_\varepsilon$.

The estimate (2.11) results from the following two lemmas.

LEMMA 2.3. *Let $(y_\varepsilon,\varphi,u)\in\mathcal{A}_\varepsilon$. Then for any $p\geq 2$,*

$$\|\beta(y_\varepsilon-\varphi)\|_{L^p(\Omega)} \leq \varepsilon C_p, \tag{2.15}$$

and consequently,

$$\|y_\varepsilon\|_{W^{2,p}(\Omega)} \leq C_p, \tag{2.16}$$

where C_p is a constant independent of $\varepsilon>0$ and $u\in\mathcal{U}$.

Proof. Define, for $r\in\mathbb{R}$, $B(r) = |\beta(r)|^{p-2}\beta(r)$. Then we have

$$B(r)\leq 0 \quad \text{in}\,\mathbb{R} \quad \text{and} \quad B(r) = 0 \quad \text{in} \quad \mathbb{R}^+, \tag{2.17}$$

$$B'(r) = (p-1)|\beta(r)|^{p-2}\beta'(r) \geq 0, \tag{2.18}$$

and, as $p \geq 2$, $\beta(0) = 0$, and $y_\varepsilon, \varphi \in W^{2,p}(\Omega) \cap W_0^{1,p}(\Omega)$,

$$B(y_\varepsilon - \varphi) \in W_0^{1,p}(\Omega) \hookrightarrow W_0^{1,p'}(\Omega), \tag{2.19}$$

where $p' = \frac{p}{p-1} \leq p$ is the conjugate number of p.

Multiplying $(2.14)_\varepsilon$ by $\varepsilon B(y_\varepsilon - \varphi)$ and integrating by parts, we obtain

$$\varepsilon a(y_\varepsilon, B(y_\varepsilon - \varphi)) + \int_\Omega |\beta(y_\varepsilon - \varphi)|^p dx = \varepsilon \int_\Omega f(x, y_\varepsilon, u) B(y_\varepsilon - \varphi) dx. \tag{2.20}$$

From (2.3), (2.17), (2.18), and the monotony of $f(x, \cdot, u)$, we see that

$$\int_\Omega f(x, y_\varepsilon, u) B(y_\varepsilon - \varphi) dx \leq \int_\Omega f(x, \varphi, u) B(y_\varepsilon - \varphi) dx \tag{2.21}$$

and

$$a(y_\varepsilon - \varphi, B(y_\varepsilon - \varphi)) \geq 0. \tag{2.22}$$

Then, by (1.2), (2.20)–(2.22), and Hölder's inequality, we have

$$\begin{aligned} &\|\beta(y_\varepsilon - \varphi)\|_{L^p(\Omega)}^p \\ &\leq \varepsilon \left\{ \int_\Omega f(x, \varphi, u) B(y_\varepsilon - \varphi) dx - a(\varphi, B(y_\varepsilon - \varphi)) \right\} \\ &= \varepsilon \int_\Omega [f(x, \varphi, u) - g(x, \varphi, u)] B(y_\varepsilon - \varphi) dx \\ &\leq \varepsilon \|f(\cdot, \varphi(\cdot), u(\cdot)) - g(\cdot, \varphi(\cdot), u(\cdot))\|_{L^p(\Omega)} \|\beta(y_\varepsilon - \varphi)\|_{L^p(\Omega)}^{p-1}. \end{aligned} \tag{2.23}$$

Thus, using (2.12) and (H_3), we get the desired estimate (2.15) with

$$C_p = \|f(\cdot, \varphi(\cdot), u(\cdot)) - g(\cdot, \varphi(\cdot), u(\cdot))\|_{L^p(\Omega)},$$

which is independent of $\varepsilon > 0$ and $u \in \mathcal{U}$.

Estimate (2.16) follows immediately from (2.15) and the elliptic L^p- estimate. □

LEMMA 2.4. *Let $(y_\varepsilon, \varphi, u) \in \mathcal{A}_\varepsilon$ and $(y, \varphi, u) \in \mathcal{A}$. Then, as $\varepsilon \to 0$,*

$$y_\varepsilon \to y \quad \textit{weakly in} W^{2,p}(\Omega) \textit{ and strongly in} W_0^{1,p}(\Omega).$$

The proof of Lemma 2.4 is similar to that in [1, Proposition 2.2(i)] with some suitable modifications. □

2.3. Continuous dependence of state on control. In the control set $\mathcal{U}$, we define the distance, called Ekeland's distance, as

$$d(u, v) = m(\{x \in \Omega | u(x) \neq v(x)\}) \quad \forall u, v \in \mathcal{U}, \tag{2.24}$$

where m denotes the Lebesgue measure. We can show that $(\mathcal{U}, d)$ is a complete metric space (cf. [13]).

The following result is concerned with the continuity of the state (y, φ) with respect to the control u under the above metric.

PROPOSITION 2.5. *Let* $(y,\varphi,u),(y_k,\varphi_k,u_k)\in\mathcal{A}\,(k=1,2,\ldots)$. *If* $d(u_k,u)\to 0$, *then for any* $p\geq 2$,

$$\|y_k-y\|_{W^{1,p}(\Omega)}\to 0 \tag{2.25}$$

and

$$\|\varphi_k-\varphi\|_{W^{1,p}(\Omega)}\to 0. \tag{2.26}$$

Proof. The convergence (2.26) is a known result in [13]. Here, we give a sketch of the proof for the reader's convenience.

Since the difference $\varphi_k-\varphi$ satisfies

$$\begin{cases} A(\varphi_k-\varphi)+a_k(x)(\varphi_k-\varphi)=g(x,\varphi,u_k)-g(x,\varphi,u),\\ (\varphi_k-\varphi)|_{\partial\Omega}=0, \end{cases} \tag{2.27}$$

where

$$a_k(x)=-\int_0^1 g_\varphi(x,\varphi(x)+\tau(\varphi_k(x)-\varphi(x)),u_k(x))d\tau,$$

satisfying (by (H_3))

$$0\leq a_k(x)\leq K\qquad\forall k, \tag{2.28}$$

we obtain, from the L^p-estimate for elliptic equations of divergence form,

$$\|\varphi_k-\varphi\|_{W^{1,p}(\Omega)}\leq C\|g(\cdot,\varphi(\cdot),u_k(\cdot))-g(\cdot,\varphi(\cdot),u(\cdot))\|_{W^{-1,p}(\Omega)}. \tag{2.29}$$

By Sobolev embedding and the duality, we have

$$\begin{cases} L^{\frac{np}{n+p}}(\Omega)\hookrightarrow W^{-1,p}(\Omega) & \text{for}\quad p>\frac{n}{n-1},\\ L^q(\Omega)\hookrightarrow W^{-1,p}(\Omega) & \text{for}\quad p=\frac{n}{n-1}. \end{cases} \tag{2.30}$$

Then, (2.29), together with (2.30), (2.12), and (H_3), gives

$$\|\varphi_k-\varphi\|_{W^{1,p}(\Omega)}\leq\begin{cases} C_p d(u_k,u)^{\frac{n+p}{np}} & \text{if}\quad p>\frac{n}{n-1},\\ C_{p,q}d(u_k,u)^{\frac{1}{q}}\quad\forall q>0 & \text{if}\quad p=\frac{n}{n-1}, \end{cases} \tag{2.31}$$

where the constants C_p and $C_{p,q}$ are independent of u and $u_k(k=1,2,\ldots)$. This proves (2.26).

From Proposition 2.2, we know that, for some subsequence,

$$y_k\to y^*\qquad\text{weakly in}\,W^{2,p}(\Omega),\quad\text{strongly in}\,W_0^{1,p}(\Omega).$$

Clearly,

$$y^*(x)\geq\varphi(x)\qquad\text{a.e.}\quad x\in\Omega. \tag{2.32}$$

Now, for any $z\in K(\varphi)$, let $z_k=z\vee\varphi_k$; we have $z_k\in K(\varphi_k)$ and $z_k\to z$ strongly in $H_0^1(\Omega)$. Note that

$$\begin{aligned}&\|f(\cdot,y_k(\cdot),u_k(\cdot))-f(\cdot,y^*(\cdot),u(\cdot))\|_{L^2(\Omega)}\\ &\leq\|f(\cdot,y_k(\cdot),u_k(\cdot))-f(\cdot,y_k(\cdot),u(\cdot))\|_{L^2(\Omega)}+\|f(\cdot,y_k(\cdot),u(\cdot))-f(\cdot,y^*(\cdot),u(\cdot))\|_{L^2(\Omega)}\\ &\leq C\{d(u_k,u)^{\frac{1}{2}}+\|y_k-y^*\|_{L^2(\Omega)}\}\to 0.\end{aligned}$$

Passing to the limit in (2.8), in which u, y, and z are replaced by u_k, y_k, and z_k, respectively, we obtain

$$a(y^*, z - y^*) \geq \int_\Omega f(x, y^*(x), u(x))(z - y^*)dx.$$

This, combined with (2.32), means that y^* is a solution of (2.8). By the uniqueness, we must have that $y^* = y$, and the whole sequence $\{y_k\}$ converges to y strongly in $W_0^{1,p}(\Omega)$. □

COROLLARY 2.6. *$J(u)$ is continuous on $(\mathcal{U}, d)$.*

3. Existence.

3.1. Cesari property and measurable selection theorem. Let us first recall the following.

DEFINITION 3.1 (see [5], [13]). *Let Y be a Banach space and Z be a metric space. Let $\Lambda\colon Z \to 2^Y$ be a multifunction. We say that Λ possesses the Cesari property at $z \in Z$ if*

$$\cap_{\delta>0} \overline{co}\Lambda(O_\delta(z)) = \Lambda(z), \tag{3.1}$$

where $\overline{co}E$ stands for the closed convex hull of the set E and $O_\delta(z)$ is the δ-neighborhood of the point z. If Λ has the Cesari property at every point $z \in Z$, we simply say that Λ has the Cesari property on Z.

DEFINITION 3.2. *Let $\Omega \subset \mathbb{R}^n$ be some Lebesgue measurable set and U be a Polish space. Let $\Gamma\colon \Omega \to 2^U$ be a multifunction. Function $u\colon \Omega \to U$ is called a selection of $\Gamma(\cdot)$ if*

$$u(x) \in \Gamma(x) \quad a.e. \quad x \in \Omega. \tag{3.2}$$

If such a u is measurable, then u is called a measurable selection of $\Gamma(\cdot)$.

The following gives the existence of measurable selections.

THEOREM 3.3. *Let $\Gamma\colon \Omega \to 2^U$ be measurable, taking closed set values. Then $\Gamma(\cdot)$ admits a measurable selection.*

We refer the readers to [13, pp. 100–101] for the proof of Theorem 3.3.

3.2. Existence of optimal controls. To establish the existence for problem (C), we first introduce the following set:

$$\Lambda(x, y, \varphi) = \{(\xi, \eta, \zeta) \in \mathbb{R}^3 | \, \xi \geq L(x, y, \varphi, u), \eta = f(x, y, u), \zeta = g(x, \varphi, u), u \in U\}. \tag{3.3}$$

We make the following assumption:

(H_5) For almost all $x \in \Omega$, the mapping $(y, \varphi) \mapsto \Lambda(x, y, \varphi)$ has the Cesari property on $\mathbb{R}^2$.

THEOREM 3.4. *Let (H_1)–(H_5) hold. Then Problem (C) admits at least one optimal control $\overline{u} \in \mathcal{U}$.*

Proof. Let $\{u_k\} \subset \mathcal{U}$ be a minimizing sequence satisfying

$$J(u_k) \leq \inf_{u \in \mathcal{U}} J(u) + \frac{1}{k}. \tag{3.4}$$

Take $p > \max\{\frac{n}{2}, 2\}$. By Proposition 2.2, we know that the corresponding state (y_k, φ_k) satisfies

$$\|y_k\|_{W^{2,p}(\Omega)} + \|\varphi_k\|_{W^{2,p}(\Omega)} \leq C_p, \tag{3.5}$$

with C_p being independent of k. Thus, we may let, extracting some subsequence if necessary,

$$\left\{\begin{array}{l} y_k \to \overline{y}, \\ \varphi_k \to \overline{\varphi}, \end{array}\right. \tag{3.6}$$

weakly in $W^{2,p}(\Omega)$, strongly in $W_0^{1,p}(\Omega)$, and $C^\alpha(\overline{\Omega})$ for some $\alpha \in (0,1)$ and some $(\overline{y},\overline{\varphi}) \in [W^{2,p}(\Omega) \cap W_0^{1,p}(\Omega)]^2$. By (3.5) and ($H_3$), the functions $f(\cdot, y_k(\cdot), u_k(\cdot))$ and $g(\cdot, \varphi_k(\cdot), u_k(\cdot))$ are uniformly bounded. Hence, we may further assume

$$\left\{\begin{array}{ll} f(\cdot, y_k(\cdot), u_k(\cdot)) \to \overline{f}(\cdot) & \text{weakly in} \quad L^p(\Omega), \\ g(\cdot, \varphi_k(\cdot), u_k(\cdot)) \to \overline{g}(\cdot) & \text{weakly in} \quad L^p(\Omega) \end{array}\right. \tag{3.7}$$

for some $\overline{f}, \overline{g} \in L^\infty(\Omega)$. Then, by the Mazur theorem, we can find $\alpha_{ij} \geq 0$, $\sum_{i\geq 1} \alpha_{ij} = 1 \, \forall j$, such that

$$\left\{\begin{array}{ll} \eta_j(\cdot) = \sum_{i\geq 1} \alpha_{ij} f(\cdot, y_{i+j}(\cdot), u_{i+j}(\cdot)) \to \overline{f}(\cdot) & \text{strongly in} \quad L^p(\Omega), \\ \zeta_j(\cdot) = \sum_{i\geq 1} \alpha_{ij} g(\cdot, \varphi_{i+j}(\cdot), u_{i+j}(\cdot)) \to \overline{g}(\cdot) & \text{strongly in} \quad L^p(\Omega). \end{array}\right. \tag{3.8}$$

Set

$$\xi_j(\cdot) = \sum_{i\geq 1} \alpha_{ij} L(\cdot, y_{i+j}(\cdot), \varphi_{i+j}(\cdot), u_{i+j}(\cdot)) \tag{3.9}$$

and

$$\overline{L}(x) = \underline{\lim}_{j\to\infty} \xi_j(x) \qquad \text{a.e. } x \in \Omega. \tag{3.10}$$

The convergence (3.6) implies that, for any $\delta > 0$, there exists a j_0 such that for $j \geq j_0$,

$$(\xi_j(x), \eta_j(x), \zeta_j(x)) \in co\Lambda(x, O_\delta((\overline{y}(x), \overline{\varphi}(x)))) \qquad \text{a.e. } x \in \Omega. \tag{3.11}$$

Thus, for any $\delta > 0$, we have

$$(\overline{L}(x), \overline{f}(x), \overline{g}(x)) \in \overline{co}\Lambda(x, O_\delta((\overline{y}(x), \overline{\varphi}(x)))) \qquad \text{a.e. } x \in \Omega, \tag{3.12}$$

and then, by (H_5),

$$(\overline{L}(x), \overline{f}(x), \overline{g}(x)) \in \Lambda(x, \overline{y}(x), \overline{\varphi}(x)) \qquad \text{a.e. } x \in \Omega. \tag{3.13}$$

Now, making use of Theorem 3.3, we can find a $\overline{u} \in \mathcal{U}$ such that

$$\left\{\begin{array}{l} \overline{L}(x) \geq L(x, \overline{y}(x), \overline{\varphi}(x), \overline{u}(x)) \\ \overline{f}(x) = f(x, \overline{y}(x), \overline{u}(x)) \\ \overline{g}(x) = g(x, \overline{\varphi}(x), \overline{u}(x)) \end{array}\right. \qquad \text{a.e. } x \in \Omega. \tag{3.14}$$

We claim that $(\overline{y}(x), \overline{\varphi}(x))$ is the state corresponding to $\overline{u}$, i.e.,

$$(\overline{y}, \overline{\varphi}, \overline{u}) \in \mathcal{A}. \tag{3.15}$$

In fact, since $(y_k, \varphi_k, u_k) \in \mathcal{A}$, from the convergence (3.6), (3.7), and (3.14), we have

$$\left\{\begin{array}{ll} A\overline{\varphi} = g(x, \overline{\varphi}, \overline{u}) & \text{in} \quad \Omega, \\ \overline{\varphi}|_{\partial\Omega} = 0 \end{array}\right. \tag{3.16}$$

and

$$\overline{y}(x) \geq \overline{\varphi}(x) \qquad \text{a.e. } x \in \Omega. \tag{3.17}$$

For any $z \in K(\overline{\varphi})$, we have $z \vee \varphi_k \in K(\varphi_k)$ and

$$z \vee \varphi_k \to z \qquad \text{strongly in} \quad H_0^1(\Omega). \tag{3.18}$$

Then, the feasibility of (y_k, φ_k, u_k) gives

$$a(y_k, z \vee \varphi_k - y_k) \geq \int_\Omega f(x, y_k, u_k)(z \vee \varphi_k - y_k)dx \quad \forall k, \tag{3.19}$$

and the convergence (3.6), (3.7), (3.14), and (3.18) yields

$$a(\overline{y}, z - \overline{y}) \geq \int_\Omega f(x, \overline{y}, \overline{u})(z - \overline{y})dx. \tag{3.20}$$

Thus, by (3.16), (3.17), and (3.20), the feasibility (3.15) is verified.

Finally, we can deduce from (3.14), (3.10), (3.9), (3.4), and Fatou's lemma that

$$\begin{aligned}
J(\overline{u}) &= \int_\Omega L(x, \overline{y}(x), \overline{\varphi}(x), \overline{u}(x))dx \\
&\leq \int_\Omega \overline{L}(x)dx = \int_\Omega \underline{\lim}_{j\to\infty} \xi_j(x)dx \\
&\leq \underline{\lim}_{j\to\infty} \int_\Omega \xi_j(x)dx = \underline{\lim}_{j\to\infty} \sum_{i\geq 1} \alpha_{ij} J(u_{i+j}) \\
&\leq \underline{\lim}_{j\to\infty} \sum_{i\geq 1} \alpha_{ij} (\inf_{u\in\mathcal{U}} J(u) + \frac{1}{j}) = \underline{\lim}_{j\to\infty} (\inf_{u\in\mathcal{U}} J(u) + \frac{1}{j}) \\
&= \inf_{u\in\mathcal{U}} J(u).
\end{aligned}$$

Hence $\overline{u}$ is an optimal control of Problem (C). □

4. Analysis of approximate problems.

4.1. Approximate functional. Let us introduce a family of approximate functional

$$J_\varepsilon(u) = J(y_\varepsilon, \varphi, u) = \int_\Omega L(x, y_\varepsilon(x), \varphi(x), u(x))dx, \tag{4.1}$$

where $(y_\varepsilon, \varphi, u) \in \mathcal{A}_\varepsilon$.

Proposition 4.1.

$$\begin{aligned}
&\text{(i)} \quad \text{For any fixed } \varepsilon > 0,\ J_\varepsilon(u) \text{is continuous on } (\mathcal{U}, d); \\
&\text{(ii)} \quad \text{For any given } u \in \mathcal{U}, \lim_{\varepsilon\to 0} J_\varepsilon(u) = J(u).
\end{aligned} \tag{4.2}$$

Proof. (i) It suffices to prove the continuous dependence of y_ε on u. Let $(y_\varepsilon, \varphi, u)$, $(y_{\varepsilon,k}, \varphi_k, u_k) \in \mathcal{A}_\varepsilon (k = 1, 2, \ldots)$. Supposing $d(u_k, u) \to 0$, we have already had that, for any $p \geq 2$, $\|\varphi_k - \varphi\|_{W^{1,p}(\Omega)} \to 0$ (see (2.26)). Adapting the proof of (2.26) to $y_{\varepsilon,k} - y_\varepsilon$, we can obtain

(4.3)

$$\|y_{\varepsilon,k} - y_\varepsilon\|_{W^{1,p}(\Omega)} \leq C \left\{ \frac{1}{\varepsilon} \|\varphi_k - \varphi\|_{L^p(\Omega)} + \|f(\cdot, y_\varepsilon(\cdot), u_k(\cdot)) - f(\cdot, y_\varepsilon(\cdot), u(\cdot))\|_{W^{-1,p}(\Omega)} \right\}.$$

Recalling (2.30), (2.16), and (H_3), we see that

$$\begin{aligned} &\|f(\cdot, y_\varepsilon(\cdot), u_k(\cdot)) - f(\cdot, y_\varepsilon(\cdot), u(\cdot))\|_{W^{-1,p}(\Omega)} \\ &\leq \begin{cases} C_p d(u_k, u)^{\frac{n+p}{np}} & \text{if} \quad p > \frac{n}{n-1}, \\ C_{p,q} d(u_k, u)^{\frac{1}{q}} \quad \forall q > 1 & \text{if} \quad p = \frac{n}{n-1}. \end{cases} \end{aligned} \tag{4.4}$$

Thus, we have

$$\|y_{\varepsilon,k} - y_\varepsilon\|_{W^{1,p}(\Omega)} \to 0. \tag{4.5}$$

(ii) (4.2) is an immediate consequence of Lemma 2.4 and (H_4). □

4.2. Taylor expansion. We note that for each $(y_\varepsilon, \varphi, u) \in \mathcal{A}_\varepsilon$, (y_ε, φ) is the solution of semilinear elliptic system $(2.14)_\varepsilon$ and (1.2). Using an argument similar to that in [13], we can present a sort of "Taylor expansion" formula for (y_ε, φ) and the approximate functional $J_\varepsilon(u)$.

PROPOSITION 4.2. *Let $(y_\varepsilon, \varphi, u) \in \mathcal{A}_\varepsilon$ and $v \in \mathcal{U}$ be fixed. Then, for any $\rho \in (0,1)$, there exists a measurable set $E^\rho \subset \Omega$ with $m(E^\rho) = \rho m(\Omega)$, such that if $u^\rho \in \mathcal{U}$ is defined by*

$$u^\rho(x) = \begin{cases} u(x) & \text{if} \quad x \in \Omega \setminus E^\rho, \\ v(x) & \text{if} \quad x \in E^\rho \end{cases} \tag{4.6}$$

and $(y_\varepsilon^\rho, \varphi^\rho, u^\rho) \in \mathcal{A}_\varepsilon$, then it holds that

$$\begin{cases} y_\varepsilon^\rho = y_\varepsilon + \rho z + r^\rho, \\ \lim_{\rho \to 0} \frac{1}{\rho} \|r^\rho\|_{W^{1,p}(\Omega)} = 0; \end{cases} \tag{4.7}$$

$$\begin{cases} \varphi^\rho = \varphi + \rho\zeta + \omega^\rho, \\ \lim_{\rho \to 0} \frac{1}{\rho} \|\omega^\rho\|_{W^{1,p}(\Omega)} = 0; \end{cases} \tag{4.8}$$

$$\begin{cases} J_\varepsilon(u^\rho) = J_\varepsilon(u) + \rho j + e^\rho, \\ \lim_{\rho \to 0} \frac{1}{\rho} |e^\rho| = 0; \end{cases} \tag{4.9}$$

where z, ζ, and j satisfy the following:

$$\begin{cases} Az - f_y(x, y_\varepsilon, u)z + \frac{1}{\varepsilon}\beta'(y_\varepsilon - \varphi)(z - \zeta) = f(x, y_\varepsilon, v) - f(x, y_\varepsilon, u) & \text{in } \Omega, \\ A\zeta - g_\varphi(x, \varphi, u)\zeta = g(x, \varphi, v) - g(x, \varphi, u) & \text{in } \Omega, \\ z|_{\partial\Omega} = 0, \qquad \zeta|_{\partial\Omega} = 0 \end{cases} \tag{4.10}$$

and

$$j = \int_\Omega [L_y(x, y_\varepsilon, \varphi, u)z + L_\varphi(x, y_\varepsilon, \varphi, u)\zeta + L(x, y_\varepsilon, \varphi, v) - L(x, y_\varepsilon, \varphi, u)]dx. \tag{4.11}$$

4.3. Convergence theorem. The main result of this section is the following convergence theorem.

THEOREM 4.3. *Let $\overline{J}_\varepsilon = \inf_{u \in \mathcal{U}} J_\varepsilon(u)$ and $\overline{J} = \inf_{u \in \mathcal{U}} J(u)$. Then*

$$\lim_{\varepsilon \to 0} \overline{J}_\varepsilon = \overline{J}. \tag{4.12}$$

To prove Theorem 4.3, we need the following:

LEMMA 4.4. *Let* $\{u_\varepsilon\} \subset \mathcal{U}$ *be any sequence,* $(y_\varepsilon, \varphi_\varepsilon, u_\varepsilon) \in \mathcal{A}_\varepsilon$, *and* $(y^\varepsilon, \varphi_\varepsilon, u_\varepsilon) \in \mathcal{A}$. *Then*

$$\lim_{\varepsilon \to 0} \|y_\varepsilon - y^\varepsilon\|_{H_0^1(\Omega)} = 0. \tag{4.13}$$

Proof. By Proposition 2.2 and Lemma 2.3, we have that, for any $p \geq 2$,

$$\|y_\varepsilon\|_{W^{2,p}(\Omega)} + \|\varphi_\varepsilon\|_{W^{2,p}(\Omega)} + \|y^\varepsilon\|_{W^{2,p}(\Omega)} \leq C_p, \tag{4.14}$$

with C_p independent of $\varepsilon > 0$. Thus, we may assume that, for some subsequence,

$$\begin{cases} y_\varepsilon \to y & \text{strongly in} \quad H_0^1(\Omega), \\ \varphi_\varepsilon \to \varphi & \text{strongly in} \quad H_0^1(\Omega) \end{cases} \tag{4.15}$$

and

$$\|y^\varepsilon\|_{H^1(\Omega) \cap L^\infty(\Omega)} \leq C \quad \text{with } C \text{ independent of } \varepsilon > 0. \tag{4.16}$$

For any $\eta \in H_0^1(\Omega)$ with $\eta \geq 0$ a.e. in Ω, we have from $(2.14)_\varepsilon$ that

$$0 \leq -\int_\Omega \beta(y_\varepsilon - \varphi_\varepsilon)\eta dx = \varepsilon \left\{ a(y_\varepsilon, \eta) - \int_\Omega f(x, y_\varepsilon, u_\varepsilon)\eta dx \right\} \to 0, \tag{4.17}$$

because the terms in $\{\,\}$ are bounded. Then, by Fatou's lemma,

$$\int_\Omega \beta(y - \varphi)\eta dx = 0 \quad \forall \eta \in H_0^1(\Omega) \text{ with } \eta \geq 0 \text{ a.e. in } \Omega. \tag{4.18}$$

This implies that

$$\beta(y - \varphi) = 0 \quad \text{a.e. in} \quad \Omega, \tag{4.19}$$

and, by the definition of $\beta(\cdot)$,

$$y \geq \varphi \quad \text{a.e. in} \quad \Omega. \tag{4.20}$$

Now, letting $z_\varepsilon = y_\varepsilon \vee \varphi_\varepsilon$, we have

$$z_\varepsilon \to y \qquad \text{strongly in} \quad H_0^1(\Omega), \tag{4.21}$$

and consequently,

$$\|z_\varepsilon - y_\varepsilon\|_{H_0^1(\Omega)} \to 0. \tag{4.22}$$

Recalling that y_ε and y^ε solve $(2.14)_\varepsilon$ and (2.8), respectively, we have

$$\begin{aligned} &a(y_\varepsilon, y_\varepsilon - y^\varepsilon) \\ &= -\tfrac{1}{\varepsilon} \int_\Omega \beta(y_\varepsilon - \varphi_\varepsilon)(y_\varepsilon - y^\varepsilon)dx + \int_\Omega f(x, y_\varepsilon, u_\varepsilon)(y_\varepsilon - y^\varepsilon)dx \end{aligned} \tag{4.23}$$

and

$$a(y^\varepsilon, z_\varepsilon - y^\varepsilon) \geq \int_\Omega f(x, y^\varepsilon, u_\varepsilon)(z_\varepsilon - y^\varepsilon)dx. \tag{4.24}$$

By the monotonicity of $f(x,\cdot,u)$ and $\beta(\cdot)$, we see that

$$\int_\Omega [f(x, y_\varepsilon, u_\varepsilon) - f(x, y^\varepsilon, u_\varepsilon)](y_\varepsilon - y^\varepsilon)dx \leq 0 \tag{4.25}$$

and

$$\int_\Omega \beta(y_\varepsilon - \varphi_\varepsilon)(y_\varepsilon - y^\varepsilon)dx \geq 0. \tag{4.26}$$

Here, we have used the fact that $y^\varepsilon \geq \varphi_\varepsilon > y_\varepsilon$ when $y_\varepsilon < \varphi_\varepsilon$. From (4.23)–(4.26), we may deduce that

$$a(y_\varepsilon - y^\varepsilon, y_\varepsilon - y^\varepsilon) \leq a(y^\varepsilon, z_\varepsilon - y_\varepsilon) - \int_\Omega f(x, y^\varepsilon, u_\varepsilon)(z_\varepsilon - y_\varepsilon)dx \leq C\|z_\varepsilon - y_\varepsilon\|_{H^1(\Omega)} \to 0,$$

where (4.16) and (4.22) have been used. This and (2.3) prove (4.13). □

Remark. The above lemma can be strengthened. As a matter of fact, using Lions's interpolation theorem (cf. [14]), we can deduce from (4.13) and (4.14) that, for any $p \geq 2$,

$$\lim_{\varepsilon \to 0} \|y_\varepsilon - y^\varepsilon\|_{W^{1,p}(\Omega)} = 0. \tag{4.27}$$

From Lemma 4.4 and (H_4), we easily get the following corollary.

COROLLARY 4.5. *For any sequence* $\{u_\varepsilon\} \subset \mathcal{U}$,

$$\lim_{\varepsilon \to 0}[J_\varepsilon(u_\varepsilon) - J(u_\varepsilon)] = 0. \tag{4.28}$$

Proof of Theorem 4.3. For any $\varepsilon > 0$, one can find $u_\varepsilon \in \mathcal{U}$, such that

$$J_\varepsilon(u_\varepsilon) < \overline{J}_\varepsilon + \varepsilon. \tag{4.29}$$

Then, by Corollary 4.5, we have

$$\underline{\lim}_{\varepsilon \to 0}\overline{J}_\varepsilon \geq \underline{\lim}_{\varepsilon \to 0} J_\varepsilon(u_\varepsilon) = \underline{\lim}_{\varepsilon \to 0}[J(u_\varepsilon) + J_\varepsilon(u_\varepsilon) - J(u_\varepsilon)] = \underline{\lim}_{\varepsilon \to 0} J(u_\varepsilon) \geq \overline{J}. \tag{4.30}$$

On the other hand, let $u_\delta \in \mathcal{U}$ be such that

$$J(u_\delta) < \overline{J} + \delta. \tag{4.31}$$

Then, using Corollary 4.5 again, we have

$$\overline{\lim}_{\delta \to 0}\overline{J}_\delta \leq \overline{\lim}_{\delta \to 0} J_\delta(u_\delta) = \overline{\lim}_{\delta \to 0}[J(u_\delta) + J_\delta(u_\delta) - J(u_\delta)] = \overline{\lim}_{\delta \to 0} J(u_\delta) \leq \overline{J}. \tag{4.32}$$

Hence, (4.12) follows from (4.30) and (4.32). □

5. Optimality condition. In this section, we first recall Ekeland's variational principle and then use it to derive some necessary conditions for optimal controls of Problem (C).

THEOREM 5.1 (Ekeland's variational principle; see [6]). *Let (V, d) be a complete metric space and $J: V \to (-\infty, +\infty]$ be a proper lower semicontinuous function bounded from below. Let $\alpha > 0, \overline{v} \in V$ such that*

$$J(\overline{v}) \le \inf_{v \in V} J(v) + \alpha^2.$$

Then there exists a $v_\alpha \in V$ satisfying

$$J(v_\alpha) \le J(\overline{v}), \tag{5.1}$$

$$d(v_\alpha, \overline{v}) \le \alpha, \tag{5.2}$$

$$-\alpha d(v, v_\alpha) \le J(v) - J(v_\alpha) \quad \forall v \in V. \tag{5.3}$$

THEOREM 5.2. *Let (H_1)–(H_4) hold and $(\overline{y}, \overline{\varphi}, \overline{u}) \in \mathcal{A}$ be an optimal triple for Problem (C). Then there exist $\overline{p}, \overline{q} \in H_0^1(\Omega)$ and $\overline{\mu} \in H^{-1}(\Omega) \cap \mathcal{M}(\overline{\Omega})$, such that*

$$\operatorname{supp} \overline{\mu} \subset \{\, x \in \Omega |\ \overline{y}(x) = \overline{\varphi}(x) \,\}, \tag{5.4}$$

$$\begin{cases} A\overline{p} - f_y(x, \overline{y}, \overline{u})\overline{p} = L_y(x, \overline{y}, \overline{\varphi}, \overline{u}) - \overline{\mu} & in\, \Omega, \\ A\overline{q} - g_\varphi(x, \overline{\varphi}, \overline{u})\overline{q} = L_\varphi(x, \overline{y}, \overline{\varphi}, \overline{u}) + \overline{\mu} & in\, \Omega, \\ \overline{p}|_{\partial\Omega} = 0, \qquad \overline{q}|_{\partial\Omega} = 0, \end{cases} \tag{5.5}$$

and

$$H(x, \overline{y}(x), \overline{\varphi}(x), \overline{u}(x), \overline{p}(x), \overline{q}(x)) = \min_{u \in U} H(x, \overline{y}(x), \overline{\varphi}(x), u, \overline{p}(x), \overline{q}(x)) \qquad a.e.\ x \in \Omega, \tag{5.6}$$

where $\mathcal{M}(\overline{\Omega})$ is the set of all regular signed measures on $\overline{\Omega}$, and

$$H(x, y, \varphi, u, p, q) = pf(x, y, u) + qg(x, \varphi, u) + L(x, y, \varphi, u)$$

for any $(x, y, \varphi, u, p, q) \in \Omega \times \mathbb{R} \times \mathbb{R} \times U \times \mathbb{R} \times \mathbb{R}$.

Condition (5.4) is understood as the following: For any $\eta \in C(\overline{\Omega})$ with $\operatorname{supp} \eta \subset \Omega_0$,

$$\langle \overline{\mu},\ \eta \rangle_{\mathcal{M}(\overline{\Omega}), C(\overline{\Omega})} = 0,$$

where

$$\Omega_0 = \{\, x \in \Omega |\ \overline{y}(x) > \overline{\varphi}(x) \,\}. \tag{5.7}$$

Proof. Given $\varepsilon > 0$ and

$$\alpha_\varepsilon = (J_\varepsilon(\overline{u}) - \overline{J}_\varepsilon + \varepsilon)^{1/2} > 0. \tag{5.8}$$

From (4.2) and (4.12), we see that (note $J(\overline{u}) = \overline{J}$)

$$\alpha_\varepsilon \to 0. \tag{5.9}$$

Since $J_\varepsilon(u)$ is continuous on $(\mathcal{U}, d)$ and

$$J_\varepsilon(\overline{u}) \le \overline{J}_\varepsilon + \alpha_\varepsilon^2 = \inf_{u\in\mathcal{U}} J_\varepsilon(u) + \alpha_\varepsilon^2 \tag{5.10}$$

by Ekeland's variational principle, there exists a $u_\varepsilon \in \mathcal{U}$, such that

$$d(u_\varepsilon, \overline{u}) \le \alpha_\varepsilon, \tag{5.11}$$

$$-\alpha_\varepsilon d(u, u_\varepsilon) \le J_\varepsilon(u) - J_\varepsilon(u_\varepsilon) \quad \forall u \in \mathcal{U}. \tag{5.12}$$

Let $(y_\varepsilon, \varphi_\varepsilon, u_\varepsilon) \in \mathcal{A}_\varepsilon$ and $v \in \mathcal{U}$ be fixed. By Proposition 4.2 we know that, for any $\rho \in (0,1)$, there exists a measurable set $E^\rho \subset \Omega$ with $m(E^\rho) = \rho m(\Omega)$, such that if we define

$$u_\varepsilon^\rho(x) = \begin{cases} u_\varepsilon(x) & \text{if} \quad x \in \Omega \setminus E^\rho, \\ v(x) & \text{if} \quad x \in E^\rho \end{cases} \tag{5.13}$$

and let $(y_\varepsilon^\rho, \varphi_\varepsilon^\rho, u_\varepsilon^\rho) \in \mathcal{A}_\varepsilon$, then

$$\begin{cases} y_\varepsilon^\rho = y_\varepsilon + \rho z_\varepsilon + r_\varepsilon^\rho, \\ \varphi_\varepsilon^\rho = \varphi_\varepsilon + \rho\zeta_\varepsilon + \omega_\varepsilon^\rho, \\ J_\varepsilon(u_\varepsilon^\rho) = J_\varepsilon(u_\varepsilon) + \rho j_\varepsilon + e_\varepsilon^\rho, \end{cases} \tag{5.14}$$

where $z_\varepsilon, \zeta_\varepsilon$, and j_ε satisfy the following:

(5.15)
$$\begin{cases} Az_\varepsilon - f_y(x, y_\varepsilon, u_\varepsilon)z_\varepsilon + \frac{1}{\varepsilon}\beta'(y_\varepsilon - \varphi_\varepsilon)(z_\varepsilon - \zeta_\varepsilon) = f(x, y_\varepsilon, v) - f(x, y_\varepsilon, u_\varepsilon) & \text{in } \Omega, \\ A\zeta_\varepsilon - g_\varphi(x, \varphi_\varepsilon, u_\varepsilon)\zeta_\varepsilon = g(x, \varphi_\varepsilon, v) - g(x, \varphi_\varepsilon, u_\varepsilon) & \text{in } \Omega, \\ z_\varepsilon|_{\partial\Omega} = 0, \qquad \zeta_\varepsilon|_{\partial\Omega} = 0 \end{cases}$$

and

(5.16)
$$j_\varepsilon = \int_\Omega [L_y(x, y_\varepsilon, \varphi_\varepsilon, u_\varepsilon)z_\varepsilon + L_\varphi(x, y_\varepsilon, \varphi_\varepsilon, u_\varepsilon)\zeta_\varepsilon + L(x, y_\varepsilon, \varphi_\varepsilon, v) - L(x, y_\varepsilon, \varphi_\varepsilon, u_\varepsilon)]dx,$$

with

$$\lim_{\rho\to 0} \frac{1}{\rho}\|r_\varepsilon^\rho\|_{W^{1,p}(\Omega)} = \lim_{\rho\to 0} \frac{1}{\rho}\|\omega_\varepsilon^\rho\|_{W^{1,p}(\Omega)} = \lim_{\rho\to 0} \frac{1}{\rho}|e_\varepsilon^\rho| = 0. \tag{5.17}$$

Now, we take $u = u_\varepsilon^\rho$ in (5.12). It follows that

$$-\alpha_\varepsilon m(\Omega) \le \frac{1}{\rho}[J_\varepsilon(u_\varepsilon^\rho) - J_\varepsilon(u_\varepsilon)] \to j_\varepsilon \qquad (\rho \to 0). \tag{5.18}$$

Let $(p_\varepsilon, q_\varepsilon) \in H_0^1(\Omega) \times H_0^1(\Omega)$ be the unique solution of the following system:

$$\begin{cases} Ap_\varepsilon + [\frac{1}{\varepsilon}\beta'(y_\varepsilon - \varphi_\varepsilon) - f_y(x, y_\varepsilon, u_\varepsilon)]p_\varepsilon = L_y(x, y_\varepsilon, \varphi_\varepsilon, u_\varepsilon) & \text{in } \Omega, \\ Aq_\varepsilon - g_\varphi(x, \varphi_\varepsilon, u_\varepsilon)q_\varepsilon - \frac{1}{\varepsilon}\beta'(y_\varepsilon - \varphi_\varepsilon)p_\varepsilon = L_\varphi(x, y_\varepsilon, \varphi_\varepsilon, u_\varepsilon) & \text{in } \Omega, \\ p_\varepsilon|_{\partial\Omega} = 0, \qquad q_\varepsilon|_{\partial\Omega} = 0. \end{cases} \tag{5.19}$$

Then, we may deduce from (5.16) and (5.18) that

$$\int_\Omega [H(x, y_\varepsilon, \varphi_\varepsilon, v, p_\varepsilon, q_\varepsilon) - H(x, y_\varepsilon, \varphi_\varepsilon, u_\varepsilon, p_\varepsilon, q_\varepsilon)]dx \ge -\alpha_\varepsilon m(\Omega). \tag{5.20}$$

In what follows, we are going to take the limits to get the final result.

Noting that $\beta' \geq 0, f_y \leq 0$, and $g_\varphi \leq 0$, we can apply the standard elliptic estimation to the system (5.19) and get

$$\|p_\varepsilon\|_{H_0^1(\Omega)} + \|q_\varepsilon\|_{H_0^1(\Omega)} \leq C. \tag{5.21}$$

Then, from (5.19), we further get

$$\|\beta'(y_\varepsilon - \varphi_\varepsilon)p_\varepsilon\|_{H^{-1}(\Omega)} \leq C\varepsilon. \tag{5.22}$$

Moreover, let $S_\delta(r) \in C^1(\mathbb{R})$ be a family of smooth approximation to $sign\ r$, satisfying the following:

$$S'_\delta(r) \geq 0 \quad \forall r \in \mathbb{R} \tag{5.23}$$

and

$$S_\delta(r) = \begin{cases} 1 & \text{if} \quad r > \delta, \\ 0 & \text{if} \quad r = 0, \\ -1 & \text{if} \quad r < -\delta. \end{cases} \tag{5.24}$$

Multiplying the first equation of (5.19) by $\varepsilon S_\delta(p_\varepsilon)$ and integrating it over Ω, we can get

$$\int_\Omega \beta'(y_\varepsilon - \varphi_\varepsilon)p_\varepsilon S_\delta(p_\varepsilon)dx \leq C\varepsilon. \tag{5.25}$$

Letting $\delta \to 0$, we have

$$\|\beta'(y_\varepsilon - \varphi_\varepsilon)p_\varepsilon\|_{L^1(\Omega)} \leq C\varepsilon. \tag{5.26}$$

In estimates (5.21), (5.22), and (5.26), the constant C is independent of $\varepsilon > 0$. Hence we may let, extracting some subsequence if necessary,

$$\begin{cases} p_\varepsilon \to \overline{p} & \text{weakly in } H_0^1(\Omega), \text{ strongly in } L^2(\Omega), \\ q_\varepsilon \to \overline{q} & \text{weakly in } H_0^1(\Omega), \text{ strongly in } L^2(\Omega), \\ \frac{1}{\varepsilon}\beta'(y_\varepsilon - \varphi_\varepsilon)p_\varepsilon \to \overline{\mu} & \text{weakly star in } H^{-1}(\Omega) \cap \mathcal{M}(\overline{\Omega}). \end{cases}$$

Let $(y^\varepsilon, \varphi_\varepsilon, u_\varepsilon) \in \mathcal{A}$. By (5.9), (5.11), and Proposition 2.5, we have, for any $p \geq 2$,

$$\begin{cases} \|y^\varepsilon - \overline{y}\|_{W^{1,p}(\Omega)} \to 0, \\ \|\varphi_\varepsilon - \overline{\varphi}\|_{W^{1,p}(\Omega)} \to 0 \end{cases} \tag{5.27}$$

and

$$\|y_\varepsilon - \overline{y}\|_{W^{1,p}(\Omega)} \to 0, \tag{5.28}$$

where the convergence $\|y_\varepsilon - y^\varepsilon\|_{W^{1,p}(\Omega)} \to 0$ (see (4.27)) has been utilized.

For any $\eta \in C(\overline{\Omega})$ with $\operatorname{supp}\eta \subset \Omega_0$, the convergence (5.27)–(5.28), combined with the compactness of $\operatorname{supp}\eta$, ensures that, for some $\varepsilon_0 > 0$,

$$y_\varepsilon(x) - \varphi_\varepsilon(x) > 0 \quad \forall x \in supp\,\eta,\ 0 < \varepsilon < \varepsilon_0,$$

which yields

$$\begin{aligned}\langle \overline{\mu},\ \eta\rangle_{\mathcal{M}(\overline{\Omega}),C(\overline{\Omega})} &= \lim_{\varepsilon\to 0}\int_\Omega \frac{1}{\varepsilon}\beta'(y_\varepsilon-\varphi_\varepsilon)p_\varepsilon\eta dx\\ &= \lim_{\varepsilon\to 0}\int_{\operatorname{supp}\eta} \frac{1}{\varepsilon}\beta'(y_\varepsilon-\varphi_\varepsilon)p_\varepsilon\eta dx\\ &= 0.\end{aligned}$$

Thus, (5.4) holds.

Passing to the limit in (5.19) and (5.20), we obtain (5.5) and

$$\int_\Omega [H(x,\overline{y}(x),\overline{\varphi}(x),v(x),\overline{p}(x),\overline{q}(x)) - H(x,\overline{y}(x),\overline{\varphi}(x),\overline{u}(x),\overline{p}(x),\overline{q}(x))]dx \geq 0.$$

Finally, by the separability of U and the continuity of the Hamiltonian H in the variable v, noting also that $v\in\mathcal{U}$ is arbitrary, we obtain the optimality condition (5.6). □

6. Convex case. In the previous sections, the control domain U is just a metric space and does not necessarily have any algebraic structure. Thus we do not talk about the convexity of U, and accordingly, we use spike perturbations in the derivation of optimality conditions.

This section will be concerned with the case in which $\mathcal{U}=L^2(\Omega)$, and then the convex variation can be performed, and a stronger regularity of optimal control can be obtained. □

6.1. Problem setting. Let all the assumptions in (H_1) and (H_2) be unchanged except for U and $\mathcal{U}$. In addition, we assume that

($\tilde{H}_3$) The function $f\colon\Omega\times\mathbb{R}\to\mathbb{R}$, has the following properties: $f(\cdot,y)$ is measurable on Ω, $f(x,\cdot)$ is in $C^1(\mathbb{R})$, and there exists a constant $K>0$, such that

$$-K\leq f_y\leq 0 \quad \text{on} \quad \Omega\times\mathbb{R}, \tag{6.1}$$

$$|f(x,0)|\leq K \quad \text{on} \quad \Omega. \tag{6.2}$$

Given $y_d\in L^2(\Omega)$, we introduce the following quadratic cost functional which we try to minimize on $u\in L^2(\Omega)$:

$$J(u)=\frac{1}{2}\int_\Omega [(y-y_d)^2+u^2]dx. \tag{6.3}$$

In (6.3), the state y is the solution of the following semilinear elliptic obstacle problem:

$$\begin{cases} y\in K(\varphi),\\ a(y,z-y)\geq \displaystyle\int_\Omega f(x,y)(z-y)dx \quad \forall z\in K(\varphi),\end{cases} \tag{6.4}$$

where $K(\varphi)$ and $a(\cdot,\cdot)$ are defined as before (see (2.9) and (2.7)) and the obstacle φ solves (1.2) with $g(x,\varphi,u)=u$, i.e.,

$$\begin{cases} A\varphi = u & \text{in} \quad \Omega,\\ \varphi|_{\partial\Omega}=0.\end{cases} \tag{6.5}$$

Obviously, the cost functional (6.3) is well defined since, under (H_1), (H_2), and ($\tilde{H}_3$), (6.4) and (6.5) are uniquely solvable for any $u \in L^2(\Omega)$.

The feasible set, denoted by $\tilde{\mathcal{A}}$, consists of all triples $(y, \varphi, u) \in H_0^1(\Omega) \times H_0^1(\Omega) \times L^2(\Omega)$ satisfying (6.4) and (6.5). Any triple $(y, \varphi, u) \in \tilde{\mathcal{A}}$ is called a feasible triple, and the corresponding (y, φ) and u are referred to as a feasible state and control, respectively. Clearly, $L^2(\Omega)$ coincides with the set of all feasible controls.

Now, we pose the following optimal control problem.

Problem ($\tilde{C}$). Find a $\overline{u} \in L^2(\Omega)$, such that

$$J(\overline{u}) = \inf_{u \in L^2(\Omega)} J(u). \tag{6.6}$$

Such a $\overline{u} \in L^2(\Omega)$, if it exists, is called an optimal control and the corresponding triple $(\overline{y}, \overline{\varphi}, \overline{u}) \in \tilde{\mathcal{A}}$ is called an optimal triple.

6.2. Compactness and existence.

PROPOSITION 6.1.

(i) *There is a unique triple* $(y, \varphi, u) \in \tilde{\mathcal{A}}$ *corresponding to each* $u \in L^2(\Omega)$, *and there exists a constant* $C > 0$, *independent of* u, *such that*

$$\|\varphi\|_{H^2(\Omega)} \leq C\|u\|_{L^2(\Omega)}, \tag{6.7}$$

$$\|y\|_{H_0^1(\Omega)} \leq C(1 + \|u\|_{L^2(\Omega)}). \tag{6.8}$$

(ii) *The solution operator* $S: u \mapsto (y, \varphi)$ *of* (6.4)–(6.5) *is compact.*

Proof. The proof of (i) is classic. For the proof of (ii), it suffices to show that if $u_k \to u$ weakly in $L^2(\Omega)$, then

$$\begin{cases} \varphi_k \to \varphi & \text{strongly in } H_0^1(\Omega), \\ y_k \to y & \text{strongly in } H_0^1(\Omega), \end{cases} \tag{6.9}$$

where (y, φ, u), $(y_k, \varphi_k, u_k) \in \tilde{\mathcal{A}}\,(k = 1, 2, \ldots)$.

In fact, by estimates (6.7) and (6.8) and the uniqueness of limit point, we have the convergence (cf. the proof of Proposition 2.5):

$$\varphi_k \to \varphi \quad \text{strongly in } H_0^1(\Omega), \tag{6.10}$$

$$y_k \to y \quad \text{weakly in } H_0^1(\Omega), \text{strongly in } L^2(\Omega). \tag{6.11}$$

To prove the strong convergency of y_k in $H_0^1(\Omega)$, we note that $z_k = y \vee \varphi_k \in K(\varphi_k)$, and

$$z_k \to y \quad \text{strongly in } H_0^1(\Omega). \tag{6.12}$$

Hence, by (6.4), convergence (6.11), and (6.12),

$$a(z_k - y_k, z_k - y_k) \leq a(z_k, z_k - y_k) - \int_\Omega f(x, y_k)(z_k - y_k)dx \to 0. \tag{6.13}$$

This implies $\|z_k - y_k\|_{H_0^1(\Omega)} \to 0$, and consequently,

$$\|y_k - y\|_{H_0^1(\Omega)} \to 0. \qquad \square \tag{6.14}$$

The existence of Problem ($\tilde{C}$) is readily obtained from the compactness of the solution operator and the lower semicontinuity of the cost functional $J(u)$.

THEOREM 6.2. *Let* (H_1), (H_2), *and* ($\tilde{H}_3$) *hold. Then Problem* ($\tilde{C}$) *admits an optimal control* $\overline{u} \in L^2(\Omega)$.

6.3. Approximate control problems. Let $(\overline{y}, \overline{\varphi}, \overline{u})$ be a fixed optimal triple of Problem $(\tilde{\mathrm{C}})$.

In this subsection, we will introduce a family of approximate control problems. First we define

$$J_\varepsilon(u) = \frac{1}{2}\int_\Omega [(y_\varepsilon - y_d)^2 + u^2 + (u - \overline{u})^2]dx, \tag{6.15}$$

where y_ε is the approximate state solving

$$\begin{cases} Ay_\varepsilon + \frac{1}{\varepsilon}\beta(y_\varepsilon - \varphi) = f(x, y_\varepsilon) & \text{in} \quad \Omega, \\ y_\varepsilon|_{\partial\Omega} = 0, \end{cases} \tag{6.16}_\varepsilon$$

with φ being the solution of (6.5) and $\beta(\cdot)$ being given in (2.13).

The set of all triples $(y_\varepsilon, \varphi, u) \in H_0^1(\Omega) \times H_0^1(\Omega) \times L^2(\Omega)$ satisfying $(6.16)_\varepsilon$ and (6.5) will be denoted by $\tilde{\mathcal{A}}_\varepsilon$.

Our approximate control problems can be stated as follows.

Problem $(\tilde{\mathrm{C}})_\varepsilon$. Find a $u_\varepsilon \in L^2(\Omega)$, such that

$$J_\varepsilon(u_\varepsilon) = \inf_{u \in L^2(\Omega)} J_\varepsilon(u). \tag{6.17}$$

It is easy to see the following.

PROPOSITION 6.3. *There exists an optimal triple* $(y_\varepsilon, \varphi_\varepsilon, u_\varepsilon) \in \tilde{\mathcal{A}}_\varepsilon$ *to Problem* $(\tilde{\mathrm{C}})_\varepsilon$.

PROPOSITION 6.4. *For any fixed* $\varepsilon > 0$, *the solution mapping* $T_\varepsilon : u \mapsto (y_\varepsilon, \varphi)$ *of* $(6.16)_\varepsilon$ *and* (6.5) *is differentiable in the following sense: Given* $u, h \in L^2(\Omega)$, *there exists a pair* $(\xi, \eta) \in H_0^1(\Omega) \times H_0^1(\Omega)$, *such that, as* $\delta \to 0$,

$$\frac{1}{\delta}[T_\varepsilon(u + \delta h) - T_\varepsilon(u)] \to (\xi, \eta) \quad \textit{weakly in } H_0^1(\Omega) \times H_0^1(\Omega).$$

Furthermore, (ξ, η) *solves*

$$\begin{cases} A\xi - f_y(x, y_\varepsilon)\xi + \frac{1}{\varepsilon}\beta'(y_\varepsilon - \varphi)(\xi - \eta) = 0 & \textit{in } \Omega, \\ A\eta = h & \textit{in } \Omega, \\ \xi|_{\partial\Omega} = 0, \quad \eta|_{\partial\Omega} = 0. \end{cases} \tag{6.18}$$

From above proposition, we can derive the characterization of optimal triples of the approximate Problem $(\tilde{\mathrm{C}})_\varepsilon$.

PROPOSITION 6.5. *Let* $(y_\varepsilon, \varphi_\varepsilon, u_\varepsilon)$ *be an optimal triple of Problem* $(\tilde{\mathrm{C}})_\varepsilon$. *Then there exists an adjoint pair* $(p_\varepsilon, q_\varepsilon) \in H_0^1(\Omega) \times H_0^1(\Omega)$, *such that*

$$\begin{cases} Ap_\varepsilon + [\frac{1}{\varepsilon}\beta'(y_\varepsilon - \varphi_\varepsilon) - f_y(x, y_\varepsilon)]p_\varepsilon = y_\varepsilon - y_d & \textit{in } \Omega, \\ Aq_\varepsilon - \frac{1}{\varepsilon}\beta'(y_\varepsilon - \varphi_\varepsilon)p_\varepsilon = 0 & \textit{in } \Omega, \\ p_\varepsilon|_{\partial\Omega} = 0, \quad q_\varepsilon|_{\partial\Omega} = 0 \end{cases} \tag{6.19}$$

and

$$q_\varepsilon + 2u_\varepsilon - \overline{u} = 0 \quad \textit{a.e. in} \quad \Omega. \tag{6.20}$$

6.4. Necessary condition. In deriving the necessary conditions of the optimal control for the original Problem ($\tilde{\mathrm{C}}$), the following lemmas are crucial.

LEMMA 6.6. *Let* $(y_\varepsilon, \varphi_\varepsilon, u_\varepsilon) \in \tilde{\mathcal{A}}_\varepsilon$. *If* $u_\varepsilon \to u$ *weakly in* $L^2(\Omega)$, *for some* $u \in L^2(\Omega)$ *(in particular, this is the case if* $u_\varepsilon = u$*), then*

$$\begin{cases} y_\varepsilon \to y & \text{strongly in } H_0^1(\Omega), \\ \varphi_\varepsilon \to \varphi & \text{strongly in } H_0^1(\Omega) \end{cases} \tag{6.21}$$

with $(y, \varphi, u) \in \tilde{\mathcal{A}}$.

Proof. The proof is quite similar to that of Proposition 6.1 (ii) with some minor modifications. □

LEMMA 6.7. *Let* $(\overline{y}, \overline{\varphi}, \overline{u})$ *be an optimal triple of the original Problem* ($\tilde{\mathrm{C}}$) (*given at the beginning of subsection* 6.3), *and* $(y_\varepsilon, \varphi_\varepsilon, u_\varepsilon)$ *be optimal triples for (approximate) Problem* $(\tilde{\mathrm{C}})_\varepsilon$. *Then*

$$u_\varepsilon \to \overline{u} \quad \text{weakly in} \quad L^2(\Omega), \tag{6.22}$$

$$y_\varepsilon \to \overline{y} \quad \text{strongly in} \quad H_0^1(\Omega), \tag{6.23}$$

and

$$\varphi_\varepsilon \to \overline{\varphi} \quad \text{strongly in} \quad H_0^1(\Omega). \tag{6.24}$$

Proof. First, we note that

$$\|u_\varepsilon\|^2_{L^2(\Omega)} \le 2J_\varepsilon(u_\varepsilon) \le 2J_\varepsilon(\overline{u}) \to 2J(\overline{u}) \qquad (\varepsilon \to 0). \tag{6.25}$$

Thus, $\{u_\varepsilon\}$ is bounded in $L^2(\Omega)$. For some subsequence, still denoted by $\{u_\varepsilon\}$, we have

$$u_\varepsilon \to u \qquad \text{weakly in } L^2(\Omega) \text{ for some } u \in L^2(\Omega). \tag{6.26}$$

Then, by Lemma 6.6,

$$y_\varepsilon \to y \qquad \text{strongly in} \quad H_0^1(\Omega), \tag{6.27}$$

$$\varphi_\varepsilon \to \varphi \qquad \text{strongly in} \quad H_0^1(\Omega) \tag{6.28}$$

with $(y, \varphi, u) \in \tilde{\mathcal{A}}$.

Next, we have the following:

$$\begin{aligned} & J(\overline{u}) \le J(u) \\ &= \tfrac{1}{2} \int_\Omega [(y - y_d)^2 + u^2] dx \\ &\le \tfrac{1}{2} \int_\Omega [(y - y_d)^2 + u^2 + (u - \overline{u})^2] dx \\ &\le \underline{\lim}_{\varepsilon \to 0} \tfrac{1}{2} \int_\Omega [(y_\varepsilon - y_d)^2 + u_\varepsilon^2 + (u_\varepsilon - \overline{u})^2] dx \\ &= \underline{\lim}_{\varepsilon \to 0} J_\varepsilon(u_\varepsilon) \\ &\le \underline{\lim}_{\varepsilon \to 0} J_\varepsilon(\overline{u}) = J(\overline{u}). \end{aligned} \tag{6.29}$$

Thus, all the equalities in (6.29) must hold. This means $J(u) = J(\overline{u})$, and $\int_\Omega (u - \overline{u})^2 dx = 0$. Hence we obtain $u = \overline{u}$ and, by uniqueness, $y = \overline{y}$ and $\varphi = \overline{\varphi}$. This completes the proof. □

Using an argument analogous to section 5, we may prove the following theorem.

THEOREM 6.8. *Let* $(\overline{y}, \overline{\varphi}, \overline{u})$ *be an optimal triple of Problem* $(\tilde{\mathrm{C}})$. *Then there exists an adjoint pair* $(\overline{p}, \overline{q}) \in H_0^1(\Omega) \times H_0^1(\Omega)$ *and* $\overline{\mu} \in H^{-1}(\Omega) \cap \mathcal{M}(\overline{\Omega})$, *such that*

$$\begin{cases} A\overline{p} - f_y(x, \overline{y})\overline{p} = \overline{y} - y_d - \overline{\mu} & \text{in } \Omega, \\ A\overline{q} = \overline{\mu} & \text{in } \Omega, \\ \overline{p}|_{\partial\Omega} = 0, \quad \overline{q}|_{\partial\Omega} = 0 \end{cases} \tag{6.30}$$

and

$$\overline{q} + \overline{u} = 0 \quad \text{a.e. in } \Omega. \tag{6.31}$$

Remark. From (6. 31), we see that any optimal control $\overline{u}$ must be an element of $H_0^1(\Omega)$. Moreover, $\overline{u}$ is the strong limit point of the corresponding approximate optimal controls u_ε in $L^2(\Omega)$. More precisely, we can deduce from (6.20) and (6.31) that

$$u_\varepsilon \to \overline{u} \qquad \text{weakly in } H_0^1(\Omega) \quad \text{and strongly in } L^2(\Omega).$$

Acknowledgments. The author is grateful to Prof. Xunjing Li and Prof. Jiongmin Yong for their instructive suggestions and also to Prof. Lishang Jiang for his helpful advice.

REFERENCES

[1] D.R. ADAMS, S.M. LENHART, AND J. YONG, *Optimal control of the obstacle for an elliptic variational inequality*, Appl. Math. Optim., 38 (1998), pp. 121–140.
[2] V. BARBU, *Optimal Control of Variational Inequalities*, Pitman, London, 1984.
[3] V. BARBU, *Analysis and Control of Nonlinear Infinite Dimensional Systems*, Academic Press, New York, 1993.
[4] J.F. BONNANS AND D. TIBA, *Pontryagin's principle in the control of semilinear elliptic variational inequalities*, Appl. Math. Optim., 23 (1991), pp. 299–312.
[5] L. CESARI, *Optimization Theory and Applications, Problems with Ordinary Differential Equations*, Springer-Verlag, New York, 1983.
[6] I. EKELAND, *Nonconvex minimization problems*, Bull. Amer. Math. Soc. (NS), 1 (1979), pp. 443–474.
[7] A. FRIEDMAN, *Variational Principles and Free-Boundary Problems*, John Wiley, New York, 1982.
[8] A. FRIEDMAN, *Optimal control for variational inequalities*, SIAM J. Control Optim., 24 (1986), pp. 439–451.
[9] A. FRIEDMAN, *Optimal control for parabolic variational inequalities*, SIAM J. Control Optim., 25 (1987), pp. 482–497.
[10] D. GILBARG AND N.S. TRUDINGER, *Elliptic Partial Differential Equations of Second Order*, 2nd ed., Springer-Verlag, Berlin, 1983.
[11] J.C. HULL, *Options, Futures and Other Derivative Securities*, 2nd ed., Prentice-Hall, Englewood Cliffs, NJ, 1993.
[12] D. KINDERLEHRER AND G.STAMPACCHIA, *An Introduction to Variational Inequalities and Their Applications*, Academic Press, New York, 1980.
[13] X. LI AND J. YONG, *Optimal Control Theory for Infinite Dimensional Systems*, Birkhäuser, Boston, 1995.
[14] J.L. LIONS, *Quelques Méthodes de Résolution des Problemes aux Limites Non Linéaires*, Dunod, Paris, 1969.
[15] F. MIGNOT AND J.P. PUEL, *Optimal control in some variational inequalities*, SIAM J. Control Optim., 22 (1984), pp. 466–476.

[16] J.F. Rodrigues, *Obstacle Problems in Mathematical Physics*, North-Holland Math. Stud. 134, North-Holland, Amsterdam, 1987.
[17] D. Tiba, *Optimal Control of Nonsmooth Distributed Parameter Systems*, Lecture Notes in Math., 1495, Springer-Verlag, Berlin, 1990.
[18] G.M. Troianiello, *Elliptic Differential Equations and Obstacle Problems*, Plenum Press, New York, 1987.
[19] J. Yong, *Pontryagin maximum principle for semilinear second order elliptic partial differential equations and variational inequalities with state constraints*, Differential Integral Equations, 5 (1992), pp. 1307–1334.
[20] G. Zhang and L. Jiang, *The free boundary problem of the stationary water cone*, J. Peking Univ., 1 (1978), pp. 1–25.

SIAM J. CONTROL OPTIM.
Vol. 38, No. 1, pp. 159–174

SOME RESULTS ON THE BELLMAN EQUATION OF ERGODIC CONTROL*

HIROAKI MORIMOTO† AND MASAMI OKADA‡

Abstract. We study the multidimensional Bellman equation of ergodic control for diffusion processes without invariant measures and establish the existence of a unique solution in a certain class.

Key words. ergodic control, Bellman equation, polynomial growth, convex

AMS subject classifications. 49C20, 93E20

PII. S0363012997325411

1. Introduction.

We deal with the d-dimensional Bellman equation of the form

$$\lambda = \frac{1}{2}\Delta\phi(x) + F(D\phi(x)) + h(x), \qquad x \in \mathbf{R^d}, \tag{1}$$

where

$$\begin{aligned} F(\xi) &= \min\{|p|^2 + (\xi, p) : |p| \le 1\} \\ &= \begin{cases} -|\xi|^2/4 & \text{if } |\xi| \le 2, \\ -|\xi| + 1 & \text{if } |\xi| > 2, \end{cases} \end{aligned} \tag{2}$$

and $|\cdot|, (,)$, and D denote the norm, the inner product of vectors, and the gradient, respectively. Now we are given a convex function $h(x)$ with polynomial growth, and the unknown is the pair of a constant λ and a C^2-function $\phi(x)$ on $\mathbf{R^d}$.

Generally speaking, the linear ergodic control problem has been investigated for the stochastic differential systems with invariant measures [3], [5]. The corresponding Bellman equation is given by

$$\lambda = \frac{1}{2}\Delta\phi + (Ax, D\phi) + F(D\phi) + h,$$

and the matrix A is assumed to satisfy a kind of Lyapunov-type stability condition for the existence result of a unique solution.

Equation (1) is related to the study of ergodic control in the linear systems without invariant measures such as manufacturing systems in the production planning problems [14]. Also, this is an extension of [9] by means of direct calculations, and [10] recently treated this from the point of view of reducing (1) to an equivalent integral equation in the 1-dimensional case.

Our aim is to show that the Bellman equation (1) admits a unique solution $(\lambda, \phi) \in \mathbf{R} \times C^2(\mathbf{R^d})$ in a certain class. To solve (1), we consider the Bellman equation for the

*Received by the editors August 4, 1997; accepted for publication (in revised form) November 25, 1998; published electronically December 7, 1999.
http://www.siam.org/journals/sicon/38-1/32541.html
†Department of Mathematical Sciences, Faculty of Science, Ehime University, Matsuyama 790, Japan (morimoto@sci.sci.ehime-u.ac.jp).
‡Division of Mathematical Sciences, G S I S, Tohoku University, Sendai 980-77, Japan.

control problem with discounted rate $\alpha > 0$:

$$\alpha u_\alpha(x) = \frac{1}{2}\Delta u_\alpha(x) + F(Du_\alpha(x)) + h(x), \qquad x \in \mathbf{R^d}. \tag{3}$$

We apply the technique of [4] to (1) with the Lipschitz continuous term $F(\xi)$, and we are focused on studying the limit of (3) as $\alpha \to 0$. The solution (λ, ϕ) is given by the limit

$$\alpha \min u_\alpha \quad \to \quad \lambda, \tag{4}$$

$$u_\alpha(x) - \min u_\alpha \quad \to \quad \phi(x). \tag{5}$$

It is known in [12] that (3) has a unique solution u_α. However, the crucial point for establishing (4), (5), and (1) is that u_α becomes a convex function with polynomial growth and we can obtain the gradient estimates of u_α in the whole space $\mathbf{R^d}$ uniformly in α.

Let us mention the works [4], [7] on the Bellman equation of ergodic control without convex and polynomial growth hypotheses and the linear quadratic case, in which $F(\xi)$ is replaced by $\min\{|p|^2 + (\xi, p) : p \in \mathbf{R^d}\}$. We refer to [13] for ergodic control of reflected diffusions.

The content of this paper is as follows. In section 2 we show the existence of a unique solution u_α to Bellman equation (3) under the convexity assumption on $h(x)$. Further, in section 3, we study the limit of u_α as $|x| \to \infty$ and the polynomial growth property. Section 4 is devoted to a priori estimates of $u_\alpha - \min u_\alpha$ for the approximation problem. Section 5 deals with the existence of the unique solution of (1). Finally, in section 6, we present an application of our results to an ergodic control problem of linear stochastic differential systems without invariant measures.

2. Bellman equations of discounted cost control. We study the existence of a unique solution u_α with polynomial growth to the Bellman equation:

$$\alpha u_\alpha(x) = \frac{1}{2}\Delta u_\alpha(x) + F(Du_\alpha(x)) + h(x), \qquad x \in \mathbf{R^d}, \tag{6}$$

where $0 < \alpha < 1/2$. We assume

$$h : \text{nonnegative, convex on } \mathbf{R^d}, \tag{7}$$

h satisfies the polynomial growth condition, i.e.,

$$\exists C > 0, \qquad m \in \mathbf{N}_+; \quad h(x) \le C(1 + |x|^m), \quad x \in \mathbf{R^d}, \tag{8}$$

$$h \in C^1(\mathbf{R^d}). \tag{9}$$

To simplify the notation, we make use of the following quantity:

$$[f]_{\delta, B_r} = \sup_{x \in B_r} |f(x)| + \sup_{x, y \in B_r,\ x \neq y} \frac{|f(x) - f(y)|}{|x - y|^\delta},$$

where $f(x)$ is the bounded Hölder continuous function with exponent δ on a ball $B_r = B_r(0)$ of $\mathbf{R^d}$. Let $t_n \in C_c^\infty(\mathbf{R^d})$ be a sequence such that $t_n = 1$ on B_n, 0 outside B_{2n} and $0 \le t_n \le 1, |Dt_n| \le C/n$, and set $h_n = t_n h \in C_c^1(\mathbf{R^d})$. It is clear that $h_n \to h$ and $0 \le h_n \le h$.

Now let us consider the Bellman equation

$$\alpha u_n(x) = \frac{1}{2}\Delta u_n(x) + F(Du_n(x)) + h_n(x), \qquad x \in \mathbf{R^d}. \tag{10}$$

Then we have Theorem 2.1.

THEOREM 2.1. *Under* (7), (8), *and* (9), *equation* (10) *admits a unique solution* $u_n \in C_0(\mathbf{R^d}) \cap C^2(\mathbf{R^d})$, *which satisfies*

$$\sup_n [u_n]_{\delta,B_r} < \infty, \tag{11}$$

$$\sup_n \sum_i [D_i u_n]_{\delta,B_r} < \infty, \tag{12}$$

$$\sup_n \sum_{i,j} [D_{ij} u_n]_{\delta,B_r} < \infty \tag{13}$$

for some $0 < \delta < 1$.

Proof. It is well known [2] that, for every $n \in \mathbf{N}_+$, (10) has a unique solution u_n of the form

$$u_n(x) = \inf \left\{ E\left[\int_0^\infty e^{-\alpha t}(h_n(x(t)) + |p(t)|^2)dt \right] : |p(t)| \le 1 \right\} \tag{14}$$

in the class $C_0(\mathbf{R^d})$ of continuous functions vanishing at infinity, where $x(t)$ is a solution of the stochastic differential equation

$$dx(t) = p(t)dt + dw(t), \qquad x(0) = x \in \mathbf{R^d},$$

defined on some probability space $(\Omega, \mathcal{F}, P; \{\mathcal{F}_t\})$ carrying a d-dimensional standard Brownian motion $w(t)$, and the infimum is taken over the class of all $\mathcal{F}_t$-progressively measurable processes $p(t)$ with $|p(t)| \le 1$.

Since the similar proof of the assertions is needed later, we shall divide it into several steps.

Step 1. We note $u_n \ge 0$ by (14) and (7). Let $\nu \ge 1$ and we shall show that

$$2\alpha \int u_n^{\nu+1}\Theta dx + \int |Du_n|^2 \nu u_n^{\nu-1}\Theta dx \le 2\int (h_n + 1)u_n^\nu \Theta dx, \tag{15}$$

where $\Theta(x) = e^{-\theta|x|}$ for a positive constant θ chosen later. Multiplying both sides of (10) by $u_n^\nu\Theta$ and integrating over $\mathbf{R^d}$, we have

$$2\alpha \int u_n^{\nu+1}\Theta dx - \int \Delta u_n u_n^\nu \Theta dx - 2\int F(Du_n)u_n^\nu\Theta dx = 2\int h_n u_n^\nu \Theta dx.$$

The second term of the left-hand side can be rewritten as

$$\begin{aligned}\int (Du_n, D(u_n^\nu\Theta))dx &= \int \left[|Du_n|^2 \nu u_n^{\nu-1}\Theta - \theta u_n^\nu \left(Du_n, \frac{x}{|x|} \right)\Theta \right] dx \\ &\ge \int [|Du_n|^2 \nu u_n^{\nu-1} - \theta u_n^\nu |Du_n|]\Theta dx.\end{aligned}$$

Since $|\xi| + 1 \ge -F(\xi) \ge |\xi| - 1$, we get (15) for a choice of θ with $0 < \theta < 2$.

Step 2. By Step 1, we have

$$\begin{aligned}\alpha \int u_n^{\nu+1}\Theta dx &\le \int (h_n + 1)u_n^\nu \Theta dx \\ &\le \int (|h| + 1)u_n^\nu \Theta dx \\ &\le \left(\int (|h| + 1)^{\nu+1}\Theta dx \right)^{1/(\nu+1)} \left(\int (u_n^\nu)^{(\nu+1)/\nu}\Theta dx \right)^{\nu/(\nu+1)}\end{aligned}$$

from which

$$\alpha\left(\int u_n^{\nu+1}\Theta dx\right)^{1/(\nu+1)} \le \left(\int (|h|+1)^{\nu+1}\Theta dx\right)^{1/(\nu+1)}.$$

Step 3. Here we take $\nu = 1$ in (15). Then

$$\begin{aligned}\alpha\int (u_n^2+|Du_n|^2)\Theta dx &\le \int (h_n+1)u_n\Theta dx\\ &\le \left(\int (|h|+1)^2\Theta dx\right)^{1/2}\left(\int u_n^2\Theta dx\right)^{1/2}\\ &\le \left(\int (|h|+1)^2\Theta dx\right)^{1/2}\left(\int (u_n^2+|Du_n|^2)\Theta dx\right)^{1/2}.\end{aligned}$$

Hence

$$\sup_n \int (u_n^2+|Du_n|^2)\Theta dx < \infty,$$

and thus,

$$\sup_n(|u_n|_{L^2(B_r)}+|Du_n|_{L^2(B_r)}) < \infty \quad \text{for each } r>0.$$

By the regularity result [11, Thm 8.8, p. 183], there exists $C>0$ such that

$$|u_n|_{W^{2,2}(B_r)} \le C(|u_n|_{W^{1,2}(B_{r+1})}+|\Delta u_n|_{L^2(B_{r+1})}).$$

Therefore, we get by (10)

$$\sup_n |u_n|_{W^{2,2}(B_r)} < \infty.$$

Step 4. We show that

$$\sup_n |u_n|_{W^{2,k}(B_r)} < \infty \qquad \text{for all } k > d. \tag{16}$$

Due to the Sobolev inequality [6, Thm IX.16, p. 169], (16) follows immediately in the case of $d = 1, 2$. Further, assuming $d > 2$, we have

$$\sup_n |Du_n|_{L^q(B_{r+1})} \le \sup_n C|Du_n|_{W^{1,2}(B_{r+1})} < \infty, \qquad \frac{1}{q} = \frac{1}{2}-\frac{1}{d}.$$

By Step 2 and (10)

$$\sup_n |\Delta u_n|_{L^q(B_{r+1})} < \infty.$$

Moreover, we know by [1] that

$$|u_n|_{W^{2,q}(B_r)} \le C(|u_n|_{W^{1,q}(B_{r+1})}+|\Delta u_n|_{L^q(B_{r+1})}).$$

Hence,

$$\sup_n |u_n|_{W^{2,q}(B_r)} < \infty.$$

By a bootstrap argument, we can obtain (16).

Step 5. Using the Sobolev inequality again, we have

$$\sup_n |u_n|_{L^\infty(B_r)} \le \sup_n C|u_n|_{W^{1,k}(B_r)} < \infty.$$

Now, we apply the Morrey theorem [6, Thm IX.12, p. 166 or p. 169],

$$|u_n(x) - u_n(y)| \le C|u_n|_{W^{1,k}(B_r)}|x-y|^\delta \qquad \forall x, y \in B_r, \qquad \delta = 1 - d/k,$$

to obtain (11). Similarly, by Step 4, we deduce (12).

Step 6. We can easily see by (9) that the derivative $D_i u_n$ satisfies

$$\alpha D_i u_n = \frac{1}{2}\Delta D_i u_n + (DF(Du_n), DD_i u_n) + D_i h_n.$$

Since $DF(\xi)$ is bounded, we have

$$\sup_n |\Delta D_i u_n|_{L^k(B_r)} < \infty,$$

and by virtue of [1]

$$\sup_n |Du_n|_{W^{2,k}(B_r)} < \infty.$$

Thus, by the same argument as Step 5, we deduce (13). The proof is complete. □

To show the existence of u_α of (6) with polynomial growth, we consider the stochastic differential equation

$$dz(t) = G(z(t))dt + dw(t), \qquad z(0) = x \in \mathbf{R^d}, \tag{17}$$

where

$$G(z) = \begin{cases} -z/|z| \text{ if } z \in \mathbf{R^d}\backslash\{0\}, \\ 0 \qquad \text{if } z = 0. \end{cases}$$

LEMMA 2.2. *For each $n \in \mathbf{N}_+$, there exists $C > 0$ such that*

$$\alpha E\left[\int_0^\infty e^{-\alpha t}|z(t)|^{2n} dt\right] \le C(1 + |x|^{2n+2}). \tag{18}$$

Proof. By an application of Ito's formula to $e^{-\alpha t}|z|^{2n+2}$, we have

$$\begin{aligned} E[e^{-\alpha t}|z(t)|^{2n+2}] - |x|^{2n+2} &= E\bigg[\int_0^t e^{-\alpha s}\{-\alpha|z(s)|^{2n+2} + (2n+2)(G(z(s)), z(s))|z(s)|^{2n} \\ &\qquad + (n+1)(2n+d)|z(s)|^{2n}\}ds\bigg] \\ &\le E\bigg[\int_0^t e^{-\alpha s}\{-2(n+1)|z(s)|^{2n+1} \\ &\qquad + (n+1)(2n+d)|z(s)|^{2n}\}ds\bigg]. \end{aligned}$$

Hence

$$E\left[\int_0^\infty e^{-\alpha s}\ g(z(s))\ ds\right] \le |x|^{2n+2},$$

where $g(z) = 2(n+1)|z|^{2n+1} - (n+1)(2n+d)|z|^{2n}$. We choose $\zeta > 0$ such that $|z|^{2n} \le g(z)$ for all $|z| \ge \zeta$. Then

$$\begin{aligned} E\left[\int_0^\infty e^{-\alpha t}|z(t)|^{2n}dt\right] &= \int_0^\infty e^{-\alpha t}E[|z(t)|^{2n}1_{(|z(t)|<\zeta)} + |z(t)|^{2n}1_{(|z(t)|\ge\zeta)}]dt \\ &\le \int_0^\infty e^{-\alpha t}\{\zeta^{2n} + E[g(z(t))]\}dt \\ &\le \zeta^{2n}/\alpha + |x|^{2n+2}. \end{aligned}$$

Thus we get (18) with $C > 0$ independent of α.

THEOREM 2.3. *Assume* (7), (8), (9). *Then there exists a unique solution* $u_\alpha \in C^2(\mathbf{R^d})$ *of equation* (6) *such that* u_α *is convex and satisfies*

$$0 \le \alpha u_\alpha(x) \le C(1+|x|^{m+3}), \qquad x \in \mathbf{R^d}, \tag{19}$$

for some constant $C > 0$.

Proof. By Theorem 2.1, it is evident that the sequences $\{u_n\}, \{Du_n\}$, and $\{\Delta u_n\}$ are uniformly bounded and equicontinuous on every B_r. By the Ascoli–Arzelà theorem, we have

$$\begin{aligned} u_n &\to u_\alpha \in C^2(\mathbf{R^d}), \\ Du_n &\to Du_\alpha, \\ \Delta u_n &\to \Delta u_\alpha \quad \text{uniformly on } B_r, \end{aligned}$$

taking a subsequence if necessary. Passing to the limit in (10), we can obtain (6).

To prove (19), we recall (14). Hence, by (8) and Lemma 2.2,

$$\begin{aligned} 0 \le \alpha u_n(x) &\le \alpha E\left[\int_0^\infty e^{-\alpha t}(h_n(z(t)) + |G(z(t))|^2)dt\right] \\ &\le C\left(1 + \alpha E\left[\int_0^\infty e^{-\alpha t}(1+|z(t)|^m)dt\right]\right) \\ &\le C(1+|x|^{m+3}). \end{aligned}$$

This implies that u_α satisfies (19). Due to (19), we have

$$E[e^{-\alpha t}u_\alpha(x(t))] \to 0 \quad \text{as} \quad t \to \infty.$$

Hence, by Ito's formula

$$u_\alpha(x) = \inf\left\{E\left[\int_0^\infty e^{-\alpha t}(h(x(t)) + |p(t)|^2)dt\right] : |p(t)| \le 1\right\}, \tag{20}$$

where the infimum is attained by the feedback law $\rho(Du_\alpha)$ with $\rho(\xi) = \arg\min F(\xi)$. Thus the convexity of u_α follows (cf. [8, p. 204]). The proof is complete. □

Remark. In the case of $d = 1$, the theorem is verified without (9), because h is Lipschitz continuous and (12) implies (13).

3. Limit at infinity and polynomial growth. We consider the limit of the solution u_α to Bellman equation (6) as $|x| \to \infty$ and also the polynomial growth property of $u_\alpha - \min u_\alpha$, denoted by v_α. We make the following assumption:

$$\text{There exists } C_0 > 0 \text{ such that } \ h(x) \geq C_0|x|. \tag{21}$$

Our objective in this section is to prove the following result.

THEOREM 3.1. *We assume* (7), (8), (9), (21). *Then we have*

$$u_\alpha(x) \quad \to \quad \infty \quad \textit{as} \quad |x| \quad \to \quad \infty \quad \textit{uniformly in } \alpha, \tag{22}$$

$$v_\alpha(x) \leq C(1 + |x|^{m+1}) \tag{23}$$

for some constant $C > 0$.

LEMMA 3.2. *If* $d = 1$, *then the assertions of Theorem* 3.1 *hold.*

Proof. By (6), (2), and the convexity of $u_\alpha \geq 0$, we have

$$\begin{aligned} h(x) &\leq \alpha u_\alpha(x) + |u_\alpha'(x)| + 1 \\ &\leq \alpha(|u_\alpha'(x)||x| + u_\alpha(0)) + |u_\alpha'(x)| + 1. \end{aligned}$$

Hence, by (21)

$$\begin{aligned} |u_\alpha'(x)| &\geq (h(x) - 1 - \alpha u_\alpha(0))/(|x| + 1) \\ &\geq C_0/2 \quad \text{for sufficiently large } x, \end{aligned}$$

and then

$$u_\alpha(x) \quad \to \quad \infty \quad \text{as} \quad |x| \quad \to \infty.$$

Thus, by convexity, we can find $C > 0$ independent of α such that $u_\alpha(x) \geq C|x| - 1/C$. This implies (22).

Now we can define $\gamma_\alpha \in \mathbf{R}$ by $u_\alpha(\gamma_\alpha) = \min_x u_\alpha(x)$. Since $u_\alpha'(\gamma_\alpha) = 0$, it is easy to see that

$$h(\gamma_\alpha) \leq \alpha u_\alpha(\gamma_\alpha) \leq \alpha u_\alpha(0).$$

By (19) and (21), this yields that

$$\sup_\alpha |\gamma_\alpha| < \infty. \tag{24}$$

Moreover,

$$\sup_\alpha \alpha \min u_\alpha < \infty.$$

From (6) it follows that

$$\alpha v_\alpha = \frac{1}{2} v_\alpha'' + F(v_\alpha') + (h - \alpha \min u_\alpha),$$

and by (2),

$$0 \leq v_\alpha'' - 2|v_\alpha'| + 2(1 + h - \alpha \min u_\alpha).$$

By (19) we note that $\lim_{y\to\infty} e^{-2y}v_\alpha(y) = 0$.

Integrating over $[x, y]$ and letting $y \to \infty$, we have

$$\begin{aligned} 0 \le v_\alpha'(x) &\le \int_0^\infty 2(1 + h(s+x) - \alpha \min u_\alpha)e^{-2s}ds \\ &\le C(1 + |x|^m) \qquad \text{for} \qquad x \ge \gamma_\alpha. \end{aligned}$$

Hence

$$0 \le v_\alpha(x) \le C(1 + |x|^{m+1}) \qquad \text{for} \qquad x \ge \gamma_\alpha.$$

By the same calculation as above over $(-\infty, \gamma_\alpha]$, we deduce (23). □

Proof of Theorem 3.1. We prove the theorem, comparing (6) with

$$\alpha v_i = \frac{1}{2}\Delta v_i + F_i(Dv_i) + h_i, \quad i = 1, 2, \tag{25}$$

where

$$F_1(\xi) = \sum_{j=1}^d \min_{|p_j|\le 1}(p_j^2 + p_j\xi_j), \qquad h_1(x) = \sum_{j=1}^d \eta_1|x_j|,$$

$$F_2(\xi) = \sum_{j=1}^d \min_{|p_j|\le 1/d}(p_j^2 + p_j\xi_j), \qquad h_2(x) = \sum_{j=1}^d \{\eta_2(|x_j|^m + 1) - \alpha \min u_\alpha\},$$

and each positive constant η_i will be chosen later. Here we note that the summation of F_i can be interchanged with the minimization on p_j. By virtue of Theorem 2.3, each equation has a unique solution v_i. Further, we can easily see that v_i is of the form

$$v_i(x) = \sum_{j=1}^d w_j^{(i)}(x_j)$$

for the solution $w_j^{(i)}(x_j)$ to (25) in the case $d = 1$. Hence, by Lemma 3.2, the following relations are fulfilled:

$$\begin{aligned} &v_1(x) \quad \to \quad \infty \quad \text{as} \quad |x| \quad \to \quad \infty \ \text{ uniformly in } \alpha, \\ &v_2(x) - \min v_2 \le \sum_j (w_j^{(2)}(x_j) - \min w_j^{(2)}) \le C(1 + |x|^{m+1}) \end{aligned}$$

for sufficiently large $\eta_2 > 0$ with $C_0|x| \le h_2$. Along the same line as (20), we have

$$v_\alpha(x) = \inf\left\{E\left[\int_0^\infty e^{-\alpha t}(h(x(t)) - \alpha \min u_\alpha + |p(t)|^2)dt\right] : |p(t)| \le 1\right\},$$

$$v_1(x) = \inf\left\{E\left[\int_0^\infty e^{-\alpha t}(h_1(x(t)) + |p(t)|^2)dt\right] : |p_j(t)| \le 1\right\},$$

$$v_2(x) - \min v_2 = \inf\left\{E\left[\int_0^\infty e^{-\alpha t}(h_2(x(t)) - \alpha \min v_2 + |p(t)|^2)dt\right] : |p_j(t)| \le 1/d\right\}.$$

Since $\{p : |p_j| \le 1/d\} \subset \{p : |p| \le 1\} \subset \{p : |p_j| \le 1\}$, we can obtain

$$\begin{aligned} u_\alpha &\ge v_1, \\ v_\alpha &\le v_2 - \min v_2 \end{aligned}$$

for a convenient choice of each η_i such that $h_1 \le h$ and $h - \alpha \min u_\alpha \le h_2 - \alpha \min v_2$. Thus the theorem is established.

4. A priori estimates for approximation. For the approximation problem of (1), we consider here the gradient estimates of v_α satisfying

$$\alpha v_\alpha = \frac{1}{2}\Delta v_\alpha + F(Dv_\alpha) + (h - \alpha \min u_\alpha), \tag{26}$$

which follows from (6).

THEOREM 4.1. *Assume* (7), (8), (9), (21). *Then there exists* $0 < \delta < 1$ *such that*

$$\sup_{0<\alpha<1/2} [v_\alpha]_{\delta,B_r} < \infty, \tag{27}$$

$$\sup_{0<\alpha<1/2} \sum_i [D_i v_\alpha]_{\delta,B_r} < \infty, \tag{28}$$

$$\sup_{0<\alpha<1/2} \sum_{i,j} [D_{ij} v_\alpha]_{\delta,B_r} < \infty \tag{29}$$

for every $r > 0$.

Proof. Multiplying both sides of (26) by $v_\alpha(x)\Theta(x)$ and integrating over $\mathbf{R^d}$, we have

$$2\int \alpha v_\alpha^2 \Theta dx - \int (\Delta v_\alpha) v_\alpha \Theta dx - 2\int F(Dv_\alpha) v_\alpha \Theta dx = 2\int (h - \alpha \min u_\alpha) v_\alpha \Theta dx,$$

where Θ is as in the proof of Theorem 2.1. The second term of the left-hand side can be rewritten as

$$\begin{aligned}\int (Dv_\alpha, D(v_\alpha\Theta))dx &= \int \left[|Dv_\alpha|^2\Theta - \theta v_\alpha \left(Dv_\alpha, \frac{x}{|x|} \right) \Theta \right] dx \\ &\geq \int [|Dv_\alpha|^2 - \theta v_\alpha |Dv_\alpha|]\Theta dx.\end{aligned}$$

By the choice of θ with $2 > \theta > 0$, we get

$$\int |Dv_\alpha|^2 \Theta dx \leq 2\int (h - \alpha \min u_\alpha + 1) v_\alpha \Theta dx.$$

Since

$$v_\alpha(\bar\gamma_\alpha) - v_\alpha(x) \geq (Dv_\alpha(x), \bar\gamma_\alpha - x) \quad \text{for} \quad \bar\gamma_\alpha := \arg\min v_\alpha,$$

and $\{\bar\gamma_\alpha\}$ is bounded by analogy with (24), we have

$$0 \leq v_\alpha(x) \leq C|Dv_\alpha(x)|(|x| + 1). \tag{30}$$

Then by the Schwarz inequality

$$\int |Dv_\alpha|^2 \Theta dx \leq 2C \left(\int (h - \alpha \min u_\alpha + 1)^2 (|x| + 1)^2 \Theta dx \right)^{1/2} \left(\int |Dv_\alpha|^2 \Theta dx \right)^{1/2}.$$

Therefore, we deduce

$$\sup_\alpha \int |Dv_\alpha|^2 \Theta dx < \infty.$$

Now we can find a constant $C > 0$ independent of α such that

$$|Dv_\alpha|_{L^2(B_r)} \leq C,$$

and by (30)

$$|v_\alpha|_{L^2(B_r)} \le C.$$

Along the same line as Step 3 of Theorem 2.1, we can obtain

$$\sup_\alpha |v_\alpha|_{W^{2,2}(B_r)} < \infty.$$

Finally, by the same bootstrap argument as Step 4 of Theorem 2.1,

$$\sup_\alpha |v_\alpha|_{W^{2,k}(B_r)} < \infty \qquad \text{for all } k > d.$$

Further, repeating Steps 5 and 6 of Theorem 2.1, we can deduce (27), (28), (29). □

5. Bellman equation of ergodic control.

5.1. Existence. In this section we shall show the existence of a unique solution $(\lambda, \phi) \in \mathbf{R} \times C^2(\mathbf{R^d})$ to the Bellman equation

$$\lambda = \frac{1}{2}\Delta\phi(x) + F(D\phi(x)) + h(x), \qquad x \in \mathbf{R^d}. \tag{31}$$

THEOREM 5.1. *We assume* (7), (8), (9), (21). *Then there exists a subsequence* $\alpha \to 0$ *such that*

$$\alpha \min u_\alpha \quad \to \quad \lambda \in \mathbf{R}_+, \tag{32}$$

$$v_\alpha(x) \quad \to \quad \phi(x) \in C^2(\mathbf{R^d}) \quad \textit{uniformly on each } B_r. \tag{33}$$

The limit (λ, ϕ) *satisfies Bellman equation* (31) *and, furthermore,*

$$\phi : \textit{convex on } \mathbf{R^d}, \tag{34}$$

$$\phi(x) \quad \to \quad \infty \quad \textit{as} \quad |x| \quad \to \quad \infty, \tag{35}$$

$$0 \le \phi(x) \le C(1 + |x|^{m+1}) \tag{36}$$

for some constant $C > 0$.

Proof. From Theorem 4.1 it follows that $\{v_\alpha\}$ and $\{Dv_\alpha\}$ are uniformly bounded and equicontinuous on each B_r, and so is $\{D^2 v_\alpha\}$. By the Ascoli–Arzelà theorem, we get

$$\begin{aligned} v_\alpha \quad &\to \quad \phi, \\ Dv_\alpha \quad &\to \quad D\phi, \\ D^2 v_\alpha \quad &\to \quad D^2\phi \qquad \text{uniformly on } B_r, \end{aligned}$$

taking a subsequence if necessary. Moreover,

$$\alpha v_\alpha(x) \quad \to \quad 0, \qquad \alpha \min u_\alpha \quad \to \quad \lambda,$$

since $\{\alpha \min u_\alpha\}$ is bounded. Passing to the limit in (26), we deduce (31). The assertions (34), (35), and (36) follow from Theorems 2.3 and 3.1 immediately. □

5.2. Uniqueness. Before going into the proof of uniqueness, we need the following property of the stochastic differential equation

$$dx^*(t) = \rho(D\phi(x^*(t)))dt + dw(t), \qquad x^*(0) = x, \tag{37}$$

where $\rho(\xi) = \arg\min F(\xi)$, i.e.,

$$\rho(\xi) = \begin{cases} -\xi/2 & \text{if } |\xi| \le 2, \\ -\xi/|\xi| & \text{if } |\xi| > 2. \end{cases} \tag{38}$$

LEMMA 5.2. *For any $n \in \mathbf{N}_+$, there exists $C > 0$ such that*

$$E[|x^*(t)|^{2n}] \le C(1+t). \tag{39}$$

Proof. We remark by (31) and (21) that

$$-F(D\phi(x)) \ge h(x) - \lambda \quad \to \quad \infty,$$

and hence

$$|D\phi(x)| \quad \to \quad \infty \quad \text{as} \quad |x| \quad \to \quad \infty.$$

Now, let us show that

$$\left(x, \frac{D\phi(x)}{|D\phi(x)|}\right) \quad \to \quad \infty \quad \text{as} \quad |x| \quad \to \quad \infty. \tag{40}$$

Suppose it were not true. Then there exists a sequence $x_k \to \infty$ such that

$$\left(x_k, \frac{D\phi(x_k)}{|D\phi(x_k)|}\right) \le C \quad \text{for some } C > 0.$$

Define $S_k = \{x : \ \phi(x) < \phi(x_k)\}$, and let T_k be the tangent plane of S_k at x_k. Since 0 belongs to S_k, we can easily find $y_k \in T_k$ such that the vector y_k coincides with $|y_k|D\phi(x_k)/|D\phi(x_k)|$. By the definition of T_k,

$$(y_k - x_k, D\phi(x_k)) = 0.$$

Hence

$$|y_k| \le C.$$

On the other hand, since S_k is convex, we have

$$\phi(y_k) \ge \phi(x_k),$$

and hence

$$\phi(y_k) \quad \to \quad \infty \quad \text{as} \quad k \quad \to \quad \infty.$$

This is a contradiction, and thus we get (40).

Next, we have by Ito's formula

$$E[|x^*(t)|^{2n}] - |x|^{2n} = E\Bigg[\int_0^t \{2n|x^*(s)|^{2(n-1)}(x^*(s), \rho(D\phi(x^*(s))) + n(d + 2(n-1))|x^*(s)|^{2(n-1)}\}ds\Bigg].$$

By (40), we can choose $R > 0$ such that

$$2(x, \rho(D\phi(x))) + d + 2(n-1) \leq 0 \quad \text{for} \quad |x| \geq R.$$

Therefore, we deduce (39). The proof is complete. □

THEOREM 5.3. *We make the assumptions of Theorem* 5.1. *Let* $(\lambda_j, \phi_j) \in \mathbf{R} \times C^2(\mathbf{R^d}), j = 1, 2,$ *be two solutions of* (31) *satisfying* (34), (35), (36). *Then we have*

$$\lambda_1 = \lambda_2, \tag{41}$$

$$D\phi_1 = D\phi_2. \tag{42}$$

Proof. We first show (41). By (31) and (38) we have

$$\lambda_1 = \frac{1}{2}\Delta\phi_1 + |\rho(D\phi_1)|^2 + (\rho(D\phi_1), D\phi_1) + h,$$
$$\lambda_2 \leq \frac{1}{2}\Delta\phi_2 + |\rho(D\phi_1)|^2 + (\rho(D\phi_1), D\phi_2) + h.$$

Denoting $\hat{\phi} = \phi_1 - \phi_2$, we get

$$\frac{1}{2}\Delta\hat{\phi} + (\rho(D\phi_1), D\hat{\phi}) \leq \lambda_1 - \lambda_2. \tag{43}$$

We apply Ito's formula to the solution $\hat{x}(t)$ of (37) with ϕ_1 replacing ϕ. Then

$$E[e^{-\alpha(t\wedge\sigma_n)}\hat{\phi}(\hat{x}(t\wedge\sigma_n))] = \hat{\phi}(x) + E\Bigg[\int_0^{t\wedge\sigma_n} \Big\{ -\alpha e^{-\alpha s}\hat{\phi}(\hat{x}(s))ds + e^{-\alpha s}(\rho(D\phi_1), D\hat{\phi})(\hat{x}(s)) + e^{-\alpha s}\frac{1}{2}\Delta\hat{\phi}(\hat{x}(s))\Big\}ds\Bigg],$$

where $\{\sigma_n\}$ is a sequence of localizing times for the local martingale. Note that

$$|\hat{x}(t)| \leq |x| + t + |w(t)|. \tag{44}$$

Letting $n \to \infty$, we have by the dominated convergence theorem

$$E[e^{-\alpha t}\hat{\phi}(\hat{x}(t))] - \hat{\phi}(x) \leq E\left[\int_0^t e^{-\alpha s}\{-\alpha\hat{\phi}(\hat{x}(s)) + (\lambda_1 - \lambda_2)\}ds\right].$$

Moreover, letting $t \to \infty$, we get

$$-\hat{\phi}(x) \leq E\left[\int_0^\infty e^{-\alpha s}\{-\alpha\hat{\phi}(\hat{x}(s))\}ds\right] + (\lambda_1 - \lambda_2)/\alpha$$

or, equivalently,

$$\lambda_1 - \lambda_2 \geq \alpha\left(\alpha E\left[\int_0^\infty e^{-\alpha s}\hat{\phi}(\hat{x}(s))ds\right] - \hat{\phi}(x)\right).$$

By (36) and Lemma 5.2 we have

$$\begin{aligned} |E[\hat{\phi}(\hat{x}(t))]| &\leq C(1 + E[|\hat{x}(t)|^{m+1}]) \\ &\leq C(1 + (E[|\hat{x}(t)|^{2(m+1)}])^{1/2}) \\ &\leq C(1 + (C(t+1))^{1/2}). \end{aligned} \tag{45}$$

Hence

$$\left|\alpha^2 E\left[\int_0^\infty e^{-\alpha s}\hat{\phi}(\hat{x}(s))ds\right]\right| \le C\left(\alpha+\alpha^2\int_0^\infty e^{-\alpha s}s^{1/2}ds\right)\to 0 \quad \text{as} \quad \alpha\to 0.$$

Passing to the limit, we can deduce $\lambda_1 \ge \lambda_2$, and thus (41).

Next, in order to prove (42), it is sufficient to show that ϕ admits a representation

$$\phi(x)-\phi(0)=\lim_{r\to 0+}\inf\left\{E\left[\int_0^{\tau_r}((h(x(t))+|p(t)|^2)-\lambda)dt\right] \;:\; |p(t)|\le 1\right\}, \tag{46}$$

where $\tau_r=\inf\{t : x(t)\in B_r\}$ for the response $x(t)$ to each control $p(t)$. By (31) and (37), it is obvious that

$$\phi(x^*(t))+\int_0^t (h(x^*(s))+|p^*(s)|^2-\lambda)ds \text{ is a local martingale,}$$

where $p^*(t)=\rho(D\phi(x^*(t)))$. Taking into account (44), we see that

$$\phi(x)=E\left[\int_0^{\tau_r\wedge t}(h(x^*(s))+|p^*(s)|^2-\lambda)ds+\phi(x^*(\tau_r\wedge t))\right].$$

Letting $t\to\infty$, we have by Fatou's lemma

$$\phi(x)\ge E\left[\int_0^{\tau_r}(h(x^*(s))+|p^*(s)|^2-\lambda)ds+\phi(x^*(\tau_r))\right].$$

Here, we note that $x^*(t)$ becomes a Brownian motion by the Girsanov transformation of probability measure P. Then $P(\tau_r<\infty)=1$ and hence $P(|x^*(\tau_r)|=r)=1$. Thus

$$\phi(x)-\inf_{|x|=r}\phi(x)\ge E\left[\int_0^{\tau_r}(h(x^*(s))+|p^*(s)|^2-\lambda)ds\right] \quad \text{for } |x|>r. \tag{47}$$

Letting $r\to 0$, we get (46) with $\ge$ replacing $=$.

On the other hand, by (26), it is easy to see that

$$e^{-\alpha t}v_\alpha(x(t))+\int_0^t e^{-\alpha s}(h(x(s))+|p(t)|^2-\alpha\min u_\alpha)ds \text{ is a submartingale.}$$

Hence

$$v_\alpha(x)\le E\left[\int_0^{\tau_r\wedge t}e^{-\alpha s}(h(x(s))+|p(s)|^2-\alpha\min u_\alpha)ds+e^{-\alpha(\tau_r\wedge t)}v_\alpha(x(\tau_r\wedge t))\right],$$

from which

$$v_\alpha(x)\le E\left[\int_0^{\tau_r}e^{-\alpha s}(h(x(s))+|p(s)|^2-\alpha\min u_\alpha)ds+e^{-\alpha\tau_r}v_\alpha(x(\tau_r))\right].$$

Let (λ_2,ϕ_2) be the limit of (32) and (33). Tending α to 0, we get

$$\phi_2(x)-\sup_{|x|=r}\phi_2(x)\le E\left[\int_0^{\tau_r}(h(x(s))+|p(s)|^2-\lambda_2)ds\right].$$

Thus we can obtain

$$\phi_2(x) - \phi_2(0) = \lim_{r\to 0+} \inf \left\{ E\left[\int_0^{\tau_r} ((h(x(t)) + |p(t)|^2) - \lambda_2) dt \right] \ : \ |p(t)| \le 1 \right\},$$
$$\phi_1(x) - \phi_1(0) \ge \phi_2(x) - \phi_2(0),$$

taking $\phi = \phi_2$ and $\phi = \phi_1$ in (47), respectively.

Define $\bar{\phi}(x) = \phi_1(x) - \phi_1(0) - (\phi_2(x) - \phi_2(0)) \ge 0$. Then, by (43)

$$\frac{1}{2}\Delta\bar{\phi} + (\rho(D\phi_1), D\bar{\phi}) \le 0.$$

By the maximal principle [11, Thm 3.5, p. 35], we have

$$\bar{\phi} = 0 \qquad \text{on } B_R \quad \forall R,$$

since $\bar{\phi}$ achieves its minimum at $x = 0$. Therefore, we conclude (46), completing the proof. □

6. An application to ergodic control. We shall study the ergodic control problem to minimize the cost

$$J(p) = \limsup_{T\to\infty} \frac{1}{T} E\left[\int_0^T \{h(x(t)) + |p(t)|^2\} dt \right] \tag{48}$$

over all $\mathcal{P}$ subject to the state equation

$$dx(t) = p(t)dt + dw(t), \qquad x(0) = x, \tag{49}$$

where $\mathcal{P}$ denotes the set of all progressively measurable $\mathcal{F}_t$-adapted processes $p(t)$ such that

$$|p(t)| \le 1,$$

$$\lim_{T\to\infty} \frac{1}{T} E[|x(t)|^{m+1}] = 0 \quad \text{for the response } x(t) \text{ to } p(t)\ . \tag{50}$$

THEOREM 6.1. *We assume* (7), (8), (9), (21). *Then the optimal control* $p^*(t)$ *is given by*

$$p^*(t) = \rho(D\phi(x^*(t)))$$

and the value is given by

$$J(p^*) = \lambda,$$

where $x^*(t)$ *is defined by* (37).

Proof. By (31) and Ito's formula, we have

$$\begin{aligned} E[\phi(x^*(T))] - \phi(x) &= E\left[\int_0^T \left\{ (D\phi(x^*(t)), \rho(D\phi(x^*(t)))) + \frac{1}{2}\Delta\phi(x^*(t)) \right\} dt \right] \\ &= E\left[\int_0^T \left(F(D\phi(x^*(t))) \ - |\rho(D\phi(x^*(t)))|^2 + \frac{1}{2}\Delta\phi(x^*(t)) \right) \ dt \right] \\ &= E\left[\int_0^T (\lambda - h(x^*(t)) - |\rho(D\phi(x^*(t)))|^2) \ dt \right]. \end{aligned}$$

In view of Lemma 5.2 and (45),

$$\lim_{T\to\infty}\frac{1}{T}E[\phi(x^*(T))]=0,$$

and also p^* belongs to $\mathcal{P}$. Hence dividing both sides by T and letting $T\to\infty$, we get

$$J(p^*)=\lim_{T\to\infty}\frac{1}{T}E\left[\int_0^T(h(x^*(t))+|p^*(t)|^2)dt\right]=\lambda.$$

Next, let $p\in\mathcal{P}$ be arbitrary. Along the same line as above, we have

$$E[\phi(x(T))]-\phi(x)\geq E\left[\int_0^T(\lambda-h(x(t))-|p(t)|^2)dt\right].$$

Also, by (36) and (50)

$$\lim_{T\to\infty}\frac{1}{T}E[\phi(x(T))]=0.$$

Thus, we deduce $J(p)\geq\lambda$, which completes the proof. □

7. Concluding remarks. In this paper we have analyzed the Bellman equation of ergodic control problems for diffusion processes that may not possess invariant measures. We have also showed that the gradient estimate of the solution to the Bellman equation in discounted cost problems is bounded in discounted rates, and the equation can be solved without the Lyapunov-type stability conditions discussed in many other sources. It should be noted that the convexity and the polynomial growth property inherited from h play important roles for the existence of a unique solution to the Bellman equation and further for a synthesis of optimal control. We can verify the optimality for the sake of preserving the moment of the optimal trajectory less than the order of time t.

The present paper is a generalization completely covering the results of earlier papers in the 1-dimensional case. The method gives some useful suggestions such that the control problem will be solvable for the long-run average cost defined on a rather wider class of admissible controls than what is known, for which the stochastic system is made unstable, even if it contains the stabilizing matrix A. It is also applicable to solve the Bellman equation with $\min\{(x,p):|p|\leq 1\}$ replacing $F(\xi)$.

The result of this paper provides an answer for some applications to the ergodic production planning problems in unstable manufacturing systems with sales returns, constant demand, and no breakdown. However, the ergodic control of the hybrid system requires the solution to the simultaneous equations with respect to the state of random demands. We need to study the convexity and the polynomial growth property of such Bellman equations.

Acknowledgments. We would like to thank anonymous referees for helpful comments that enabled us to improve the manuscript.

REFERENCES

[1] S. Agmon, A. Douglis, and L. Nirenberg, *Estimates near the boundary for solutions of elliptic partial differential equations satisfying general boundary conditions*, I, Comm. Pure Appl. Math., 12 (1959), pp. 623–727.
[2] A. Bensoussan, *Stochastic Control by Functional Analysis Methods*, North–Holland, New York, 1982.
[3] A. Bensoussan, *Perturbation Methods in Optimal Control*, John Wiley, New York, 1988.
[4] A. Bensoussan and J. Frehse, *On Bellman equations of ergodic control in R^n*, J. Reine Angew. Math., 429 (1992), pp. 125–160.
[5] V. S. Borkar, *Optimal Control of Diffusion Processes*, Pitman Res. Notes Math. Ser. 203, Longman, Essex, 1989.
[6] H. Brezis, *Analyse Fonctionnelle; Théorie et applications*, Masson, Paris, 1992.
[7] G. Da Prato and A. Ichikawa, *Quadratic control for linear time-varying systems*, SIAM J. Control Optim., 28 (1990), pp. 359–381.
[8] W. H. Fleming and H. M. Soner, *Controlled Markov Processes and Viscosity Solutions*, Springer-Verlag, Berlin, 1993.
[9] Y. Fujita and H. Morimoto, *Ergodic control of stochastic differential systems with controller constraints*, Stochastics Stochastics Rep., 58 (1996), pp. 245–257.
[10] Y. Fujita and H. Morimoto, *On Bellman equations in quadratic ergodic control with controller constraints*, Appl. Math. Optim., 39 (1999), pp. 1–15.
[11] D. Gilbarg and N. S. Trudinger, *Elliptic Partial Differential Equations of Second Order*, Springer-Verlag, Berlin, 1983.
[12] M. K. Ghosh, A. Arapostathis, and S. I. Marcus, *Optimal control of switching diffusions with application to flexible manufacturing systems*, SIAM J. Control Optim., 31 (1993), pp. 1183–1204.
[13] J. L. Menaldi and M. Robin, *Ergodic control of reflected diffusions with jumps*, Appl. Math. Optim., 35 (1997), pp. 117–137.
[14] S. P. Sethi and Q. Zhang, *Hierarchical Decision Making in Stochastic Manufacturing Systems*, Birkhäuser, Boston, 1994.

SIAM J. CONTROL OPTIM.
Vol. 38, No. 1, pp. 175–196

H^2 AND H^∞ DESIGN OF SAMPLED-DATA SYSTEMS USING LIFTING. PART I: GENERAL FRAMEWORK AND SOLUTIONS*

LEONID MIRKIN†, HÉCTOR P. ROTSTEIN‡, AND ZALMAN J. PALMOR†

Abstract. This paper presents a complete solution to the H^2 and H^∞ problems for sampled-data systems. As opposed to previous works in the area, it is assumed here that all or some of the sampling function, the discrete-time controller, and the hold function are available for design.

The solution is obtained by transforming the problem to discrete time using the well-known lifting technique. It is then shown that the desired components of the sampled-data controller can be "peeled-off" from the inherently infinite-dimensional description in the lifted-domain. The procedure for doing this last step is central to the approach in this paper.

Both new and revised solutions are presented in this paper. The solution to the H^2 problems is completely new. The solution to the H^∞ problems is presented in a unifying framework and is more transparent than the previous existing solutions in the literature. Transparency pays in the form of clearer results. In particular, a separation structure is established between the design of (sub)optimal hold and sampler.

Key words. sampled-data control, generalized sampler and hold, H^2 optimization, H^∞ optimization, lifting technique

AMS subject classifications. 93C57, 49N10, 93B36

PII. S0363012997329603

1. Introduction.

Since the early 90's, much attention has been paid to H^2 and H^∞ optimal control of continuous-time systems using a sampled-data controller, namely, a controller implemented by a digital computer connected to a plant via A/D (sampler) and D/A (hold) converters. See [17, 4, 30, 8, 16, 27, 3, 25, 29] for an introduction to the subject and pointers to relevant literature. In a great majority of these works, only the discrete part of the sampled-data controller is designed; the sampler and the hold, on the other hand, are selected without taking into consideration the plant dynamics or the control objectives. In most cases, an approximate "ideal" sampler is used to obtain the discrete-time measurements, while zero- or first-order holds are the devices of choice for converting the output of the discrete part of the controller to a continuous-time control signal. Having fixed sampler and hold has several advantages: the behavior of these devices is fairly well understood and hence can be incorporated (at least heuristically) in the controller design process, the design of hardware is considerably simplified, and adequate simulation tools exist, etc. At the same time, fixing the sampler and hold devices irrespective of control considerations may limit the closed-loop performance, especially when the sampling rate is not "fast" enough with respect to the plant dynamics.

Numerous attempts have been made to incorporate the sampler and, especially, the hold into the design process specially by considering the discrete-time performance [15, 1]. Loosely speaking, these designs are based on the open-loop compensation of undesirable plant dynamics (e.g., nonminimum-phase zeros) and manage to achieve

*Received by the editors November 5, 1997; accepted for publication (in revised form) December 14, 1998; published electronically December 15, 1999.

http://www.siam.org/journals/sicon/38-1/32960.html

†Faculty of Mechanical Engineering, Technion—IIT, Haifa 32000, Israel (mersglm@tx.technion.ac.il, palmor@tx.technion.ac.il). The work of the first author was partially supported by the Theodore and Mina Bargman Academic Lectureship.

‡RAFAEL and Department of Electrical Engineering, Technion—IIT, Haifa 32000, Israel (hector@ee.technion.ac.il).

significant improvements of discrete-time performance, usually at the expense of a poor intersampling behavior. As a consequence, these works give rise to criticism [10, 7] and the suspicion that, for instance, a generalized hold device will generically exhibit undesirable robustness properties. It becomes apparent that the design of sampling and/or hold devices should take into account intersampling behavior, even when sampling is fast.

The design of A/D and D/A converters on the basis of continuous-time performance poses some serious technical difficulties. It is not surprising then that few results in this direction are available in the literature. For instance, LQG design of the hold function was discussed by Juan and Kabamba [14] (see also [12]) assuming that the discrete-time part of the controller is fixed. The necessary conditions for optimality of the hold function derived in [14] are quite involved and apply only when rather restrictive constraints (fixed monodromy) on the hold function are imposed. The design of the H^∞ suboptimal hold function is considered by Başar [6] for the state-feedback and by Sun, Nagpal, and Khargonekar [26] for the output feedback cases. In a remarkable work, Tadmor [27] treated the sampled-data H^∞ problem in a general setting, including the design of sampling and/or hold functions. In a more restrictive setting, [22] proposed a simple approach to the H^∞ control when *both* sampler *and* hold are design parameters. In all the H^∞ results above, explicit formulae for the suboptimal sampler and/or hold are obtained. Finally, Bamieh [2] proposed an approximate solution to the ℓ^1 design of the sampling and hold functions.

An important breakthrough in the treatment of sampled-data systems was the introduction of "lifting," an operation that reduces the time-varying sampled data problem into a time-invariant, albeit inherently infinite-dimensional, discrete-time one. The idea is conceptually simple and involves three steps: (i) lift the problem to the discrete time, (ii) solve the resulting formal discrete-time problem in the so-called lifted domain, and (iii) "peel-off" the result back to continuous time. The complexity of dealing with systems in the lifted domain thus far has prevented finding a complete solution to the relevant design problems. The main contribution of the current research is to show how the three design steps can effectively be carried out. In order to do this, a better understanding of the lifted domain and its connection with continuous time is required. In order to streamline the presentation, we have divided the paper into two parts. In the first one, the framework is described and the main solutions are presented, based on the technical results derived in the second part. The second part presents some technical developments which are relevant for the problems under consideration, but also have independent interest.[1] For clarity, we have tried to make the two parts as self-contained as possible. We believe that with this structure we have achieved an adequate tradeoff between readability of the material and completeness.

This paper is organized as follows. In section 2 the problems to be considered throughout the paper are formulated. The next two sections are devoted to the solutions of these problems in the lifted domain. In particular, in section 3 it is shown how a wide class of sampled-data problems can be reduced to a unified "standard problem" in the lifted domain, while in section 4 the lifted solutions to the sampled-data H^2 (section 4.1) and H^∞ (section 4.2) problems are presented and some interesting properties of these solutions are discussed (section 4.3). The lifted solutions are then "peeled-off" (based on the technical results obtained in the companion paper [23]) in

[1] This order reflects our personal taste; a reader needing all facts established before proceeding forward may want to read Part II first.

section 5, which contains complete solutions to the sampled-data H^2 and H^∞ problems, including the (sub)optimal sampling and hold. In section 6 various properties and interpretations of the optimal sampling and hold functions are discussed. In section 6.1 some qualitative conjectures concerning the control oriented design of the D/A converters are also presented.

1.1. Notation. The notation throughout the paper is as follows. As usual, $\mathbb{C}^-$ and $\mathbb{D}$ stand for the open left half plane and the open unit disc, respectively. $\mathbb{R}^n$ denotes the n-dimensional Euclidean space and $L_n^2[0,h]$ denotes the (Hilbert) space of square integrable $\mathbb{R}^n$-valued functions on the interval $[0,h]$. When the dimensions are irrelevant the dimension index is dropped, by simply writing $\mathbb{R}$ and $L^2[0,h]$. For operators on $\mathbb{R} \oplus L^2[0,h]$, $\|\cdot\|_2$ denotes the induced operator norm, while $\|\cdot\|_{HS}$ — the Hilbert–Schmidt operator norm. M' means the transpose of a matrix M and O^* — the adjoint of a Hilbert space operator O. The notations $\sigma(M)$, $\rho(M)$, and $\bar{\sigma}(M)$ stand for the spectrum, the spectral radius, and the maximum singular value of a square matrix M, respectively. $O^{1/2}$ means the square root of $O = O^* \geq 0$.

A "bar" above a variable ($\bar{\zeta}$) denotes discrete-time signals in $\mathbb{R}^n$, while "breve" ($\breve{\zeta}$) denotes discrete-time signals in the lifted domain. Also, we put forward the following operator notation which improves the readability of formulae when both finite- and infinite-dimensional input/output spaces are involved: a bar indicates an operator $\bar{O}$ with both input and output spaces finite dimensional; grave accent — $\grave{O}$, when the input space is finite-dimensional and the output infinite-dimensional one; acute accent — $\acute{O}$, when the input space is infinite-dimensional and the output finite-dimensional one; and finally breve — $\breve{O}$, when both input and output spaces are infinite dimensional.

The compact block notation

$$\left[\begin{array}{c|c} A & B \\ \hline C & D \end{array}\right]$$

denotes (matrix- or operator-valued) transfer functions either in s or in z domain in terms of their state–space realization. To distinguish linear time-invariant (LTI) systems in the time domain from the corresponding transfer functions, the former are denoted by script capital letters, so $G(s)$ implies the transfer function of a continuous-time LTI system $\mathcal{G}$. The lower linear fractional transformation of $\mathcal{K}$ over $\mathcal{P}$ is denoted as $\mathcal{F}_\ell(\mathcal{P},\mathcal{K})$. Finally, $Ric_{\mathbb{C}^-}$ and $Ric_{\mathbb{D}}$ denote the continuous- and discrete-time Riccati functions. These functions are not the standard ones usually found in the literature [32], but rather are defined over product spaces:

$$Ric_{\mathbb{S}} : \mathbb{R}^{(2n+m)\times(2n+m)} \times \mathbb{R}^{(2n+m)\times(2n+m)} \to \mathbb{R}^{n\times n} \times \mathbb{R}^{m\times n},$$

where $\mathbb{S} = \mathbb{C}^-$ or $\mathbb{D}$. There are good reasons for considering these generalized forms; see, for instance, [24, 13] and the Appendix in the companion paper [23] for definitions and properties. The functions $Ric_{\mathbb{S}}$ are not defined over the whole $\mathbb{R}^{(2n+m)\times(2n+m)}$ but rather on a subset called $\operatorname{dom} Ric_{\mathbb{S}}$. It is worth stressing that the function $Ric_{\mathbb{S}}$ has a one-to-one correspondence with the stabilizing solution of appropriate Riccati equations.

2. Problem statement. Consider the single rate sampled-data control systems illustrated in Figure 1, where $\mathcal{P}_c$ is a continuous-time generalized plant and w, z, y, and u are (continuous-time) exogenous input, regulated output, measured output,

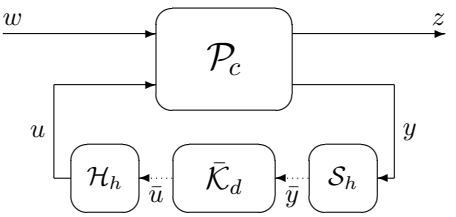


FIG. 1. *General sampled-data setup in time domain.*

and control input, respectively. The sampled-data controller consists of a discrete-time part $\bar{\mathcal{K}}_d$, a sampler $\mathcal{S}_h$, and a hold $\mathcal{H}_h$, assumed to be synchronized and with a sampling period h. Throughout this paper the generalized plant $\mathcal{P}_c$ is assumed to be LTI with the following state–space realization:

$$P_c(s) = \left[\begin{array}{c|cc} A & B_1 & B_2 \\ \hline C_1 & 0 & D_{12} \\ C_2 & D_{21} & 0 \end{array}\right]. \tag{1}$$

The matrices D_{11} and D_{22} are both taken to be zero; this assumption both simplifies the derivations and leads to more transparent results. Moreover, for the H^2 problem the assumption $D_{11} = 0$ is also a necessary condition for the cost function to be finite.

The hold and sampler act on the output of the controller $\bar{u}[k]$ and the measurement $y(t)$, respectively, to generate [15, 1]

$$\big(\mathcal{H}_h \bar{u}\big)(kh + \tau) = \phi_H(\tau)\bar{u}[k] \qquad \forall \tau \in [0, h) \tag{2a}$$

and

$$\big(\mathcal{S}_h y\big)[k] = \int_0^{h^-} \phi_S(\tau) y(kh^- - \tau)\, d\tau. \tag{2b}$$

Here $\phi_H(\tau)$ and $\phi_S(\tau)$ are generalized hold and sampling functions, respectively, defined on the interval $[0, h)$. During the intersample, the hold function shapes the form of the control signal while the sampling function is used to weight the measurements. Note that the integration limit for $\mathcal{S}_h$ is chosen to be h^- rather than h. This is motivated by the fact that any implementable $\bar{\mathcal{K}}_d$ cannot process information instantaneously. As shown in [22], this assumption considerably simplifies the treatment of sampled-data systems.

Depending on whether the sampler and hold devices are considered as design variables or not, there are four possible classes of sampled-data control problems:

C_a: Both sampler and hold are free for design;
C_b: The sampler is fixed but the hold is to be designed;
C_c: The sampler is to be designed but the hold is fixed;
C_d: Both sampler and hold are fixed.

The last case has been extensively treated in the literature, especially when $\mathcal{S}_h$ is the ideal sampler and $\mathcal{H}_h$ is the zero-order hold (see [8] and the references therein). For the class C_d the solutions usually are based on conversions of the sampled-data problem into an equivalent pure discrete one; the latter problem can then be solved by using known techniques. Such an approach requires several intermediate steps and does not give rise to closed-form solutions to the sampled-data H^∞ problem. Hence,

although this paper addresses basically cases C_{a-c}, which involve the design of the sampling and/or hold functions, the results for C_d will also be presented.

The control problems to be dealt with in the paper are the H^2 and H^∞ optimization problems. Under minor reasonable constraints on the functions ϕ_S and ϕ_H in (2) and any LTI $\bar{\mathcal{K}}_d$, the system in Figure 1 is h-periodic in continuous time. For such systems the notions of the H^2 and H^∞ system norms can be introduced in a natural manner [8]. Thus, the following optimization problems can be posed:

OP_{H^2}: Find an LTI $\bar{\mathcal{K}}_d$ and, possibly, a sampling function $\phi_S(\tau)$ and/or a hold function $\phi_H(\tau)$ so that the resulting sampled-data controller internally stabilizes the system in Figure 1 and minimizes the H^2 norm of the closed-loop operator from w to z.

OP_{H^∞}: Given a scalar $\gamma > 0$, find (if they exist) an LTI $\bar{\mathcal{K}}_d$ and, possibly, a sampling function $\phi_S(\tau)$ and/or a hold function $\phi_H(\tau)$ so that the resulting sampled-data controller internally stabilizes the system in Figure 1 and makes the H^∞ norm of the closed-loop operator from w to z less than γ.

Remark 2.1. It is worth stressing that the discrete-time part $\bar{\mathcal{K}}_d$ of the sampled-data controller is always a design parameter. This is in contrast to the approach in [14, 12], where the generalized hold is designed for a fixed $\bar{\mathcal{K}}_d$. Nevertheless, the $\mathcal{H}_h$ and $\mathcal{S}_h$ resulting from the solution to OP_{H^2} and OP_{H^∞} are referred to as H^2-*optimal* and H^∞-*suboptimal* hold and sampler, respectively. The H^2-optimality of $\mathcal{H}_h$ ($\mathcal{S}_h$) here is understood as the ability to design $\bar{\mathcal{K}}_d$ and, possibly, $\mathcal{S}_h$ ($\mathcal{H}_h$) so that the H^2 performance achieved by the sampled-data controller $\mathcal{H}_h\bar{\mathcal{K}}_d\mathcal{S}_h$ supersedes the one achieved using any other hold (sampling) device. Similarly, H^∞-suboptimal hold (sampler) refers to the feasibility of designing $\bar{\mathcal{K}}_d$ and, possibly, $\mathcal{S}_h$ ($\mathcal{H}_h$) so that the overall sampled-data controller be γ-suboptimal.

3. Lifted "standard problem." The treatment of OP_{H^2} and OP_{H^∞} is complicated by their hybrid continuous/discrete nature and their inherent periodicity. To circumvent these difficulties the so-called *lifting technique* of [31, 5, 3] (see also [28, 27]) can be applied.

The notion of lifting is based on a conversion of real valued signals in continuous time into *functional space valued* sequences, that is, sequences that take values not from $\mathbb{R}$ but rather from some general Banach space ($L^2[0,h]$ in this paper). Formally, let $\ell_{L^2[0,h]}$ be the space of sequences of the form $\{\breve{\xi}[k]\}$, where each $\breve{\xi}[k]$ is a function in $L^2[0,h]$, that is,

$$\ell_{L^2[0,h]} = \left\{ \breve{\xi} : \breve{\xi}[k] \in L^2[0,h] \quad \forall k \in \mathbb{Z}_+ \right\}.$$

Then, given any $h > 0$, the lifting operator $\mathcal{W}_h : L^2_e \mapsto \ell_{L^2[0,h]}$ is defined [5, 3] through

$$\breve{\xi} = \mathcal{W}_h\xi \quad \Longleftrightarrow \quad \big(\breve{\xi}[k]\big)(\tau) = \xi(kh+\tau) \quad \forall \tau \in [0,h].$$

It is easy to see that the lifting operator is a linear bijection between L^2_e and $\ell_{L^2[0,h]}$. Moreover, if the domain of $\mathcal{W}_h$ is restricted to the Hilbert space L^2, then the lifting operator can be made an isometry by endowing $\ell_{L^2[0,h]}$ with an appropriate norm. Hence, treating a system $\zeta = \mathcal{G}\omega$ not as a mapping from ω to ζ but rather as a mapping from $\breve{\omega}$ to $\breve{\zeta}$ gives essentially the same system (as an input–output mapping). Indeed, since lifting preserves stability and induced norms, the system $\mathcal{G}$ and its *lifting*,

$$\breve{\mathcal{G}} \doteq \mathcal{W}_h\mathcal{G}\mathcal{W}_h^{-1} : \ell_{L^2[0,h]} \mapsto \ell_{L^2[0,h]},$$

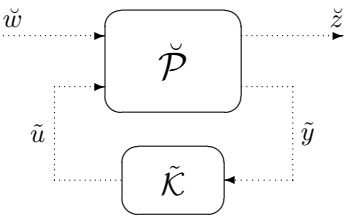


FIG. 2. *"Standard problem" in the lifted domain.*

are equivalent. The advantage of treating systems in the lifting domain stems from the fact that $\breve{\mathcal{G}}$ is time-invariant in discrete time even if $\mathcal{G}$ is h-periodic in continuous time. Hence, any periodic problem in continuous time can be made time-invariant in discrete time.

The application of this idea to the sampled-data setup in Figure 1 is straightforward. In order to convert this setup to a pure discrete time-invariant one, one of the operators $\mathcal{P}_c$, $\mathcal{S}_h$, and $\mathcal{H}_h$ in Figure 1 should be lifted to $\breve{\mathcal{P}}_c \doteq \mathcal{W}_h \mathcal{P}_c \mathcal{W}_h^{-1}$, $\acute{\mathcal{S}}_h \doteq \mathcal{S}_h \mathcal{W}_h^{-1}$, and $\grave{\mathcal{H}}_h \doteq \mathcal{W}_h \mathcal{H}_h$, respectively. The lifted plant $\breve{\mathcal{P}}_c$ is LTI and has a state–space realization

$$\breve{P}_c(z) = \left[\begin{array}{c|cc} \bar{A} & \acute{B}_1 & \acute{B}_2 \\ \hline \grave{C}_1 & \breve{D}_{11} & \breve{D}_{12} \\ \grave{C}_2 & \breve{D}_{21} & \breve{D}_{22} \end{array}\right]. \tag{3}$$

Notice that although $\bar{A}$ is a square matrix with the same dimensions as A, the remaining entries in this state–space representation are operators acting from or/and to infinite-dimensional spaces. The lifted hold $\grave{\mathcal{H}}_h$ is a *memoryless* gain with transfer function

$$\grave{H}_h(z) = \grave{\Phi}_H,$$

while the lifted sampler includes a backward shift [22]:

$$\acute{S}_h(z) = z^{-1} \acute{\Phi}_S.$$

The expressions for the parameters of the plant $\breve{\mathcal{P}}_c$, the hold $\grave{\mathcal{H}}_h$, and the sampler $\acute{\mathcal{S}}_h$ can be derived using standard lifting arguments [3, 21]. They, however, are not essential for the discussion in this section and hence are postponed to the companion paper [23].

What is important in the discussion that follows is the fact that lifting puts all blocks in Figure 1 on equal footing, making them discrete-time LTI systems. Consequently, the four cases $\mathsf{C}_{a\text{–}d}$ of the sampled-data interconnection in Figure 1 with generalized sampler and hold can be reduced to the standard setup of the linear optimal control in Figure 2, where the generalized plant $\breve{\mathcal{P}}$ is *given*, while the controller $\tilde{\mathcal{K}}$ is *to be designed*. The lifted sampler $\acute{\mathcal{S}}_h$ and hold $\grave{\mathcal{H}}_h$ can be absorbed either into the generalized plant or to the controller, depending on whether they are fixed or treated as free design parameters. The only fixed part that will always be absorbed into the controller is the backward shift z^{-1} of $\acute{\mathcal{S}}_h$. The reason for this is twofold: first, the state–space dimension of $\breve{\mathcal{P}}$ is preserved, and second, the characterization of $\tilde{\mathcal{K}}$ is considerably simplified if the operator is *strictly* causal [22]. An optimization problem can therefore always be formulated as a "standard problem," with a *strictly*

proper controller. More specifically, for the four cases considered above, one gets the following:

C_a: Here, both $\acute{\Phi}_S$ and $\grave{\Phi}_H$ are absorbed into the controller, giving

$$\breve{P}(z) = \breve{P}_c(z) \qquad \text{and} \qquad \tilde{K}(z) = \breve{K}(z) \doteq z^{-1}\grave{\Phi}_H \bar{K}_d(z)\acute{\Phi}_S.$$

Any designed controller can always be factorized as follows:

$$\breve{K}(z) = \left[\begin{array}{c|c} \bar{A}_K & \acute{B}_K \\ \hline \grave{C}_K & 0 \end{array}\right] = \grave{\Phi}_H \left[\begin{array}{c|c} \bar{A}_K & \bar{B}_K \\ \hline C_K & 0 \end{array}\right] \acute{\Phi}_S,$$

where $\grave{\Phi}_H$ and $\bar{C}_K$ are any operators of appropriate dimensions such that $\grave{\Phi}_H \bar{C}_K = \grave{C}_K$ and $\bar{B}_K$ and $\acute{\Phi}_S$ are such that $\bar{B}_K \acute{\Phi}_S = \acute{B}_K$. This yields $\bar{\mathcal{K}}_d$, $\mathcal{H}_h$, and $\mathcal{S}_h$.

C_b: In this case $\acute{\Phi}_S$ is absorbed into the lifted plant, but $\grave{\Phi}_H$ is absorbed into the controller, giving

$$\breve{P}(z) = \left[\begin{array}{c|cc} \bar{A} & \acute{B}_1 & \acute{B}_2 \\ \hline \grave{C}_1 & \breve{D}_{11} & \breve{D}_{12} \\ \bar{C}_2 & \acute{D}_{21} & \acute{D}_{22} \end{array}\right] \qquad \text{and} \qquad \tilde{K}(z) = \grave{K}(z) \doteq z^{-1}\grave{\Phi}_H \bar{K}_d(z),$$

where

$$\left[\begin{array}{ccc} \bar{C}_2 & \acute{D}_{21} & \acute{D}_{22} \end{array}\right] \doteq \acute{\Phi}_S \left[\begin{array}{ccc} \grave{C}_2 & \breve{D}_{21} & \breve{D}_{22} \end{array}\right].$$

Any designed controller can always be factorized as follows:

$$\grave{K}(z) = \left[\begin{array}{c|c} \bar{A}_K & \bar{B}_K \\ \hline \grave{C}_K & 0 \end{array}\right] = \grave{\Phi}_H \left[\begin{array}{c|c} \bar{A}_K & \bar{B}_K \\ \hline C_K & 0 \end{array}\right],$$

where $\grave{\Phi}_H$ and $\bar{C}_K$ are any operators of appropriate dimensions such that $\grave{\Phi}_H \bar{C}_K = \grave{C}_K$. This yields both $\bar{\mathcal{K}}_d$ and $\mathcal{H}_h$.

C_c: Now $\grave{\Phi}_H$ is absorbed into the lifted plant, but $\acute{\Phi}_S$ is absorbed into the controller, giving

$$\breve{P}(z) = \left[\begin{array}{c|cc} \bar{A} & \acute{B}_1 & \bar{B}_2 \\ \hline \grave{C}_1 & \breve{D}_{11} & \grave{D}_{12} \\ \grave{C}_2 & \breve{D}_{21} & \grave{D}_{22} \end{array}\right] \qquad \text{and} \qquad \tilde{K}(z) = \acute{K}(z) \doteq z^{-1}\bar{K}_d(z)\acute{\Phi}_S,$$

where

$$\left[\begin{array}{c} \bar{B}_2 \\ \grave{D}_{12} \\ \grave{D}_{22} \end{array}\right] \doteq \left[\begin{array}{c} \acute{B}_2 \\ \breve{D}_{12} \\ \breve{D}_{22} \end{array}\right] \grave{\Phi}_H.$$

Any designed controller can always be factorized as follows:

$$\acute{K}(z) = \left[\begin{array}{c|c} \bar{A}_K & \acute{B}_K \\ \hline C_K & 0 \end{array}\right] = \left[\begin{array}{c|c} \bar{A}_K & \bar{B}_K \\ \hline C_K & 0 \end{array}\right] \acute{\Phi}_S,$$

where $\bar{B}_K$ and $\acute{\Phi}_S$ are any operators of appropriate dimensions such that $\bar{B}_K \acute{\Phi}_S = \acute{B}_K$. This yields both $\bar{\mathcal{K}}_d$ and $\mathcal{S}_h$.

C_d: In this case, both $\acute{\Phi}_S$ and $\grave{\Phi}_H$ are absorbed into the lifted plant, giving

$$\breve{P}(z) = \left[\begin{array}{c|cc} \bar{A} & \acute{B}_1 & \bar{B}_2 \\ \hline \grave{C}_1 & \breve{D}_{11} & \grave{D}_{12} \\ \bar{C}_2 & \acute{D}_{21} & \bar{D}_{22} \end{array}\right] \qquad \text{and} \qquad \tilde{K}(z) = \bar{K}(z) \doteq z^{-1}\bar{K}_d(z),$$

where

$$\bar{D}_{22} \doteq \acute{\Phi}_S \breve{D}_{22} \grave{\Phi}_H.$$

Given a designed $\bar{\mathcal{K}}$, the discrete-time part of the controller is

$$\bar{K}_d(z) = z\bar{K}(z).$$

Thus a designed $\tilde{\mathcal{K}}$ now contains information not only about $\bar{\mathcal{K}}_d$, but also possibly about $\grave{\Phi}_H$ and/or $\acute{\Phi}_S$. If the hold function is the design parameter, then it is completely characterized by the "C"-part of the controller. Likewise, if the sampling function is the design parameter, then it is completely characterized by the "B"-part of the controller.

It is worth stressing that although for all the cases above the problem fits into the unified framework in Figure 2, it is seen that some of the parameters of $\breve{\mathcal{P}}$ and $\tilde{\mathcal{K}}$ might have either finite- or infinite-dimensional input and/or output spaces depending on the situation. In this respect, it is assumed here that the generalized plant takes the form

$$\breve{P}(z) = \left[\begin{array}{c|cc} \bar{A} & \acute{B}_1 & \tilde{B}_2 \\ \hline \grave{C}_1 & \breve{D}_{11} & \tilde{D}_{12} \\ \tilde{C}_2 & \tilde{D}_{21} & \tilde{D}_{22} \end{array}\right] = \left[\begin{array}{c|cc} \bar{A} & \acute{B}_1 & \tilde{B}_2 \\ \hline \grave{C} & \breve{D}_{\bullet 1} & \tilde{D}_{\bullet 2} \end{array}\right] = \left[\begin{array}{c|c} \bar{A} & \acute{B} \\ \hline \grave{C}_1 & \breve{D}_{1\bullet} \\ \tilde{C}_2 & \tilde{D}_{2\bullet} \end{array}\right], \tag{4}$$

while a designed controller is necessarily of the form

$$\tilde{K}(z) = \left[\begin{array}{c|c} \bar{A}_K & \tilde{B}_K \\ \hline \tilde{C}_K & 0 \end{array}\right].$$

The tilde is used to highlight that the corresponding operators may have either finite or infinite dimension depending on whether $\mathcal{S}_h$ and $\mathcal{H}_h$ are fixed or not.

4. Optimal design in the lifted domain. Having reduced the sampled-data setup to the discrete-time LTI "standard problem" in Figure 2, the next step is to reformulate and solve OP_{H^∞} and OP_{H^2} in terms of $\breve{\mathcal{P}}$ and $\tilde{\mathcal{K}}$. This is the purpose of this section.

Start by imposing the following assumptions on the generalized plant (4):

(A1): The operator $\left[\begin{array}{cc} \bar{A} - \lambda I & \tilde{B}_2 \end{array}\right]$ is right invertible $\forall\, |\lambda| \geq 1$.

(A2): The operator $\left[\begin{smallmatrix} \bar{A} - \lambda I \\ \tilde{C}_2 \end{smallmatrix}\right]$ is left invertible $\forall\, |\lambda| \geq 1$.

(A3): The operator $\left[\begin{smallmatrix} \bar{A} - e^{j\theta} I & \tilde{B}_2 \\ \grave{C}_1 & \tilde{D}_{12} \end{smallmatrix}\right]$ is left invertible $\forall \theta \in [0, 2\pi)$.

(A4): The operator $\left[\begin{smallmatrix} \bar{A} - e^{j\theta} I & \acute{B}_1 \\ \tilde{C}_2 & \tilde{D}_{21} \end{smallmatrix}\right]$ is right invertible $\forall \theta \in [0, 2\pi)$.

These assumptions are the counterparts of the standard assumptions imposed on a discrete-time generalized plant, to guarantee input–output stabilizability and nonsingularity of the H^2 and H^∞ problems. Furthermore, the following assumption about $\breve{\mathcal{P}}$ has to be made:

(A5): $\tilde{D}_{21}$ is a bounded operator.

For the standard discrete-time systems with matrix valued parameters this assumption is obviously redundant. For lifted systems, however, this is not always true since sampling operations might be unbounded in the L^2 sense [8, Theorem 9.3.1]. Thus the assumption (A5), together with the assumption $D_{22} = 0$, guarantees that $\mathcal{S}_h$ operates over "proper" signals. Actually, it means that prefiltering by an anti-aliasing filter is provided if necessary.

4.1. H^2 optimization. The notion of the H^2 system norm can be extended in a natural manner to LTI systems in the lifted domain [4] (see also [17]) although it is not an induced norm. More specifically, the H^2 norm of an LTI system $\breve{\mathcal{G}} : \ell_{L^2[0,h]} \mapsto \ell_{L^2[0,h]}$ is defined as follows:

$$\|\breve{\mathcal{G}}\|_{H^2}^2 \doteq \frac{1}{h}\int_0^{2\pi} \|\breve{G}(e^{j\theta})\|_{HS}^2 \, d\theta.$$

This definition is consistent with both the deterministic and the stochastic interpretations of the H^2 system norm in the time domain, and reduces to the usual definition when $\breve{\mathcal{G}}$ is the lifting of an LTI continuous-time system [4].

Thus, OP_{H^2} can now be reformulated in the lifted domain as follows:
$\mathsf{OP}_{H^2}^{eq}$: For the plant $\breve{\mathcal{P}}$ given in (4), find a *strictly causal* $\tilde{\mathcal{K}}$ which internally stabilizes $\breve{\mathcal{P}}$ and minimizes the performance index:

$$J_{H^2} \doteq \|\mathcal{F}_\ell\big(\breve{\mathcal{P}}, \tilde{\mathcal{K}}\big)\|_{H^2}^2.$$

The solution to this H^2 problem is presented next without a proof. This is because, in principle, problem $\mathsf{OP}_{H^2}^{eq}$ is a discrete-time LTI H^2 problem which can be solved by using existing techniques [8, 30]. Detailed treatment of the H^2 problem in the lifted domain in the case when both sampler and hold are fixed and the controller is not constrained to be strictly proper can be found in [20].

The solution of $\mathsf{OP}_{H^2}^{eq}$ requires the following two H^2 DAREs:

$$\begin{aligned}(5a)\quad \bar{X} &= \bar{A}'\bar{X}\bar{A} + \grave{C}_1^*\grave{C}_1 \\ &\quad - (\tilde{B}_2^*\bar{X}\bar{A} + \tilde{D}_{12}^*\grave{C}_1)^*(\tilde{D}_{12}^*\tilde{D}_{12} + \tilde{B}_2^*\bar{X}\tilde{B}_2)^{-1}(\tilde{B}_2^*\bar{X}\bar{A} + \tilde{D}_{12}^*\grave{C}_1)\end{aligned}$$

and

$$\begin{aligned}(5b)\quad \bar{Y} &= \bar{A}\bar{Y}\bar{A}' + \acute{B}_1\acute{B}_1^* \\ &\quad - (\bar{A}\bar{Y}\tilde{C}_2^* + \acute{B}_1\tilde{D}_{21}^*)(\tilde{D}_{21}\tilde{D}_{21}^* + \tilde{C}_2\bar{Y}\tilde{C}_2^*)^{-1}(\bar{A}\bar{Y}\tilde{C}_2^* + \acute{B}_1\tilde{D}_{21}^*)^*.\end{aligned}$$

Notice that for all the four cases described in section 3, both $\bar{X}$ and $\bar{Y}$ of (5) are square matrices (e. g., *finite dimensional*) with the same dimensions as $\bar{A}$. Also, observe that the choice of the hold device affects only the control Riccati equation (5a), while the choice of sampler affects only the filtering Riccati equation (5b).

The main result for the $\mathsf{OP}_{H^2}^{eq}$ can now be stated.

THEOREM 1. *Let* (A1)–(A5) *be satisfied. Then the DAREs* (5) *have the stabilizing solutions* $\bar{X}_2 \geq 0$ *and* $\bar{Y}_2 \geq 0$*, the optimal value of the performance index* J_{H^2} *is*

$$J_{H^2}^{opt} = \frac{1}{h}\Big(\|\breve{D}_{11}\|_{HS}^2 + \mathrm{tr}\big\{\bar{X}_2\acute{B}_1\acute{B}_1^* + \grave{C}_1^*\grave{C}_1\bar{Y}_2 + (\bar{A}'\bar{X}_2\bar{A} - \bar{X}_2)\bar{Y}_2\big\}\Big),$$

and the unique strictly proper controller which achieves the optimal cost $J_{H^2}^{opt}$ *has the state–space representation*

$$\tilde{K}_2(z) = \left[\begin{array}{c|c} \bar{A} + \tilde{B}_2\tilde{F}_2 + \tilde{L}_2\tilde{C}_2 + \tilde{L}_2\tilde{D}_{22}\tilde{F}_2 & -\tilde{L}_2 \\ \hline \tilde{F}_2 & 0 \end{array}\right],$$

where

$$\tilde{F}_2 \doteq -\big(\tilde{D}_{12}^*\tilde{D}_{12} + \tilde{B}_2^*\bar{X}_2\tilde{B}_2\big)^{-1}(\tilde{B}_2^*\bar{X}_2\bar{A} + \tilde{D}_{12}^*\grave{C}_1), \tag{6a}$$

$$\tilde{L}_2 \doteq -(\bar{A}\bar{Y}_2\tilde{C}_2^* + \acute{B}_1\tilde{D}_{21}^*)\big(\tilde{D}_{21}\tilde{D}_{21}^* + \tilde{C}_2\bar{Y}_2\tilde{C}_2^*\big)^{-1}. \tag{6b}$$

Note again that although the parameters of $\breve{\mathcal{P}}$ might be infinite dimensional, the solutions to the Riccati equations as well as the "A" parameter of the optimal controller are always finite-dimensional matrices. The state feedback $\tilde{F}_2$ and the output injection $\tilde{L}_2$ "gains," however, might operate over *infinite-dimensional* output and input spaces, respectively. Consider, for instance, the operator $\tilde{F}_2$. When $\mathcal{H}_h$ is fixed (cases C_c and C_d), the operator compositions $\tilde{D}_{12}^*\grave{C}_1 = \grave{D}_{12}^*\grave{C}_1$ and $\tilde{D}_{12}^*\tilde{D}_{12} = \grave{D}_{12}^*\grave{D}_{12}$, as well as the operator $\tilde{B}_2 = \bar{B}_2$, are finite-dimensional matrices, which can easily be computed [4, 21]. Yet when $\mathcal{H}_h$ is to be designed (cases C_a and C_b), both (5a) and $\tilde{F}_2$ involve quite complicated infinite-dimensional operators. The Riccati equations, like (5a), can in principle be dealt with via the associated Hamiltonians [9], that enables to circumvent some problems. Nevertheless, in order to obtain the optimal state feedback "gain" one has to handle infinite-dimensional operators like $\breve{D}_{12}^*\breve{D}_{12} + \acute{B}_2^*\bar{X}\acute{B}_2$. The techniques for performing such manipulations over the lifted operators were developed recently in [21]. As shown in the companion paper [23], the resulting $\grave{F}_2$ (and $\acute{L}_2$) can be obtained in an elegant form.

4.2. H^∞ optimization. Since for continuous-time systems the H^∞ norm is the induced L^2/L^2 norm, the notion of the H^∞ system norm can be extended to sampled-data systems by defining it as the $\ell^2_{L^2[0,h]}/\ell^2_{L^2[0,h]}$-induced operator norm. As shown in [3], the H^∞ norm can be defined in the frequency domain as follows:

$$\|\breve{\mathcal{G}}\|_{H^\infty} \doteq \max_{\theta\in[0,2\pi)} \|\breve{G}(e^{j\theta})\|_2.$$

Correspondingly, the lifted equivalent of OP_{H^∞} for the setup in Figure 2 takes the following form:

$\mathsf{OP}_{H^\infty}^{eq}$: For the plant $\breve{\mathcal{P}}$ given in (4) and a number $\gamma > 0$, find a *strictly causal* internally stabilizing controller $\tilde{\mathcal{K}}$ such that

$$\|\mathcal{F}_\ell(\breve{\mathcal{P}}, \tilde{\mathcal{K}})\|_{H^\infty} < \gamma,$$

or show that no such controller exists.

Using the same reasoning as in the $\mathsf{OP}_{H^2}^{eq}$ case, the $\mathsf{OP}_{H^\infty}^{eq}$ problem is a discrete-time LTI H^∞ problem [5], with the additional constraint that the controller must be strictly proper. This type of problem was discussed in detail for the case of finite-dimensional parameters in [19].

The solution of the $\mathsf{OP}_{H^\infty}^{eq}$ requires the following two H^∞ DAREs:

$$\begin{aligned}\bar{X} = {} & \bar{A}'\bar{X}\bar{A} + \grave{C}_1^*\grave{C}_1 \\ & - (\breve{D}_{1\bullet}^*\grave{C}_1 + \acute{B}^*\bar{X}\bar{A})^*(\breve{D}_{1\bullet}^*\breve{D}_{1\bullet} - \gamma^2 E_{11} + \acute{B}^*\bar{X}\acute{B})^{-1}(\breve{D}_{1\bullet}^*\grave{C}_1 + \acute{B}^*\bar{X}\bar{A})\end{aligned} \tag{7a}$$

and

$$\text{(7b)}\quad \bar{Y} = \bar{A}\bar{Y}\bar{A}' + \acute{B}_1\acute{B}_1^* - (\acute{B}_1\breve{D}_{\bullet 1}^* + \bar{A}\bar{Y}\grave{C}^*)(\breve{D}_{\bullet 1}\breve{D}_{\bullet 1}^* - \gamma^2 E_{11} + \grave{C}\bar{Y}\grave{C}^*)^{-1}(\acute{B}_1\breve{D}_{\bullet 1}^* + \bar{A}\bar{Y}\grave{C}^*)^*,$$

where $E_{11} \doteq \left[\begin{smallmatrix} I & 0 \\ 0 & 0 \end{smallmatrix}\right]$. As in the H^2 case, both $\bar{X}$ and $\bar{Y}$ solving the H^∞ DAREs (7) are square matrices of the same dimension as $\bar{A}$. Also, the choices of $\mathcal{H}_h$ and $\mathcal{S}_h$ affect only (7a) and (7b), respectively.

Using (7), necessary and sufficient conditions for the existence of a solution to the OP_{H^∞} as well as a particular solution may be established:

THEOREM 2. *Given plant* (4) *such that the assumptions* (A1)–(A5) *hold, then the following statements are equivalent:*

(i) *There exists a controller* $\tilde{\mathcal{K}}$ *which solves the* $\mathsf{OP}_{H^\infty}^{eq}$.

(ii) *The DAREs* (7) *have stabilizing solutions* $\bar{X}_\gamma \geq 0$ *and* $\bar{Y}_\gamma \geq 0$ *such that*

$$\text{(8)}\qquad \left\| \begin{bmatrix} \bar{X}_\gamma^{1/2} & 0 \\ 0 & I \end{bmatrix} \begin{bmatrix} \bar{A} & \acute{B}_1 \\ \grave{C}_1 & \breve{D}_{11} \end{bmatrix} \begin{bmatrix} \bar{Y}_\gamma^{1/2} & 0 \\ 0 & I \end{bmatrix} \right\|_2 < \gamma.$$

Given that the conditions of part (ii) *hold, then the matrix* $\bar{Z}_\gamma \doteq (I - \gamma^{-2}\bar{Y}_\gamma\bar{X}_\gamma)^{-1}$ *is well defined and one controller which solves* $\mathsf{OP}_{H^\infty}^{eq}$ *is*

$$\tilde{K}_\gamma(z) = \left[\begin{array}{c|c} \bar{A} + \acute{B}\grave{F}_\gamma + \bar{Z}_\gamma\tilde{L}_{\gamma 2}(\tilde{C}_2 + \tilde{D}_{2\bullet}\grave{F}_\gamma) & -\bar{Z}_\gamma\tilde{L}_{\gamma 2} \\ \hline \tilde{F}_{\gamma 2} & 0 \end{array}\right],$$

where

$$\text{(9a)}\qquad \grave{F}_\gamma \doteq -(\breve{D}_{1\bullet}^*\breve{D}_{1\bullet} - \gamma^2 E_{11} + \acute{B}^*\bar{X}_\gamma\acute{B})^{-1}(\breve{D}_{1\bullet}^*\grave{C}_1 + \acute{B}^*\bar{X}_\gamma\bar{A}) = \begin{bmatrix} \grave{F}_{\gamma 1} \\ \tilde{F}_{\gamma 2} \end{bmatrix},$$

$$\text{(9b)}\qquad \acute{L}_\gamma \doteq -(\acute{B}_1\breve{D}_{\bullet 1}^* + \bar{A}\bar{Y}_\gamma\grave{C}^*)(\breve{D}_{\bullet 1}\breve{D}_{\bullet 1}^* - \gamma^2 E_{11} + \grave{C}\bar{Y}_\gamma\grave{C}^*)^{-1} = \begin{bmatrix} \acute{L}_{\gamma 1} & \tilde{L}_{\gamma 2} \end{bmatrix}.$$

As seen, the direct solution to the H^∞ problem involves infinite-dimensional operators even when both sampling and hold functions are fixed (case C_d above). Nevertheless, as shown in the companion paper [23] the techniques developed in [21] enable the reduction of all these operator compositions to finite-dimensional matrices, which can be easily computed.

4.3. Discussion. Although the lifted solutions of Theorems 1 and 2 are not readily implementable, some insight into the solutions to OP_{H^2} and OP_{H^∞} can be gained already at this stage. This is shown by the following remarks.

Remark 4.1. As noted in section 3, if the hold (sampling) function is the design parameter, then it is characterized by the "C" ("B")-part of the lifted controller. Hence, the H^2-optimal hold is characterized by the operator $\tilde{F}_2$, which becomes $\grave{F}_2$ in that case, while the H^2-optimal sampler — by the operator $\tilde{L}_2$ ($\acute{L}_2$). By inspecting (6) and (5) one can see that $\tilde{F}_2$ depends only on the subsystem from $\tilde{u}$ to $\breve{z}$, while $\tilde{L}_2$ is completely characterized by the properties of the subsystem from $\breve{w}$ to $\tilde{y}$. Therefore, there is complete separation between the design of $\mathcal{H}_h$ and $\mathcal{S}_h$ in the H^2 case. This separation is reminiscent of that between the state feedback and the state estimation in the standard H^2 (LQG) design and thus is not a surprise.

Remark 4.2. More surprising is the fact that similar separation arguments apply to the design of the H^∞ suboptimal hold and sampler as well. This fact is worth

stressing, given the coupling which exists between the full information and the output estimation problems in the H^∞ output feedback control [32]. Although the "B"-part of the lifted H^∞ suboptimal controller contains the coupling term $\bar{Z}_\gamma$, the latter is *finite dimensional* and hence can always be absorbed into the discrete-time part of the controller. Consequently, the H^∞ suboptimal hold and sampler are characterized by the operators $\tilde{F}_{\gamma 2}$ and $\tilde{L}_{\gamma 2}$, respectively. It is seen from (9) and (7) that these operators are not completely independent, since both of them are affected by the subsystem from $\breve{w}$ to $\breve{z}$. Nevertheless, $\tilde{F}_{\gamma 2}$ does *not* depend on the measurement $\tilde{y}$ and $\tilde{L}_{\gamma 2}$ does *not* depend on the control $\tilde{u}$. Hence, *both* the hold *and* the sampling functions can be designed independently from one another in the H^∞ case also. This is in contrast to previous works in the literature [28, 22], where the H^∞ suboptimal sampler depended on the hold and it is not clear how to recover the separation.

Remark 4.3. It is not difficult to verify that as $\gamma \to \infty$ the H^∞ Riccati equations (7) reduce to the corresponding H^2 ones (5). Moreover, as $\gamma \to \infty$, the conditions of statement (ii) of Theorem 2 hold automatically, $\bar{Z}_\gamma = I$ (since both $\bar{X}_\gamma$ and $\bar{Y}_\gamma$ are uniformly bounded),

$$\lim_{\gamma\to\infty} \grave{F}_\gamma = \begin{bmatrix} 0 \\ \tilde{F}_2 \end{bmatrix} \quad \text{and} \quad \lim_{\gamma\to\infty} \acute{L}_\gamma = \begin{bmatrix} 0 & \tilde{L}_2 \end{bmatrix}.$$

Thus in the limit case $\gamma \to \infty$ the H^∞-suboptimal controller $\tilde{K}_\gamma(z)$ approaches $\tilde{K}_2(z)$, the H^2-optimal one.

5. Main results. The solutions to the OP_{H^2} and OP_{H^∞} can now be presented. Again, these solutions are strongly based on the technical machinery developed in the companion paper, so that some readers may prefer to read the material there before proceeding. To keep the presentation clear, it is assumed that a zero-order hold is used in cases C_c and C_d, while an ideal sampler is used in cases C_b and C_d. From Remarks 4.1 and 4.2 it follows that such simplifications do not affect the results concerning the (sub)optimal sampler and hold. More general $\mathcal{S}_h$ and $\mathcal{H}_h$ can be treated in a similar fashion. Also, as follows from Remark 4.3, the solution of Theorem 2 approaches that of Theorem 1 as $\gamma \to \infty$. For that reason, only the H^∞ results are presented below. The H^2 results can be obtained from the H^∞ ones by the simple substitution $\gamma^{-1} = 0$. The only part of the H^2 solution which needs to be treated independently is the calculation of the optimal performance $J^{opt}_{H^2}$. See section 5.5 for further details.

Let

$$\Gamma_X \doteq \begin{bmatrix} -D_{12}'C_1 & -B_2' & -D_{12}'D_{12} \\ A & \gamma^{-2}B_1B_1' & B_2 \\ -C_1'C_1 & -A' & -C_1'D_{12} \end{bmatrix},$$

$$\Gamma_Y \doteq \begin{bmatrix} -D_{21}B_1' & -C_2 & -D_{21}D_{21}' \\ A' & \gamma^{-2}C_1'C_1 & C_2' \\ -B_1B_1' & -A & -B_1D_{21}' \end{bmatrix},$$

and

$$\Sigma \doteq \exp\left(\begin{bmatrix} 0 & -D_{12}'C_1 & -B_2' & -D_{12}'D_{12} \\ 0 & A & \gamma^{-2}B_1B_1' & B_2 \\ 0 & -C_1'C_1 & -A' & -C_1'D_{12} \\ 0 & 0 & 0 & 0 \end{bmatrix} h\right) = \begin{bmatrix} I & \Sigma_{12} & \Sigma_{13} & \Sigma_{14} \\ 0 & \Sigma_{22} & \Sigma_{23} & \Sigma_{24} \\ 0 & \Sigma_{32} & \Sigma_{33} & \Sigma_{34} \\ 0 & 0 & 0 & I \end{bmatrix}.$$

Define the matrices

$$\Sigma_X \doteq \begin{bmatrix} \Sigma_{12} & \Sigma_{13} & \Sigma_{14} \\ \Sigma_{22} & \Sigma_{23} & \Sigma_{24} \\ \Sigma_{32} & \Sigma_{33} & \Sigma_{34} \end{bmatrix},$$

$$\Sigma_Y \doteq \begin{bmatrix} 0 & C_2 & 0 \\ \Sigma'_{22} & -\gamma^{-2}\Sigma'_{32} & \Sigma'_{22}C'_2 \\ -\gamma^2\Sigma'_{23} & \Sigma'_{33} & -\gamma^2\Sigma'_{23}C'_2 \end{bmatrix},$$

and the matrix function of an argument $\nu \in \mathbb{Z}^+$

$$\Lambda_\nu \doteq \begin{bmatrix} 0 & 0 & 0 \\ I_n & 0 & 0 \\ 0 & I_n & 0 \end{bmatrix} \in \mathbb{R}^{(\nu+2n)\times(2n+\nu)}.$$

The following quantity is also required:

$$\gamma_0 \doteq \|\mathcal{P}_{c,11}\|_{L^2[0,h]}, \tag{10}$$

where $P_{c,11}(s) = C_1(sI - A)^{-1}B_1$ (in other words, γ_0 is the $L^2[0,h]$-induced norm of the subsystem from w to z). This quantity can be computed as described in [8, Section 13.5] or [11]. Since $\gamma_0 = \|\breve{D}_{11}\|_2$, it is clear that γ_0 is the lower bound for the achievable H^∞ performance in sampled-data systems under any choice of $\mathcal{S}_h$ and $\mathcal{H}_h$. Hence it is natural to consider only the cases where $\gamma > \gamma_0$.

Assumptions (A1)–(A5) are formulated in terms of the system description in the lifted domain. Results in the companion paper [23] allow us to replace them with more readily checkable conditions. In particular, when $\mathcal{H}_h$ is the design parameter, assumptions (A1) and (A3) can be equivalently formulated as

(A1′): The pair (A, B_2) is $\mathbb{C}^-$-stabilizable;

(A3′): The matrix $\left[\begin{smallmatrix} A-j\omega I & B_2 \\ C_1 & D_{12} \end{smallmatrix}\right]$ is left invertible $\forall\omega \in \mathbb{R}$ and $D'_{12}D_{12} > 0$,

while when $\mathcal{H}_h$ is the zero-order hold ($\phi_H(\tau) = I$) they can be replaced with

(A1″): The pair $(\Sigma_{22}, \Sigma_{24})$ is $\mathbb{D}$-stabilizable for any $\gamma > \gamma_0$;

(A3″): The matrix

$$\begin{bmatrix} \Sigma_{12} & \Sigma_{14} \\ \Sigma_{22} - e^{j\theta}I & \Sigma_{24} \\ \Sigma_{32} & \Sigma_{34} \end{bmatrix}$$

is left invertible $\forall\theta \in [0, 2\pi)$ and any $\gamma > \gamma_0$.

Analogously, when $\mathcal{S}_h$ is is the design parameter, assumptions (A2) and (A4) can be equivalently expressed as

(A2′): The pair (C_2, A) is $\mathbb{C}^-$-detectable;

(A4′): The matrix $\left[\begin{smallmatrix} A-j\omega I & B_1 \\ C_2 & D_{21} \end{smallmatrix}\right]$ is right invertible $\forall\omega \in \mathbb{R}$ and $D_{21}D'_{21} > 0$

and (A5) is redundant, while when it is the ideal sampler ($\phi_S(\tau) = \delta(\tau)$), (A2) and (A4) can be replaced with

(A2″): The pair (C_2, Σ_{22}) is $\mathbb{D}$-detectable for any $\gamma > \gamma_0$;

(A4″): The matrix $\left[\begin{smallmatrix} \Sigma_{22}-e^{j\theta}I & \gamma^2\Sigma_{23} \\ C_2 & 0 \end{smallmatrix}\right]$ is right invertible $\forall\theta \in [0, 2\pi)$ and any $\gamma > \gamma_0$;

(A5″): $D_{21} = 0$.

The following remarks are in order.

Remark 5.1. Although (A1″) is equivalent to (A1) only when $\gamma \to \infty$, OP_{H^∞} will have no solutions whenever (A1″) is violated. Hence, (A1) can be replaced by (A1″)

without loss of generality. Just note that unless $\gamma = \infty$, the violation of (A1″) does not necessarily imply that $(\bar{A}, \bar{B}_2)$ is not $\mathbb{D}$-stabilizable. The same is true regarding (A2″) and the $\mathbb{D}$-detectability of $(\bar{C}_2, \bar{A})$.

Remark 5.2. Assumptions (A1″)–(A4″) are expressed in terms of the parameters of the matrix exponential Σ. This, in principle, does not complicate the computations, since Σ is required for the solution of OP_{H^2} and OP_{H^∞} in any case. In some cases, however, it is possible to verify (A1)–(A4) directly in terms of the parameters of the continuous-time plant $\mathcal{P}$. In particular, if the sampling period h is nonpathological with respect to the eigenvalues of A [8], then (A1) and (A2) are equivalent to (A1′) and (A2′), respectively, even when $\mathcal{H}_h$ and $\mathcal{S}_h$ are fixed. Furthermore, it is shown in [30, section 5] that

- if the continuous-time subsystem from u to z, $\mathcal{P}_{12}$, is left invertible and $\mathcal{H}_h$ is the zero-order hold, then (A3″) holds iff (C_2, A) has no unobservable $j\omega$-axis eigenvalues and 0 is not a zero of $P_{12}(s)$;
- if the continuous-time subsystem from w to y is right invertibe and $\mathcal{S}_h$ is the ideal sampler, then (A4″) holds iff (A, B_2) has no unobservable $j\omega$-axis eigenvalues.

It is worth stressing, however, that (A1) and/or (A2) may hold even if h is pathological and (A3) ((A4)) may hold even if the continuous-time subsystem from u to z (from w to y) is not left (right) invertible.

5.1. Both hold and sampler are free. In case C_a the solution to OP_{H^∞} is as follows.

THEOREM 3 (C_a). *Given plant* (1) *such that assumptions* (A1′)–(A4′) *hold, then for any* $\gamma > \gamma_0$ *the matrix* Σ_{33} *is nonsingular and the following statements are equivalent:*

(i) *There exist* $\bar{\mathcal{K}}_d$, $\mathcal{H}_h$, *and* $\mathcal{S}_h$ *which solve* OP_{H^∞}.

(ii) $(\Gamma_X, \Lambda_m) \in \mathrm{dom}(Ric_{\mathbb{C}^-})$ *and* $(\Gamma_Y, \Lambda_r) \in \mathrm{dom}(Ric_{\mathbb{C}^-})$ *and the following conditions hold:*

(a) $X_\gamma \ge 0$ *and* $\rho(X_\gamma \Sigma_{23} \Sigma_{33}^{-1}) < 1$,

(b) $Y_\gamma \ge 0$ *and* $\rho(\Sigma_{33}^{-1} \Sigma_{32} Y_\gamma) < \gamma^2$,

(c) $\rho\big(Y_\gamma(\Sigma_{33} + \gamma^{-2}\Sigma_{32}Y_\gamma)^{-1} X_\gamma (\Sigma'_{33} - \Sigma'_{23}X_\gamma)^{-1}\big) < \gamma^2$,

where $(X_\gamma, F_\gamma) = Ric_{\mathbb{C}^-}(\Gamma_X, \Lambda_m)$ *and* $(Y_\gamma, L'_\gamma) = Ric_{\mathbb{C}^-}(\Gamma_Y, \Lambda_r)$.

Furthermore, if the conditions of part (ii) *hold, then* $Z_\gamma \doteq (I - \gamma^{-2} Y_\gamma X_\gamma)^{-1}$ *is well defined, and one possible choice for* $\bar{\mathcal{K}}_d$, $\mathcal{H}_h$, *and* $\mathcal{S}_h$ *is*

$$\bar{K}_d(z) = z\left[\begin{array}{c|c} Z_\gamma \Theta_{12} + \Theta_{22} & Z_\gamma \\ \hline I & 0 \end{array}\right],$$

where

$$\left[\begin{array}{cc} \Theta_{11} & \Theta_{12} \\ 0 & \Theta_{22} \end{array}\right] \doteq \exp\left(\left[\begin{array}{cc} A + \gamma^{-2} Y_\gamma C_1' C_1 + L_\gamma C_2 & L_\gamma (C_2 + \gamma^{-2} D_{21} B_1' X_\gamma) \\ 0 & A + \gamma^{-2} B_1 B_1' X_\gamma + B_2 F_\gamma \end{array}\right] h\right)$$

and

$$\phi_H(\tau) = F_\gamma\, e^{(A + \gamma^{-2} B_1 B_1' X_\gamma + B_2 F_\gamma)\tau}, \tag{11a}$$

$$\phi_S(\tau) = -e^{(A + \gamma^{-2} Y_\gamma C_1' C_1 + L_\gamma C_2)\tau}\, L_\gamma. \tag{11b}$$

Proof (outline). Actually, it suffices to prove that the solution given in the theorem is equivalent to the one in Theorem 2. To this end, note that by [23, Lemma 3]

and its dual, the DAREs (7) have stabilizing solutions iff $(\Gamma_X, \Lambda_m) \in \mathrm{dom}(Ric_{\mathbb{C}^-})$ and $(\Gamma_Y, \Lambda_r) \in \mathrm{dom}(Ric_{\mathbb{C}^-})$. Then items (a)–(c) are equivalent to (8) by [23, Lemma 8].

Now consider the lifted controller $\bar{K}_\gamma(z)$ $(= \breve{K}_\gamma(z))$ in Theorem 2. It is clear that $\grave{\Phi}_H = \grave{F}_{\gamma 2}$ and $\acute{\Phi}_S = \acute{L}_{\gamma 2}$ can be chosen. Then (11) follow directly from [23, Lemma 3] and its dual. It also follows from that lemma that

$$\bar{A} + \acute{B}\grave{F}_\gamma = \Theta_{22} \qquad \text{and} \qquad \acute{L}_2(\grave{C}_2 + \breve{D}_{2\bullet}\grave{F}_\gamma) = \Theta_{12},$$

from which the formula for $\bar{K}_d(z)$ follows. □

Note that the H^∞ problem for case C_a was already treated by Tadmor [27] (see also [22]). The solution in Theorem 3, however, is simpler and has the intriguing separation structure between $\mathcal{H}_h$ and $\mathcal{S}_h$ (the suboptimal sampler in [27] does depend on the suboptimal hold).

5.2. Free hold and fixed sampler. In case C_b the solution to OP_{H^∞} is as follows.

THEOREM 4 (C_b). *Given plant* (1) *such that assumptions* (A1′), (A2″), (A3′), (A4″), (A5″) *hold and let the sampling function be* $\phi_S(\tau) = \delta(\tau)$. *Then for any* $\gamma > \gamma_0$ *the following statements are equivalent:*

(i) *There exist* $\bar{\mathcal{K}}_d$ *and* $\mathcal{H}_h$ *which solve* OP_{H^∞}.

(ii) $(\Gamma_X, \Lambda_m) \in \mathrm{dom}(Ric_{\mathbb{C}^-})$ *and* $(\Sigma_Y, \Lambda_r) \in \mathrm{dom}(Ric_{\mathbb{D}})$ *and the following conditions hold:*

(a) $X_\gamma \geq 0$ *and* $\rho(X_\gamma \Sigma_{23} \Sigma_{33}^{-1}) < 1$,

(b) $\bar{Y}_\gamma \geq 0$ *and* $\rho(\Sigma_{33}^{-1}\Sigma_{32}\bar{Y}_\gamma) < \gamma^2$,

(c) $\rho\big(\bar{Y}_\gamma(\Sigma_{33} + \gamma^{-2}\Sigma_{32}\bar{Y}_\gamma)^{-1} X_\gamma (\Sigma'_{33} - \Sigma'_{23}X_\gamma)^{-1}\big) < \gamma^2$,

where $(X_\gamma, F_\gamma) = Ric_{\mathbb{C}^-}(\Gamma_X, \Lambda_m)$ *and* $(\bar{Y}_\gamma, \bar{L}'_\gamma) = Ric_{\mathbb{D}}(\Sigma_Y, \Lambda_r)$.

Furthermore, if the conditions of part (ii) *hold, then* $Z_\gamma \doteq (I - \gamma^{-2}\bar{Y}_\gamma X_\gamma)^{-1}$ *is well defined and one possible choice for* $\bar{\mathcal{K}}_d$ *and* $\mathcal{H}_h$ *is*

$$\bar{K}_d(z) = z\left[\begin{array}{c|c} (I + Z_\gamma \bar{L}_\gamma C_2) e^{(A+\gamma^{-2}B_1B_1'X_\gamma + B_2F_\gamma)h} & -Z_\gamma \bar{L}_\gamma \\ \hline I & 0 \end{array}\right]$$

and $\phi_H(\tau)$ *as in* (11a).

Proof (outline). The first part can be proven in a similar fashion as the first part of Theorem 3, except that the DARE (7b) is solved by using [23, Lemma 7]. To get $\bar{K}_d(z)$ just note that if $\mathcal{S}_h$ is the ideal sampler and $D_{21} = 0$, then

$$\left[\begin{array}{cc} \bar{C}_2 & \acute{D}_{2\bullet} \end{array}\right] = C_2 \left[\begin{array}{cc} \bar{A} & \acute{B} \end{array}\right], \tag{12}$$

which completes that proof. □

The H^∞ problem for case C_b was considered in [26], where the suboptimal hold function has also the form (11a). Yet the Riccati equations there are nontrivially coupled and the existence conditions are computationally more complicated than conditions (a)–(c) in Theorem 4. In particular, the nonsingularity of a matrix function of time t on the whole interval $[0, h]$ should be verified in [26] for each γ. On the other hand, a similar test in our case is to be performed only once, when γ_0 is calculated [8, section 13.5].

5.3. Fixed hold and free sampler. In case C_c, the solution to OP_{H^∞} is as follows.

THEOREM 5 (C_c). *Given plant* (1) *such that assumptions* (A1″), (A2′), (A3″), (A4′) *hold and let the hold function be* $\phi_H(\tau) = I$. *Then for any* $\gamma > \gamma_0$ *the following statements are equivalent:*

(i) *There exist* $\bar{\mathcal{K}}_d$ *and* $\mathcal{S}_h$ *which solve* OP_{H^∞}.
(ii) $(\Sigma_X, \Lambda_m) \in \mathrm{dom}(Ric_{\mathbb{D}})$ *and* $(\Gamma_Y, \Lambda_r) \in \mathrm{dom}(Ric_{\mathbb{C}^-})$ *and the following conditions hold:*
  (a) $\bar{X}_\gamma \geq 0$ *and* $\rho(\bar{X}_\gamma \Sigma_{23} \Sigma_{33}^{-1}) < 1$,
  (b) $Y_\gamma \geq 0$ *and* $\rho(\Sigma_{33}^{-1} \Sigma_{32} Y_\gamma) < \gamma^2$,
  (c) $\rho\big(Y_\gamma(\Sigma_{33} + \gamma^{-2}\Sigma_{32}Y_\gamma)^{-1}\bar{X}_\gamma(\Sigma'_{33} - \Sigma'_{23}\bar{X}_\gamma)^{-1}\big) < \gamma^2$,
  where $(\bar{X}_\gamma, \bar{F}_\gamma) = Ric_{\mathbb{D}}(\Sigma_X, \Lambda_m)$ *and* $(Y_\gamma, L'_\gamma) = Ric_{\mathbb{C}^-}(\Gamma_Y, \Lambda_r)$.

Furthermore, if the conditions of part (ii) *hold, then* $Z_\gamma \doteq (I - \gamma^{-2} Y_\gamma \bar{X}_\gamma)^{-1}$ *is well defined and one possible choice for* $\bar{\mathcal{K}}_d$ *and* $\mathcal{S}_h$ *is*

$$\bar{K}_d(z) = z\left[\begin{array}{c|c} \Psi_{11} + \Psi_{12}\bar{F}_\gamma Z_\gamma & I \\ \hline \bar{F}_\gamma Z_\gamma & 0 \end{array}\right],$$

where

$$\left[\begin{array}{cc} \Psi_{11} & \Psi_{12} \\ 0 & I \end{array}\right] \doteq \exp\left(\left[\begin{array}{cc} A + \gamma^{-2} Y_\gamma C_1' C_1 + L_\gamma C_2 & B_2 + \gamma^{-2} Y_\gamma C_1' D_{12} \\ 0 & 0 \end{array}\right] h\right)$$

and $\phi_S(\tau)$ *as in* (11b).

Proof (outline). The first part can be proven in a way similar to the first part of Theorem 3, except that the DARE (7a) is solved by [23, Lemma 7]. In order to obtain the formula for $\bar{K}_d(z)$ it is more convenient to use the dual form of the H^∞ suboptimal controller, i.e.,

$$\tilde{K}(z) = \left[\begin{array}{c|c} \bar{A} + \acute{L}_\gamma \grave{C} + (\tilde{B}_2 + \acute{L}_\gamma \tilde{D}_{\bullet 2})\tilde{F}_{\gamma 2}\bar{Z}_\gamma & -\tilde{L}_{\gamma 2} \\ \hline \tilde{F}_{\gamma 2}\bar{Z}_\gamma & 0 \end{array}\right]$$

(which becomes $\acute{K}(z)$ in this case). Then, using the dual result to [23, Lemma 3] one can get that

$$\bar{A} + \acute{L}_\gamma \grave{C} = \Psi_{11} \qquad \text{and} \qquad \bar{B}_2 + \acute{L}_\gamma \grave{D}_{\bullet 2} = \Psi_{12},$$

from which the formula for $\bar{K}_d(z)$ follows. □

5.4. Both hold and sampler are fixed. In case C_d the solution to OP_{H^∞} is as follows.

THEOREM 6 (C_d). *Given plant* (1) *such that assumptions* ($\mathsf{A1}''$)–($\mathsf{A5}''$) *hold and let the hold function be* $\phi_H(\tau) = I$ *and the sampling function be* $\phi_S(\tau) = \delta(\tau)$. *Then for any* $\gamma > \gamma_0$ *the following statements are equivalent:*
(i) *There exists* $\bar{\mathcal{K}}_d$ *which solves* OP_{H^∞}.
(ii) $(\Sigma_X, \Lambda_m) \in \mathrm{dom}(Ric_{\mathbb{D}})$ *and* $(\Sigma_Y, \Lambda_r) \in \mathrm{dom}(Ric_{\mathbb{D}})$ *and the following conditions hold:*
  (a) $\bar{X}_\gamma \geq 0$ *and* $\rho(\bar{X}_\gamma \Sigma_{23} \Sigma_{33}^{-1}) < 1$,
  (b) $\bar{Y}_\gamma \geq 0$ *and* $\rho(\Sigma_{33}^{-1} \Sigma_{32} \bar{Y}_\gamma) < \gamma^2$,
  (c) $\rho\big(\bar{Y}_\gamma(\Sigma_{33} + \gamma^{-2}\Sigma_{32}\bar{Y}_\gamma)^{-1}\bar{X}_\gamma(\Sigma'_{33} - \Sigma'_{23}\bar{X}_\gamma)^{-1}\big) < \gamma^2$,
  where $(\bar{X}_\gamma, \bar{F}_\gamma) = Ric_{\mathbb{D}}(\Sigma_X, \Lambda_m)$ *and* $(\bar{Y}_\gamma, \bar{L}'_\gamma) = Ric_{\mathbb{D}}(\Sigma_Y, \Lambda_r)$.

Furthermore, if the conditions of part (ii) *hold, then* $Z_\gamma \doteq (I - \gamma^{-2} \bar{Y}_\gamma \bar{X}_\gamma)^{-1}$ *is well defined and one possible choice for* $\bar{\mathcal{K}}_d$ *is*

$$\bar{K}_d(z) = z\left[\begin{array}{c|c} (I + Z_\gamma \bar{L}_\gamma C_2)(\Sigma_{22} + \Sigma_{23}\bar{X}_\gamma + \Sigma_{24}\bar{F}_\gamma) & -Z_\gamma \bar{L}_\gamma \\ \hline \bar{F}_\gamma & 0 \end{array}\right].$$

Proof (outline). The first part can be proven in a way similar to the first part of Theorem 3, except that instead of [23, Lemma 3] one has to use [23, Lemma 7]. The formula for $\bar{K}_d(z)$ then follows by (12) and the equality

$$\bar{A} + \acute{B}\grave{F}_\gamma = \Sigma_{22} + \Sigma_{23}\bar{X}_\gamma + \Sigma_{24}\bar{F}_\gamma,$$

which, in turn, follows from the fact that $(\bar{X}_\gamma, \bar{F}_\gamma) = Ric_{\mathbb{D}}(\Sigma_X, \Lambda_m)$. □

Although this case has been extensively studied in the literature in both H^2 and H^∞ cases (see the references in [8]), the conventional approach to the solution is based on the following two-step procedure: first, a sampled-data problem is converted to an equivalent discrete-time finite-dimensional one; and second, the latter problem is solved for $\bar{\mathcal{K}}_d$ using standard discrete H^2 or H^∞ methods. The available solutions thus are quite involved. Moreover, because of the presence of the intermediate step, the effect of the original parameters of the continuous-time problem is difficult to trace back.

To the best of our knowledge, Theorem 6 yields the first closed-form solution to the sampled-data H^∞ problem. It is worth stressing that from the computational point of view the solutions to both H^2 and H^∞ sampled-data problems given in Theorem 6 are compatible with those for pure discrete-time treatments.

5.5. The optimal H^2 performance. In this subsection the formula for the H^2 optimal performance index $J^{opt}_{H^2}$ given in Theorem 1, is expressed in terms of the matrices A, B_1, and C_1. Such a formula can, in principle, be assembled from the results in [4, p. 11]. The result below, however, is simpler in the sense that the matrix exponential of smaller dimension is to be computed.

Form the matrix exponential

$$\Delta \doteq \exp\left(\begin{bmatrix} -A' & C_1'C_1 & 0 \\ 0 & A & B_1B_1' \\ 0 & 0 & -A' \end{bmatrix} h\right) = \begin{bmatrix} \Delta_{11} & \Delta_{12} & \Delta_{13} \\ 0 & \Delta_{22} & \Delta_{23} \\ 0 & 0 & \Delta_{11} \end{bmatrix}.$$

Then this lemma follows.

LEMMA 1. *Let X_2 and Y_2 be the stabilizing solutions to the Riccati equations* (5)*, then the optimal value of the performance index J_{H^2} is*

$$J^{opt}_{H^2} = \tfrac{1}{h}\operatorname{tr}\big(\Delta_{22}'(\Delta_{13} + X_2\Delta_{23} + \Delta_{12}Y_2 + X_2\Delta_{22}Y_2) - X_2Y_2\big).$$

Proof. As follows from [23, Props. 8 and 9] (subject to (s.t.) $B_\gamma = 0$) and their duals:

$$\bar{A} = \Delta_{22}, \quad \acute{B}_1\acute{B}_1^* = \Delta_{23}\Delta_{22}', \quad \text{and} \quad \grave{C}_1^*\grave{C}_1 = \Delta_{22}'\Delta_{12}.$$

Then, using the integral expression for the Hilbert–Schmidt norm [4] one can write

$$\begin{aligned} \|\breve{D}_{11}\|^2_{HS} &= \operatorname{tr}\left(\int_0^h \int_0^t C_1 e^{A(t-s)}B_1B_1'e^{A'(t-s)}C_1' ds\, dt\right) \\ &= \operatorname{tr}\left(e^{A'h}\int_0^h e^{-A'(h-t)}C_1' \int_0^t C_1 e^{A(t-s)}B_1B_1'e^{-A's} ds\, dt\right) \\ &= \operatorname{tr}\big(\Delta_{22}'\Delta_{13}\big), \end{aligned}$$

where the latter equality follows from the fact that the second trace above is just the trace of the impulse response of a continuous-time system with the following transfer matrix:

$$\left[\begin{array}{c|c} -A' & C' \\ \hline \Delta_{22}' & 0 \end{array}\right]\left[\begin{array}{c|c} A & B \\ \hline C & 0 \end{array}\right]\left[\begin{array}{c|c} -A' & I \\ \hline B' & 0 \end{array}\right]$$

at $t = h$. This completes the proof. □

6. Discussion. Theorems 3–5 give not only the discrete-time part $\bar{\mathcal{K}}_d$ of the sampled-data controller, but also the H^∞ suboptimal (or the H^2 optimal) sampling and/or hold functions[2]. The purpose of the present section is to present some generic properties of the (sub)optimal sampler and hold functions, together with their interpretation. The first property is the separation between the design of the sampler and the hold, which has already been discussed in section 4.3. The separation property has a clear explanation: both sampler and hold are, in a sense, open-loop devices. Consequently, the design of the hold does not depend on the measurement $y(t)$ and the design of the sampler does not depend on the control action $u(t)$. The remaining properties refer to the $\mathcal{S}_h$ and $\mathcal{H}_h$, which are discussed separately below.

6.1. Optimal hold. Consider the H^∞ control CARE

$$0 = A'X + XA + C_1'C_1 + \tfrac{1}{\gamma^2}XB_1B_1'X - (B_2'X + D_{12}'C_1)'(D_{12}'D_{12})^{-1}(B_2'X + D_{12}'C_1), \tag{13}$$

which reduces to the H^2 one as $\gamma \to \infty$. The H^∞ state feedback gain is then

$$F_2 \doteq -(D_{12}'D_{12})^{-1}(B_2'X + D_{12}'C_1).$$

The following corollary can be derived from Theorems 3 and 4.

COROLLARY 1. *Let $X = X' \geq 0$ be the stabilizing solution to the CARE* (13). *Then one possible H^∞ suboptimal hold function is*

$$\phi_H(\tau) = F_2\, e^{(A+\gamma^{-2}B_1B_1'X+B_2F_2)\tau} \quad \forall \tau \in [0, h), \tag{14}$$

independent of the choice of the sampler and the measured output $y(t)$. When $\gamma^{-2} = 0$, this becomes the unique H^2 optimal solution.

The following remarks are in order.

Remark 6.1. Notice that the optimal hold for the *nonsingular* H^2 and H^∞ problems is always "asymptotically stable," in the sense that its "A" matrix is Hurwitz. In other words, the continuation of the hold in (14) over the whole interval $[0, \infty)$ belongs to L^2. The "stability" property is intuitively reasonable: the generalized hold being an "open-loop" device, it should prevent any source of instability during the intersample.

Remark 6.2. As a curious corollary to the previous remark, the standard zero-order hold (ZOH) can *never* appear as part of an optimal hold for nonsingular problems; this is because the "A" matrix of a ZOH is the zero-matrix. The ZOH does appear as part of an optimal hold for *singular* H^2 and H^∞ problems, for which $A + \gamma^{-2}B_1B_1'X + B_2F_2$ has an eigenvalue at the origin. Such problems may arise when integral control is required; see [32, section 17.3]. In the sampled-data case, the

[2]See Remark 2.1 for what is meant by *optimal* sampling and hold functions.

integral control action is actually "redistributed" between the discrete-time part of the controller $\bar{\mathcal{K}}_d$ and the hold $\mathcal{H}_h$.

Remark 6.3. It is worth mentioning that the design of H^2 and H^∞ (sub)optimal holds does *not* depend on the sampling period h. The optimal hold function remains the same for all h and only the interval changes on which $\phi_H(\tau)$ is defined. Also, it can easily be verified that with the optimal $\mathcal{H}_h$ the system is stabilizable *for all* h. In other words, if $\mathcal{H}_h$ is either H^2 or H^∞ optimal, then no sampling period can result in a loss of controllability and hence there exists no *pathological sampling*.

The stability property discussed in Remarks 6.1 and 6.2 becomes particularly interesting when compared with previous works on the design of generalized hold functions. Indeed, for most hold devices designed in the literature from the discrete-time performance point of view (see [1] and the references therein), the continuation of ϕ_H on the whole $\mathbb{R}^+$ does *not* necessarily belong to L^2 (and even L^∞). For example, if the hold function of the form $\phi_H(\tau) = C_H e^{A_H \tau} B_H$ is designed, then the typical choice is $A_H = -A'$, where A is the "A" matrix of a plant. Hence, the stability of such a hold depends on the open-loop plant dynamics. In the face of the previous remark, it is desirable to understand what is the underlying reason that rules out the use of "unstable" hold functions. To this end, consider the continuous-time LQR problem for the process

$$\dot{x}(t) = Ax(t) + B_2 u(t)$$

and the cost function

$$\mathcal{J} = \int_0^\infty \big(C_1 x(t) + D_{12} u(t)\big)' \big(C_1 x(t) + D_{12} u(t)\big) dt.$$

The solution to the problem of minimizing $\mathcal{J}$ is given by the feedback law [18] $u(t) = F_2 x(t)$, so that for a given t_k and any $\tau > 0$, the closed-loop state vector satisfies the equation

$$x(t_k + \tau) = e^{(A + B_2 F_2)\tau} x(t_k).$$

Substituting this equation into the formula for the control signal gives

$$u(t_k + \tau) = F_2 e^{(A + B_2 F_2)\tau} x(t_k).$$

On the other hand, it follows from (2a) and (14) that the H^2 optimal hold produces the control signal

$$u(kh + \tau) = F_2 e^{(A + B_2 F_2)\tau} \bar{u}_k \qquad \forall \tau \in [0, h).$$

The comparison of the latter two expressions prompts the following interpretation of the H^2 optimal hold.

INTERPRETATION 1. *The H^2 optimal hold, given by* (14) *subject to $\gamma^{-1} = 0$, attempts to "reconstruct" the LQR feedback control law, assuming that $\bar{\mathcal{K}}_d$ produces at the kth sampling instant an estimation of state vector of the plant at $t = kh$.*

Similarly to the H^2 case, the interpretation of the H^∞ suboptimal hold can be obtained from the continuous-time H^∞ state feedback problem as follows.

INTERPRETATION 2. *The H^∞ suboptimal hold, given by* (14)*, attempts to "reconstruct" the H^∞ suboptimal state feedback control law assuming that* (i) *$\bar{\mathcal{K}}_d$ produces at the kth sampling instant an estimate of the state vector of the plant at $t = kh$; and* (ii) *the disturbance w is the worst-case one (in an H^∞ sense), i.e., $w(t) = \frac{1}{\gamma^2} B_1' X\, x(t)$.*

These interpretations explain the "stability" of the optimal hold. Indeed, optimal continuous-time control must guarantee the internal stability of the closed-loop system. Consequently, the reconstruction of the continuous-time control signal in an open-loop fashion lead to the "stable dynamics" of the hold. When the integral control results from the continuous-time design, the "A" matrix of the hold function has an eigenvalue at the origin. This is consistent with the discussion in Remark 6.2.

The main conclusion of the previous interpretations is that the optimal hold attempts to reconstruct a "good" LTI continuous-time control law. Notice that this differs from some previous approaches to the design of $\mathcal{H}_h$, which seek to circumvent basic limitations of linear continuous-time control and attempt to outperform a continuous-time controller. The results in this paper suggest that the design of the hold should be based on the understanding that the optimal continuous-time control is the "best" possible choice; consequently, the sampled-data controller should mimic it as good as possible. We believe that this idea can be extended for designing controllers in problems with no known analytic solution but with computable optimal continuous-time controller. For example, the H^2 or H^∞ problems when $\mathcal{H}_h$ or $\mathcal{S}_h$ are constrained to be scalar.

6.2. Optimal sampler. Consider now the H^∞ filtering CARE

$$0 = AY + YA' + B_1B_1' + \tfrac{1}{\gamma^2}YC_1'C_1Y - (YC_2' + B_1D_{21}')(D_{21}D_{21}')^{-1}(YC_2' + B_1D_{21}')', \tag{15}$$

which again reduces to the H^2 one when $\gamma \to \infty$. The H^∞ filter gain is then as follows:

$$L_2 \doteq -(YC_2' + B_1D_{21}')(D_{21}D_{21}')^{-1}.$$

The following corollary can be derived from Theorems 3 and 5.

COROLLARY 2. *Let $Y = Y' \geq 0$ be the stabilizing solution of the CARE* (15). *Then one possible H^∞ suboptimal sampling function is*

$$\phi_S(\tau) = -e^{(A+\gamma^{-2}YC_1'C_1+L_2C_2)\tau}\,L_2 \quad \forall \tau \in [0,h), \tag{16}$$

independent of the choice of the hold and the control input $u(t)$. When $\gamma^{-2} = 0$, this becomes the unique H^2 optimal solution.

The following remarks are in order.

Remark 6.4. Analogously to the optimal hold case, the optimal sampling function (16) is asymptotically stable in the sense defined in Remark 6.1. Moreover, the design of $\mathcal{S}_h$ does not depend on the sampling period and it produces a detectable system for all h. In other words, if $\mathcal{S}_h$ is either H^2 or H^∞ optimal, then no sampling period can lead to the loss of observability and hence there exists no *pathological sampling*.

Remark 6.5. Another interesting property of the optimal sampler is that is does not require prefiltering of the measurement output $y(t)$ by an anti-aliasing filter. This is because, loosely speaking, the sampler itself serves as an anti-aliasing filter. Mathematically, this means that the generalized sampler $\mathcal{S}_h$ with sampling function (16) is bounded as an operator $L^2 \mapsto \ell^2$.

Remark 6.6. The anti-aliasing capability of the optimal sampler is actually a consequence of the fact that the ideal sampler (the sampling function $\phi_S(\tau) = \delta(\tau)$) can never appear as a part of an optimal sampler for *nonsingular* H^2 and H^∞ problems. The crucial assumption here is the full row rank of the matrix D_{21}. The latter

means that all of the measurement channels are corrupted by noise. For such a $y(t)$ the instantaneous sampling is an illegal operation that accounts for the absence of the impulse component in the optimal sampling function. In the case when an anti-aliasing filter is present, the matrix D_{21} necessarily loses row rank and the problem becomes singular. It can be shown that in such a case the optimal sampler is of the form $D_S\delta(\tau) + C_S e^{A_S\tau}$ for some A_S, C_S, and $D_S \neq 0$ (but still, $D_S D_{21} = 0$). This result gives rise to an interesting observation. It is well known that if $\mathcal{S}_h$ contains the ideal sampler, then the measured output must be prefiltered by an anti-aliasing filter. The discussion above implies that the opposite is also true: *whenever the measured output is prefiltered by an anti-aliasing filter, an optimal sampler contains the ideal sampler.*

7. Concluding remarks. This paper proposes a framework for treating a rather general class of sampled-data control problems. The cases considered were those in which the sampler, the hold, or both the sampler and the hold are available for design. For all these cases, necessary and sufficient conditions for the existence of controllers as well as state–space formulas for all the blocks involved have been provided. Conditions and the formulas for state–space matrices can be readily implemented in a computer.

The essence of this framework is to convert the hybrid periodically time-varying problems to a unified discrete-time LTI "standard problem" via the lifting transformation. The solution of the problem is then found directly in the lifted domain. The lifting procedure is then carried one step forward by peeling off the representations in the lifted domain back to continuous, or more specifically sampled-data, time. The sampled-data problems completely solved in this paper are the following:

- The sampled-data H^2 problems when sampling and/or hold functions are the design parameters. The solution to this problem is, to the best of our knowledge, completely new.
- The sampled-data H^∞ problems under similar conditions. These solutions, although not new, are considerably more transparent than the ones previously existing in the literature. In exchange for this transparency, we have found out that: (a) a separation structure between the design of H^∞ suboptimal sampler and hold; and (b) a solution to the H^∞ problem which, to the best of our knowledge, is the first closed-form solution.

In the final section, several comments which reconcile and compare the present result with some previous approaches to the design of generalized hold functions are discussed.

This paper is complemented with [23], where all the technical results required to establish the main results in this paper are developed. The reader is therefore referred to this work for details and interesting additional results concerning sampled-data systems in the lifted domain.

REFERENCES

[1] M. Araki, *Recent developments in digital control theory*, in Proceedings of 12th IFAC Word Congress, vol. IX, Sydney, Australia, 1993, pp. 251–260.

[2] B. Bamieh, *Optimal samplers and optimal hold functions in sampled-data problems*, in Proceedings of 3rd IEEE Mediterranean Symposium on New Directions in Control and Automation, vol. 2, 1995, Limassol, Cyprus, pp. 63–68.

[3] B. Bamieh and J. B. Pearson, *A general framework for linear periodic systems with applications to H^∞ sampled-data control*, IEEE Trans. Automat. Control, 37 (1992), pp. 418–435.

[4] B. Bamieh and J. B. Pearson, *The H^2 problem for sampled-data systems*, Systems Control Lett., 19 (1992), pp. 1–12.

[5] B. Bamieh, J. B. Pearson, B. A. Francis, and A. Tannenbaum, *A lifting technique for linear periodic systems with applications to sampled-data control*, Systems Control Lett., 17 (1991), pp. 79–88.

[6] T. Başar, *Optimum H^∞ designs under sampled state measurements*, Systems Control Lett., 16 (1991), pp. 399–409.

[7] J. H. Braslavsky, *Frequency Domain Analysis of Sampled-data Control Systems*, Ph.D. thesis, Department of Electrical and Computer Eng., The University of Newcastle, Newcastle, Australia, 1995.

[8] T. Chen and B. A. Francis, *Optimal Sampled-Data Control Systems*, Springer-Verlag, London, 1995.

[9] P. Colaneri, *Hamiltonian matrices for lifted systems and periodic Riccati equations in H^2/H^∞ analysis and control*, in Proceedings of the 30th IEEE Conference on Decision and Control, vol. 2, Brighton, UK, 1991, pp. 1914–1919.

[10] A. Feuer and G. C. Goodwin, *Generalized sample hold function: Frequency domain analysis of robustness, sensitivity and intersample difficulties*, IEEE Trans. Automat. Control, 39 (1994), pp. 1042–1045.

[11] G. Gu, J. Chen, and O. Toker, *Computation of $L^2[0,h]$ induced norms*, in Proceedings of the 35th IEEE Conference on Decision and Control, Kobe, Japan, 1996, pp. 4046–4051.

[12] E. S. Hamby, Y.-C. Juan, and P. T. Kabamba, *Optimal hold functions for digital control systems*, in Control and Dynamic Systems, C. T. Leondes, ed., vol. 79, Academic Press, 1996, pp. 1–50.

[13] V. Ionescu and M. Weiss, *Continuous and discrete-time Riccati theory: A Popov-function approach*, Linear Algebra and its Applications, 193 (1993), pp. 173–209.

[14] Y.-C. Juan and P. T. Kabamba, *Optimal hold functions for sampled data regulation*, Automatica, 27 (1991), pp. 177–181.

[15] P. T. Kabamba, *Control of linear systems using generalized sampled-data hold functions*, IEEE Trans. Automat. Control, 32 (1987), pp. 772–783.

[16] P. T. Kabamba and S. Hara, *Worst-case analysis and design of sampled-data control systems*, IEEE Trans. Automat. Control, 38 (1993), pp. 1337–1357.

[17] P. P. Khargonekar and N. Sivashankar, *H^2 optimal control for sampled-data systems*, Systems Control Lett., 17 (1991), pp. 425–436.

[18] H. Kwakernaak and R. Sivan, *Linear Optimal Control Systems*, Wiley & Sons, NY, 1972.

[19] L. Mirkin, *On discrete-time H^∞ problem with a strictly proper controller*, Internat. J. Control, 66 (1997), pp. 747–765.

[20] L. Mirkin and Z. J. Palmor, *Mixed discrete/continuous specifications in sampled-data H^2-optimal control*, Automatica, 33 (1997), pp. 1997–2014.

[21] L. Mirkin and Z. J. Palmor, *A new representation of the parameters of lifted systems*, IEEE Trans. Automat. Control, 44 (1999), pp. 833–840.

[22] L. Mirkin and H. Rotstein, *On the characterization of sampled-data controllers in the lifted domain*, Systems Control Lett., 29 (1997), pp. 269–277.

[23] L. Mirkin, H. Rotstein, and Z. J. Palmor, *H^2 and H^∞ design of sampled-data systems using lifting. Part* II*: Properties of systems in the lifted domain*, SIAM J. Control Optim., 38 (1999), pp. 197–218.

[24] H. P. Rotstein, *Constrained H^∞-Optimization for Discrete-Time Control*, Ph.D. thesis, California Institute of Technology, Pasadena, CA, 1992.

[25] N. Sivashankar and P. P. Khargonekar, *Characterization of the L^2-induced norm for linear systems with jumps with applications to sampled-data systems*, SIAM J. Control Optim., 32 (1994), pp. 1128–1150.

[26] W. Sun, K. M. Nagpal, and P. P. Khargonekar, *H^∞ control and filtering for sampled-data systems*, IEEE Trans. Automat. Control, 38 (1993), pp. 1162–1174.

[27] G. Tadmor, *H^∞ optimal sampled-data control in continuous-time systems*, Internat. J. Control, 56 (1992), pp. 99–141.

[28] H. T. Toivonen, *Sampled-data control of continuous-time systems with an H^∞ optimality criterion*, Automatica, 28 (1992), pp. 45–54.

[29] H. T. Toivonen, *Digital control with H^∞ optimality criteria*, in Control and Dynamic Systems, C. T. Leondes, ed., Academic Press, 71 (1995), pp. 215–262.

[30] H. L. Trentelman and A. A. Stoorvogel, *Sampled-data and discrete-time H^2 optimal control*, SIAM J. Control Optim., 33 (1995), pp. 834–862.

[31] Y. Yamamoto, *A function space approach to sampled data control systems and tracking problems*, IEEE Trans. Automat. Control, 39 (1994), pp. 703–713.

[32] K. Zhou, J. C. Doyle, and K. Glover, *Robust and Optimal Control*, Prentice-Hall, Englewood Cliffs, NJ, 1995.

SIAM J. CONTROL OPTIM.
Vol. 38, No. 1, pp. 197–218

H^2 AND H^∞ DESIGN OF SAMPLED-DATA SYSTEMS USING LIFTING. PART II: PROPERTIES OF SYSTEMS IN THE LIFTED DOMAIN*

LEONID MIRKIN†, HÉCTOR P. ROTSTEIN‡, AND ZALMAN J. PALMOR†

Abstract. An important class of sampled-data systems becomes time-invariant when the systems are "lifted" into an appropriate domain. The purpose of the present paper is to study some basic properties of systems represented by state space equations in the lifted domain. Hidden modes, invariant zeros, and Riccati equations are investigated, and the connection between representations in continuous and lifted domains are clarified. The material treated here is important for solving optimal control problems of generalized sampled-data systems.

Key words. sampled-data control, lifting technique, generalized sampler and hold

AMS subject classifications. 93C57, 93C25, 93B05, 93B07, 93B60

PII. S0363012997329615

1. Introduction.

In the companion paper [12], the authors presented a solution to the H^2 and H^∞ problems for a general class of sampled-data systems, including the cases of sampling and/or the hold functions available for design. The steps involved in the solution were the following: (i) lift the problem to the "discrete-time" lifted domain; (ii) find a formal solution to the resulting discrete-time problem; and (iii) "peel-off" the result back to continuous time. Although this approach is conceptually straightforward, it presents some technical difficulties, particularly when attempting to perform the calculations involved in step (iii). For this reason, the approach has rarely been followed in the literature, with the exception of [2] and [9], where sampled-data H^∞ analysis and H^2 design problems were solved, respectively, both for the case of fixed sampler and hold. Also note that Tadmor in [15, section 4] performed the H^∞ design directly for the lifted system, yet his solution was left in the lifted domain and no hint was given of how to "peel-off" the resulting solution.

One of the reasons that make the third step problematic is numerous properties that are well known for continuous and discrete-time systems have not been worked out for the representation of continuous-time systems in the lifted domain. It is worth mentioning that these properties are not directly inherited from the continuous-time representations and that the connection between a continuous-time representation and its lifted counterpart has not completely been clarified. The purpose of this paper is to bridge this gap.

Let $\mathcal{G}$, $\mathcal{H}_h$, and $\mathcal{W}_h$ denote a linear time-invariant (LTI) continuous-time system, a generalized hold with period h and the so-called LTI hold function of the form $\phi_H(\tau) = D_H\delta(\tau) + C_H e^{A_H\tau}B_H$, and the lifting operator, respectively (see [12]). This paper focuses on the LTI lifted systems $\check{\mathcal{G}} \doteq \mathcal{W}_h\mathcal{G}\mathcal{W}_h^{-1}$ and $\dot{\mathcal{G}} \doteq \mathcal{W}_h\mathcal{G}\mathcal{H}_h$; the aim is to characterize the hidden modes and invariant zeros of the state-space realizations obtained when lifting the continuous-time representations and to study the discrete

*Received by the editors November 5, 1997; accepted for publication (in revised form) December 14, 1998; published electronically December 15, 1999.
http://www.siam.org/journals/sicon/38-1/32961.html
†Faculty of Mechanical Engineering, Technion—Israel Institute of Technology, Technion City, Haifa 32000, Israel (mersglm@tx.technion.ac.il, merzpal@tx.technion.ac.il).
‡RAFAEL and Faculty of Electrical Engineering, Technion—Israel Institute of Technology, Technion City, Haifa 32000, Israel (hector@ee.technion.ac.il).

algebraic Riccati equations (DAREs) associated with these systems. Corresponding results for systems of the form $\acute{\mathcal{G}} \doteq \mathcal{S}_h \mathcal{G} \mathcal{W}_h^{-1}$, where $\mathcal{S}_h$ is a generalized sampler, follow from those for $\grave{\mathcal{G}}$ by duality arguments. (To this end, $\mathcal{H}_h$ includes the impulse action $D_H \delta(\tau)$; this enables one to cover corresponding results for the ideal sampler.)

Since the systems $\mathcal{G}$ and $\breve{\mathcal{G}}$ are equivalent in an input-output sense [1], one would expect a close connection between the properties of their state-space realizations. However, establishing these connections is far from straightforward, since the properties cannot always be characterized adequately in terms of input-output relations. To illustrate the difficulties, recall that [10] establishes that a DARE associated with $\breve{\mathcal{G}}$ may have more solutions than the corresponding continuous-time algebraic Riccati equation (CARE) associated with $\mathcal{G}$. This result suggests that the connections between state-space properties of continuous-time LTI systems and their lifted counterparts are indeed worth study. In spite of this remark, this paper shows that a strong connection between state-space properties of $\mathcal{G}$ and $\breve{\mathcal{G}}$ does exist, much like the connection between their poles. In particular, it will be shown that any hidden mode (invariant zero) λ_i of $\mathcal{G}$ corresponds to the hidden mode (invariant zero) $e^{\lambda_i h}$ of $\breve{\mathcal{G}}$, even when their state-space realizations are not minimal.

Contrary to $\breve{\mathcal{G}}$, the operator $\grave{\mathcal{G}}$ has a finite rank. Consequently, hidden modes and invariant zeros of $\grave{\mathcal{G}}$ can, in principle, be routinely characterized in terms of matrix-valued parameters of the *finite-dimensional* operator $\grave{\mathcal{G}}^* \grave{\mathcal{G}}$ [3]. Such an approach fits well into the H^2 problem, since computations are based on precisely the same matrices. Unfortunately, the same is not true for the sampled-data H^∞ problem. For that reason, in this paper a new characterization of singularities of $\grave{\mathcal{G}}$ is proposed in terms of the H^∞ data. Although the derivations in this case are more involved, the final formulae are not more complicated than those based on the H^2 data. As it turns out, Riccati equations play an important role in the characterization of certain properties of systems in the lifted-domain, and consequently they are also investigated. It will be shown that the treatment of DAREs in terms of extended symplectic pairs (ESPs) associated with a *lifted* system enables one to obtain surprisingly simple solutions. In particular, the H^∞ DARE associated with $\grave{\mathcal{G}}$ reduces to ESP, which is not more complicated than the ESP associated with the H^2 DARE.

The material considered here was originally intended for solving the optimization problems presented in [12]; however, the results have independent interest since numerous important properties of systems in the lifted-domain are studied. The reader is referred to [12] for an introduction to lifting and additional motivation.

This paper consists of four sections and an appendix. Section 2 reviews the representation for lifted systems introduced in [11]. Section 3 studies a continuous-time system in the lifted domain, including the connection between the continuous-time and the lifted domain descriptions. Section 4 studies the cascade of a continuous time system and a generalized hold, also in the lifted-domain. Section 5 discusses the meaning of the coupling condition for H^∞. When studying the properties of systems in the lifted domain, one is naturally led to considering the DARE. For completeness, an appendix has been included, which reviews relevant results collected from various sources.

1.1. Notation. The notation throughout the paper is consistent with that of [12], with the following additions. The Redheffer star product [13] of two suitably partitioned 2×2 block operators O and P is denoted as

$$O \star P \doteq \begin{bmatrix} \mathcal{F}_\ell(O, P_{22}) & O_{12}(I - P_{22}O_{22})^{-1}P_{21} \\ P_{12}(I - O_{22}P_{22})^{-1}O_{21} & \mathcal{F}_\ell(P, O_{22}) \end{bmatrix}.$$

Continuous, discrete, or lifted LTI system $\tilde{\mathcal{G}}$ will be represented by the transfer function

$$\tilde{G}(\cdot) = \left[\begin{array}{c|c} \tilde{A} & \tilde{B} \\ \hline \tilde{C} & \tilde{D} \end{array}\right],$$

such that $\tilde{A} \in \mathbb{R}^{n\times n}$, associated with a state-space realization. This realization is in turn associated with the following pencils: the *controllability pencil*

$$\boldsymbol{C}_{\tilde{\mathcal{G}}}(\lambda) \doteq \left[\begin{array}{cc} \tilde{A} - \lambda I & \tilde{B} \end{array}\right]$$

and the *system pencil*

$$\boldsymbol{S}_{\tilde{\mathcal{G}}}(\lambda) \doteq \left[\begin{array}{cc} \tilde{A} - \lambda I & \tilde{B} \\ \tilde{C} & \tilde{D} \end{array}\right].$$

As is conventional, $z \in \mathbb{C}$ is said to be (a) an uncontrollable mode of $\tilde{A}$ if $\operatorname{rank} \boldsymbol{C}_{\tilde{\mathcal{G}}}(z) < n$ and (b) an invariant zero of the realization of $\tilde{\mathcal{G}}$ if $\operatorname{rank} \boldsymbol{S}_{\tilde{\mathcal{G}}}(z) < \operatorname{normalrank} \boldsymbol{S}_{\tilde{\mathcal{G}}}(\lambda)$. Finally, the special matrices $E_1 \doteq \left[\begin{smallmatrix} I \\ 0 \end{smallmatrix}\right]$, $E_2 \doteq \left[\begin{smallmatrix} 0 \\ I \end{smallmatrix}\right]$, $E_{11} \doteq \left[\begin{smallmatrix} I & 0 \\ 0 & 0 \end{smallmatrix}\right]$, and $E_{22} \doteq \left[\begin{smallmatrix} 0 & 0 \\ 0 & I \end{smallmatrix}\right]$ are used. (The block dimensions are always clear from the context.)

2. A representation of lifted systems. This section presents a brief exposition of the results of [11] concerning the representation of the parameters of lifted systems. This representation plays a central role in the reasonings that follow. Conventionally [17, 2, 1, 3], the parameters of lifted systems are described by integral operators over $L^2[0,h]$. This representation follows naturally from the lifting procedure but unfortunately makes manipulations quite cumbersome. To remedy this difficulty, [11] considered a different representation of the parameters of the lifted systems, which considerably simplifies manipulations over these parameters. This representation builds on three components: systems with two-point boundary conditions (STPBC) operating on the time interval $[0,h]$, the impulse operator $\mathcal{I}_\theta$, and the sampling operator $\mathcal{I}_\theta^*$.

(i) STPBC are linear continuous-time operators $\breve{O} : L^2[0,h] \mapsto L^2[0,h]$, described by the state equations [7, 4]:

$$\breve{O} : \left\{\begin{array}{ll} \dot{x}(t) = Ax(t) + B\omega(t), & \Omega x(0) + \Upsilon x(h) = 0, \\ \zeta(t) = Cx(t) + D\omega(t), & \end{array}\right. \tag{2.1}$$

where the square matrices Ω and Υ shape the boundary conditions of the state vector x. The boundary conditions are said to be well-posed if $\det(\Omega + \Upsilon e^{Ah}) \neq 0$ and in this case the map $\zeta = \breve{O}\omega$ is well defined $\forall \omega \in L^2[0,h]$, namely,

$$\zeta(t) = D\omega(t) + Ce^{At}(\Omega + \Upsilon e^{Ah})^{-1}\left(\Omega \int_0^t e^{-As}B\omega(s)ds - \Upsilon \int_t^h e^{A(h-s)}B\omega(s)ds\right).$$

In this paper, STPBC are denoted by using the compact block notation:

$$\breve{O} = \left(\begin{array}{c|c|c} A & \boxed{\Omega \rightleftharpoons \Upsilon} & B \\ \hline C & & D \end{array}\right).$$

The term STPBC is reserved for systems with well-posed boundary conditions only. In the case when $\Omega = I$ and $\Upsilon = 0$, the boundary condition window can be omitted so that the notation becomes $\left(\begin{array}{c|c} A & B \\ \hline C & D \end{array}\right)$. Notice that this case corresponds to a causal STPBC.

(ii) The impulse operator $\mathcal{I}_\theta$ transforms a vector $\eta \in \mathbb{R}^n$ into a modulated δ-impulse:

$$\big(\mathcal{I}_\theta \eta\big)(t) = \delta(t-\theta)\eta.$$

(iii) The sampling operator[1] $\mathcal{I}_\theta^*$ transforms a function $\zeta \in C_n[0,h]$ into a vector from $\mathbb{R}^n$:

$$\mathcal{I}_\theta^* \zeta = \zeta(\theta).$$

With the aid of these operators, we now consider the representation of basic components of sampled-data systems in the lifted domain. Let $\mathcal{G}$ be an LTI continuous-time system with the transfer matrix

$$G(s) = \left[\begin{array}{c|c} A & B \\ \hline C & D \end{array}\right], \tag{2.2}$$

$\mathcal{H}_h$ be a generalized zero-order hold of the form [12, eq. 2a] with the hold function

$$\phi_H(\tau) = D_H \delta(\tau) + C_H e^{A_H \tau} B_H, \tag{2.3}$$

and $\mathcal{S}_h$ be a generalized zero-order sampler of the form [12, eq. 2b] with the sampling function

$$\phi_S(\tau) = D_S \delta(\tau) + C_S e^{A_S \tau} B_S. \tag{2.4}$$

The systems $\breve{\mathcal{G}} \doteq \mathcal{W}_h \mathcal{G} \mathcal{W}_h^{-1}$, $\grave{\mathcal{H}}_h \doteq \mathcal{W}_h \mathcal{H}_h$, and $\acute{\mathcal{S}}_h \doteq \mathcal{S}_h \mathcal{W}_h^{-1}$ are LTI with transfer functions

$$\breve{G}(z) = \left[\begin{array}{c|c} \bar{A} & \acute{B} \\ \hline \grave{C} & \breve{D} \end{array}\right],$$

$$\grave{H}_h(z) = \left[\begin{array}{c|c} 0 & 0 \\ \hline 0 & \grave{\Phi}_H \end{array}\right],$$

and

$$\acute{S}_h(z) = \left[\begin{array}{c|c} 0 & \acute{\Phi}_S \\ \hline I & 0 \end{array}\right],$$

respectively, where

$$\left[\begin{array}{cc} \bar{A} & \acute{B} \\ \grave{C} & \breve{D} \end{array}\right] = \left[\begin{array}{cc} \mathcal{I}_h^* & 0 \\ 0 & I \end{array}\right] \left(\begin{array}{c|cc} A & I & B \\ \hline I & 0 & 0 \\ C & 0 & D \end{array}\right) \left[\begin{array}{cc} \mathcal{I}_0 & 0 \\ 0 & I \end{array}\right],$$

$$\grave{\Phi}_H = \left(\begin{array}{c|c} A_H & B_H \\ \hline C_H & D_H \end{array}\right) \mathcal{I}_0,$$

and

$$\acute{\Phi}_S = \mathcal{I}_h^* \left(\begin{array}{c|c} A_S & B_S \\ \hline C_S & D_S \end{array}\right).$$

[1] It is worth stressing that $\mathcal{I}_\theta^*$ is not the adjoint of $\mathcal{I}_\theta$. See [11] for a justification of this abuse of notation.

The representation via STPBC has several advantages. First, algebraic manipulations over STPBC can be performed in state space, much like manipulations over standard LTI systems. This fact is established in the next proposition.

PROPOSITION 2.1 (see [11]). *Let*

$$\breve{O}_i = \left(\begin{array}{c|c|cc} A_i & \Omega_i \rightleftharpoons \Upsilon_i & B_{i_1} & B_{i_2} \\ \hline C_{i_1} & & D_{i_{11}} & D_{i_{12}} \\ C_{i_2} & & D_{i_{21}} & D_{i_{22}} \end{array}\right), \qquad i = 1, 2.$$

Then

$$\breve{O}_1 \star \breve{O}_2 = \left(\begin{array}{c|c|c} A_\star & \Omega_\star \rightleftharpoons \Upsilon_\star & B_\star \\ \hline C_\star & & D_\star \end{array}\right),$$

where

$$\begin{aligned}
A_\star &\doteq \begin{bmatrix} A_1 & 0 \\ B_{2_2}C_{1_2} & A_2 \end{bmatrix} + \begin{bmatrix} B_{1_2} \\ B_{2_2}D_{1_{22}} \end{bmatrix}(I - D_{2_{22}}D_{1_{22}})^{-1}\begin{bmatrix} D_{2_{22}}C_{1_2} & C_{2_2} \end{bmatrix}, \\
B_\star &\doteq \begin{bmatrix} B_{1_1} & 0 \\ B_{2_2}D_{1_{21}} & B_{2_1} \end{bmatrix} + \begin{bmatrix} B_{1_2} \\ B_{2_2}D_{1_{22}} \end{bmatrix}(I - D_{2_{22}}D_{1_{22}})^{-1}\begin{bmatrix} D_{2_{22}}D_{1_{21}} & D_{2_{21}} \end{bmatrix}, \\
C_\star &\doteq \begin{bmatrix} C_{1_1} & 0 \\ D_{2_{12}}C_{1_2} & C_{2_1} \end{bmatrix} + \begin{bmatrix} D_{1_{12}} \\ D_{2_{12}}D_{1_{22}} \end{bmatrix}(I - D_{2_{22}}D_{1_{22}})^{-1}\begin{bmatrix} D_{2_{22}}C_{1_2} & C_{2_2} \end{bmatrix}, \\
D_\star &\doteq \begin{bmatrix} D_{1_{11}} & 0 \\ D_{2_{12}}D_{1_{21}} & D_{2_{11}} \end{bmatrix} + \begin{bmatrix} D_{1_{12}} \\ D_{2_{12}}D_{1_{22}} \end{bmatrix}(I - D_{2_{22}}D_{1_{22}})^{-1}\begin{bmatrix} D_{2_{22}}D_{1_{21}} & D_{2_{21}} \end{bmatrix}, \\
\Omega_\star &\doteq \begin{bmatrix} \Omega_1 & 0 \\ 0 & \Omega_2 \end{bmatrix}, \\
\Upsilon_\star &\doteq \begin{bmatrix} \Upsilon_1 & 0 \\ 0 & \Upsilon_2 \end{bmatrix},
\end{aligned}$$

and the star product exists iff $\det(I - D_{2,22}D_{1,22}) \neq 0$ *and* $\det(\Omega_\star + \Upsilon_\star e^{A_\star h}) \neq 0$.

REMARK 2.1. *Since addition, multiplication, inversion, and LFT operations are just special cases of the star product, Proposition* 2.1 *covers a wide spectrum of manipulations over STPBC. In particular, it follows from Proposition* 2.1 *that* $\breve{O}_1 + \breve{O}_2$ *and* $\breve{O}_1\breve{O}_2$ *are well defined for every* $\breve{O}_1$ *and* $\breve{O}_2$ *of appropriate dimensions, and a square STPBC given by* (2.1) *is invertible iff* $\det(D) \neq 0$ *and* $\det(\Omega + \Upsilon e^{(A - BD^{-1}C)h}) \neq 0$.

Second, the operators $\mathcal{I}_\theta$ and $\mathcal{I}_\theta^*$ fit nicely into the state-space framework. In particular, the adjoint of any composition of STPBC, $\mathcal{I}_\theta$ and $\mathcal{I}_\theta^*$ can be computed componentwise. For instance,

$$\begin{bmatrix} \bar{A} & \acute{B} \\ \grave{C} & \breve{D} \end{bmatrix}^* = \begin{bmatrix} \mathcal{I}_0^* & 0 \\ 0 & I \end{bmatrix}\left(\begin{array}{c|c|cc} -A' & 0 \rightleftharpoons I & I & C' \\ \hline -I & & 0 & 0 \\ -B' & & 0 & D' \end{array}\right)\begin{bmatrix} \mathcal{I}_h & 0 \\ 0 & I \end{bmatrix}.$$

Moreover, the operators $\mathcal{I}_\theta$ and $\mathcal{I}_\theta^*$ can be absorbed into STPBC in an elegant manner.

PROPOSITION 2.2 (see [11]). *Let*

$$\begin{bmatrix} \acute{B}_i \\ \breve{D}_i \end{bmatrix} \doteq \begin{bmatrix} \mathcal{I}_h^* & 0 \\ 0 & I \end{bmatrix}\left(\begin{array}{c|c} A & B_i \\ \hline I & 0 \\ C & D_i \end{array}\right), \qquad i = 1, 2.$$

Then for any appropriately dimensioned matrix M

$$\breve{D}_1^*\breve{D}_2 + \acute{B}_1^* M \acute{B}_2 = \left(\begin{array}{cc|c|c} A & 0 & \left[\begin{smallmatrix} I & 0 \\ 0 & 0 \end{smallmatrix}\right] \rightleftharpoons \left[\begin{smallmatrix} 0 & 0 \\ -M & I \end{smallmatrix}\right] & B_2 \\ -C'C & -A' & & -C'D_2 \\ \hline D_1'C & B_1' & & D_1'D_2 \end{array}\right).$$

Finally, the computations of matrices involving infinite-dimensional parameters of lifted systems can be reduced to the computation of matrix exponentials.

PROPOSITION 2.3 (see [11]).

$$\left[\begin{smallmatrix} \mathcal{I}_h^* \\ \mathcal{I}_0^* \end{smallmatrix}\right]\left(\begin{array}{c|c|c} A & \Omega \rightleftharpoons \Upsilon & B \\ \hline C & & 0 \end{array}\right)\left[\begin{smallmatrix} \mathcal{I}_0 & \mathcal{I}_h \end{smallmatrix}\right] = \left[\begin{smallmatrix} Ce^{Ah} \\ C \end{smallmatrix}\right](\Omega + \Upsilon e^{Ah})^{-1}\left[\begin{smallmatrix} \Omega B & -\Upsilon B \end{smallmatrix}\right],$$

where $CB = 0$ is assumed for the $(2,1)$ and the $(1,2)$ entries to guarantee well-posedness.

3. System of the form $\mathcal{W}_h\mathcal{G}\mathcal{W}_h^{-1}$. The system $\breve{\mathcal{G}} \doteq \mathcal{W}_h\mathcal{G}\mathcal{W}_h^{-1}$ has a natural state space representation inherited from the state space representation of $\mathcal{G}$. The purpose of this section is to study the relationship between the uncontrollable modes and invariant zeros of the two representations. It is clear that such a relationship must exist. After all, $\mathcal{G}$ and $\breve{\mathcal{G}}$ are equivalent in an *input-output* sense, and the state vector of $\breve{\mathcal{G}}$ is the sampled state vector of $\mathcal{G}$. However, as shown next, some subtleties are involved and the topic must be investigated with care. For the sake of completeness, a result from [10] is also reviewed, stating that an H^∞ DARE associated with $\breve{\mathcal{G}}$ and an H^∞ CARE associated with $\mathcal{G}$ are actually equivalent.

3.1. Preliminary results. For given matrices A_a, B_a, and C_a of appropriate dimensions and a scalar h form the Hamiltonian matrix

$$H_a \doteq \left[\begin{smallmatrix} A_a & -B_aB_a' \\ -C_a'C_a & -A_a' \end{smallmatrix}\right] = \left[\begin{smallmatrix} H_{a,11} & H_{a,12} \\ H_{a,21} & H_{a,22} \end{smallmatrix}\right]$$

and the symplectic matrix

$$S_a \doteq e^{H_a h} = \left[\begin{smallmatrix} S_{a,11} & S_{a,12} \\ S_{a,21} & S_{a,22} \end{smallmatrix}\right].$$

The following two propositions will be useful later.

PROPOSITION 3.1. *The block $S_{a,22}$ verifies* $\det S_{a,22} \neq 0$.

Proof. Consider the STPBC

$$\breve{O}_a \doteq \left(\begin{array}{c|c} A_a & B_a \\ \hline C_a & 0 \end{array}\right).$$

From here, the operator $I + \breve{O}_a\breve{O}_a^*$ has the representation

$$I + \breve{O}_a\breve{O}_a^* = \left(\begin{array}{cc|c|c} A_a & -B_aB_a' & \left[\begin{smallmatrix} I & 0 \\ 0 & 0 \end{smallmatrix}\right] \rightleftharpoons \left[\begin{smallmatrix} 0 & 0 \\ 0 & I \end{smallmatrix}\right] & 0 \\ 0 & -A_a' & & C_a' \\ \hline C_a & 0 & & I \end{array}\right),$$

and the STPBC in the right-hand side above is invertible iff the matrix

$$E_{11} + E_{22}\exp\left(\left(\left[\begin{smallmatrix} A_a & -B_aB_a' \\ 0 & -A_a' \end{smallmatrix}\right] - \left[\begin{smallmatrix} 0 \\ C_a' \end{smallmatrix}\right]\left[\begin{smallmatrix} C_a & 0 \end{smallmatrix}\right]\right)h\right) = \left[\begin{smallmatrix} I & 0 \\ S_{a,21} & S_{a,22} \end{smallmatrix}\right]$$

is nonsingular. On the other hand, the operator $I+\breve{O}_a\breve{O}_a^*$ is invertible by construction. This concludes the proof. □

PROPOSITION 3.2. *The following two statements are equivalent:*

(i) *The pencil* $\left[\begin{smallmatrix} H_{a,11}-sI \\ H_{a,21}\end{smallmatrix}\right]$ *has reduced column rank at* $s=\lambda+j\frac{2\pi}{h}k$ *for some* $k\in\mathbb{Z}$.

(ii) *The pencil* $\left[\begin{smallmatrix} S_{a,11}-zI \\ S_{a,21}\end{smallmatrix}\right]$ *has reduced column rank at* $z=e^{\lambda h}$.

Proof. Let $P_H(s)$ and $P_S(z)$ denote the pencils in (i) and (ii), respectively.

(i)⇒(ii): Assume that $P_H(\lambda)$ has reduced column rank. Then there exists a vector $\eta\neq 0$ so that $P_H(\lambda)\eta=0$. It means that $E_1\eta$ is the eigenvector of H_a associated with the eigenvalue λ. Hence, $E_1\eta$ is also the eigenvector of $e^{H_a h}$ associated with the eigenvalue $e^{\lambda h}$, which in turn leads to $P_S(e^{\lambda h})\eta=0$.

(ii) ⇒ (i): Assume that $P_S(e^{\lambda h})$ has reduced column rank. Then there exists a vector $\eta\neq 0$ so that $P_S(e^{\lambda h})\eta=0$. Since an eigenvector of $e^{H_a h}$ is not necessarily an eigenvector of H_a [5, §2.11], one cannot now apply the same reasoning as in the first part of the proof. Instead, the STPBC arguments are used below. To this end, we introduce the following operators:

$$\left[\begin{array}{cc}\grave{C}_a & \breve{D}_a\end{array}\right] \doteq \left(\begin{array}{c|cc} A_a & I & B_a \\ \hline C_a & 0 & 0\end{array}\right)\left[\begin{array}{cc}\mathcal{I}_0 & 0\\ 0 & I\end{array}\right].$$

Then

$$\begin{aligned}\grave{C}_a^*(I+\breve{D}_a\breve{D}_a^*)^{-1}\grave{C}_a &= \left[\begin{array}{cc} I & 0\end{array}\right]\left(\left[\begin{array}{cc}0 & \grave{C}_a^*\\ 0 & -\breve{D}_a^*\end{array}\right]\star\left[\begin{array}{cc}0 & 0\\ \grave{C}_a & \breve{D}_a\end{array}\right]\right)\left[\begin{array}{c}0\\ I\end{array}\right]\\ &= \mathcal{I}_0^*\left(\begin{array}{c|c|c} H_a & E_{11}\rightleftharpoons E_{22} & E_1\\ \hline E_2' & & 0\end{array}\right)\mathcal{I}_0 \qquad \text{(by Proposition 2.1)}\\ &= -S_{a,22}^{-1}S_{a,21} \qquad \text{(by Proposition 2.3)}\end{aligned}$$

and hence $\ker S_{a,21}=\ker\grave{C}_a$. Since $\eta\in\ker S_{a,21}$, then $C_a e^{A_a\tau}\eta\equiv 0$. Therefore, η belongs to the unobservable subspace of the pair (C_a,A_a). Now, it can be shown (e.g., by using the Kalman canonical decomposition of (C_a,A_a)) that there exists a $k\in\mathbb{Z}$ so that $P_H(\lambda+j\frac{2\pi}{h}k)\eta=0$. □

Now, assume that $\det D'D\neq 0$, and associate with the systems $\mathcal{G}$ and $\breve{\mathcal{G}}$ the following Hamiltonian matrices:

$$H \doteq \left[\begin{array}{cc} A & 0\\ -C'C & -A'\end{array}\right]-\left[\begin{array}{c} B\\ -C'D\end{array}\right](D'D)^{-1}\left[\begin{array}{cc} D'C & B'\end{array}\right]=\left[\begin{array}{cc} H_{11} & H_{12}\\ H_{21} & H_{22}\end{array}\right] \tag{3.1a}$$

and

$$\bar{H} \doteq \left[\begin{array}{cc} \bar{A} & 0\\ -\grave{C}^*\grave{C} & -\bar{A}'\end{array}\right]-\left[\begin{array}{c} \acute{B}\\ -\grave{C}^*\breve{D}\end{array}\right](\breve{D}^*\breve{D})^{-1}\left[\begin{array}{cc} \breve{D}^*\grave{C} & \acute{B}^*\end{array}\right]=\left[\begin{array}{cc} \bar{H}_{11} & \bar{H}_{12}\\ \bar{H}_{21} & \bar{H}_{22}\end{array}\right]. \tag{3.1b}$$

The next proposition establishes a relationship between H and $\bar{H}$.

PROPOSITION 3.3. *If* $\det D'D\neq 0$, *then the matrix* $\bar{H}$ *is well defined and can be expressed as*

$$\bar{H}=\left[\begin{array}{cc} S_{11} & S_{12}\\ 0 & -I\end{array}\right]\left[\begin{array}{cc} I & 0\\ S_{21} & S_{22}\end{array}\right]^{-1}=\left[\begin{array}{cc} I & S_{12}\\ 0 & S_{22}\end{array}\right]^{-1}\left[\begin{array}{cc} S_{11} & 0\\ S_{21} & -I\end{array}\right],$$

where S_{ij} *are the subblocks of the matrix* $S\doteq e^{Hh}$.

Proof. Consider the 2×2 operator

$$\breve{O} \doteq \left(\begin{array}{c|cc} A & I & B \\ \hline I & 0 & 0 \\ \left[\begin{array}{c} C \\ 0 \end{array}\right] & \left[\begin{array}{c} 0 \\ 0 \end{array}\right] & \left[\begin{array}{c} D \\ I \end{array}\right] \end{array}\right).$$

Then it is straightforward to verify that

$$\begin{aligned} \bar{H} &= \left[\begin{array}{cc} \mathcal{I}_h^* & 0 \\ 0 & -\mathcal{I}_0^* \end{array}\right] \breve{O} \star \breve{O}^* \left[\begin{array}{cc} \mathcal{I}_0 & 0 \\ 0 & \mathcal{I}_h \end{array}\right] \\ &= \left[\begin{array}{cc} \mathcal{I}_h^* & 0 \\ 0 & -\mathcal{I}_0^* \end{array}\right] \left(\begin{array}{c|c|c} H & \boxed{E_{11} \rightleftharpoons E_{22}} & I \\ \hline I & & 0 \end{array}\right) \left[\begin{array}{cc} \mathcal{I}_0 & 0 \\ 0 & -\mathcal{I}_h \end{array}\right] \qquad \text{(by Proposition 2.1)} \end{aligned}$$

and the proof is now completed by applying Proposition 2.3. ∎

3.2. Uncontrollable modes. From the input-output equivalence, a strong correspondence exists between the poles of $\mathcal{G}$ and those of $\breve{\mathcal{G}}$. This is clearly exhibited by the fact that the "A" matrix of the realization of $\breve{\mathcal{G}}$ inherited from a corresponding one of $\mathcal{G}$ is $\bar{A} = e^{Ah}$. Consequently, $e^{\lambda h}$ is a pole of the realization of $\breve{\mathcal{G}}$ iff either of $\lambda + j\frac{2\pi}{h}k$, $k \in \mathbb{Z}$, is a pole of the realization of $\mathcal{G}$. The lemma below establishes that precisely the same relationship holds between the uncontrollable modes of the realizations of $\mathcal{G}$ and $\breve{\mathcal{G}}$.

LEMMA 3.4. *$\boldsymbol{C}_{\breve{\mathcal{G}}}(e^{\lambda h})$ is right invertible iff $\boldsymbol{C}_{\mathcal{G}}(\lambda + j\frac{2\pi}{h}k)$ is right invertible $\forall k \in \mathbb{Z}$.*

Proof. Without loss of generality, assume that $C = 0$ and $D = I$. Then in (3.1a) $H_{21} = S_{21} = 0$ and then, for any μ

$$\left[\begin{array}{cc} S_{11} - \mu I & S_{12} \end{array}\right] = \boldsymbol{C}_{\breve{\mathcal{G}}}(\mu) \left[\begin{array}{cc} I & 0 \\ 0 & \acute{B}^* \end{array}\right] \left[\begin{array}{cc} I & 0 \\ 0 & -S_{22} \end{array}\right]$$

by Proposition 3.3. Together with Proposition 3.1, this leads to the equality

$$\operatorname{rank} \boldsymbol{C}_{\breve{\mathcal{G}}}(\mu) = \operatorname{rank} \left[\begin{array}{cc} S_{11} - \mu I & S_{12} \end{array}\right].$$

The same arguments show that for any μ

$$\operatorname{rank} \boldsymbol{C}_{\mathcal{G}}(\mu) = \operatorname{rank} \left[\begin{array}{cc} H_{11} - \mu I & H_{12} \end{array}\right].$$

Thus, the lemma follows by applying Proposition 3.2 to $A_a = A'$, $B_a = 0$, and $C_a = B'$. ∎

The following two corollaries to Lemma 3.4 are of particular interest for the solutions of the sampled-data H^2 and H^∞ problems with free hold or sampler, respectively:

COROLLARY 3.5. *The pair $(\bar{A}, \acute{B})$ is $\mathbb{D}$-stabilizable iff the pair (A, B) is $\mathbb{C}^-$-stabilizable.*

COROLLARY 3.6. *The pair $(\grave{C}, \bar{A})$ is $\mathbb{D}$-detectable iff the pair (C, A) is $\mathbb{C}^-$-detectable.*

3.3. Invariant zeros. The relationship between the invariant zeros of the realizations of $\mathcal{G}$ and $\breve{\mathcal{G}}$ is given in the next lemma.

LEMMA 3.7. *$\boldsymbol{S}_{\breve{\mathcal{G}}}(e^{\lambda h})$ is left invertible iff $\det D'D \neq 0$ and $\boldsymbol{S}_{\mathcal{G}}(\lambda + j\frac{2\pi}{h}k)$ is left invertible $\forall k \in \mathbb{Z}$.*

Proof. First, note that the operator $\boldsymbol{S}_{\breve{\mathcal{G}}}(z)$ is left invertible only if $\breve{O}_D \doteq \breve{D}^*\breve{D} + \acute{B}^*\acute{B}$ is invertible on $L^2[0,h]$. It follows from Proposition 2.2 that

$$\breve{O}_D = \left(\begin{array}{cc|c|c} A & 0 & \left[\begin{smallmatrix} I & 0 \\ 0 & 0 \end{smallmatrix}\right] \rightleftharpoons \left[\begin{smallmatrix} 0 & 0 \\ -I & I \end{smallmatrix}\right] & B \\ -C'C & -A' & & -C'D \\ \hline D'C & B' & & D'D \end{array}\right),$$

from which the necessity of $D'D > 0$ for the left invertibility of $\boldsymbol{S}_{\breve{\mathcal{G}}}(z)$ is established.

If D is left invertible, then so is $\breve{D}$ and for any z the equation

$$\begin{bmatrix} \bar{H}_{11} - zI & ? \\ \bar{H}_{21} & ? \\ 0 & \breve{D}^*\breve{D} \end{bmatrix} = \begin{bmatrix} I & 0 \\ 0 & \grave{C}^* \\ 0 & \breve{D}^* \end{bmatrix} \boldsymbol{S}_{\breve{\mathcal{G}}}(z) \begin{bmatrix} I & 0 \\ -(\breve{D}^*\breve{D})^{-1}\breve{D}^*\grave{C} & I \end{bmatrix}$$

holds. Therefore, taking into account the relationship between $\bar{H}$ and S from Proposition 3.3,

$$\dim\ker \boldsymbol{S}_{\breve{\mathcal{G}}}(z) = \dim\ker \begin{bmatrix} \bar{H}_{11} - zI \\ \bar{H}_{21} \end{bmatrix} = \dim\ker \begin{bmatrix} S_{11} - zI \\ S_{21} \end{bmatrix}.$$

Analogously, for any $s \in \mathbb{C}$

$$\dim\ker \boldsymbol{S}_{\mathcal{G}}(s) = \dim\ker \begin{bmatrix} H_{11} - sI \\ H_{21} \end{bmatrix}.$$

The proof is now completed by applying Proposition 3.2 to matrices $A_a = A - B(D'D)^{-1}D'C$, $B_a = B(D'D)^{-1/2}$, and $C_a = (I - D(D'D)^{-1}D')^{1/2}C$. □

In the solution of the sampled-data H^2 and H^∞ problems one is concerned with invariant zeros on the unit circle only. Consequently, the following two corollaries of Lemma 3.7 are formulated.

COROLLARY 3.8. *$\boldsymbol{S}_{\breve{\mathcal{G}}}(e^{j\theta})$ is left invertible $\forall\theta \in [0, 2\pi]$ iff $\det D'D \neq 0$ and $\boldsymbol{S}_{\mathcal{G}}(j\omega)$ is left invertible $\forall\omega \in \mathbb{R}$.*

COROLLARY 3.9. *$\boldsymbol{S}_{\breve{\mathcal{G}}}(e^{j\theta})$ is right invertible $\forall\theta \in [0, 2\pi]$ iff $\det DD' \neq 0$ and $\boldsymbol{S}_{\mathcal{G}}(j\omega)$ is right invertible $\forall\omega \in \mathbb{R}$.*

3.4. Riccati equation. The last topic of this section is to study the H^∞ DARE associated with $\breve{\mathcal{G}}$. For that purpose, consider the operator

$$\begin{bmatrix} \acute{B}_\gamma \\ \breve{D}_\gamma \end{bmatrix} \doteq \begin{bmatrix} \mathcal{I}_h^* & 0 \\ 0 & I \end{bmatrix} \left(\begin{array}{c|c} A & B_\gamma \\ \hline I & 0 \\ C & 0 \end{array}\right), \tag{3.2}$$

where B_γ is a given matrix of appropriate dimensions. Now, augment the "B" and "D" parameters of $\breve{\mathcal{G}}$ as follows:

$$\acute{B}_\alpha \doteq \begin{bmatrix} \acute{B}_\gamma & \acute{B} \end{bmatrix} \qquad \text{and} \qquad \breve{D}_\alpha \doteq \begin{bmatrix} \breve{D}_\gamma & \breve{D} \end{bmatrix}.$$

Consider the H^∞ DARE:

$$X = \bar{A}'X\bar{A} + \grave{C}^*\grave{C} - (\breve{D}_\alpha^*\grave{C} + \acute{B}_\alpha^*X\bar{A})^*(\breve{D}_\alpha^*\breve{D}_\alpha - E_{11} + \acute{B}_\alpha^*X\acute{B}_\alpha)^{-1}(\breve{D}_\alpha^*\grave{C} + \acute{B}_\alpha^*X\bar{A}), \tag{3.3}$$

where the partitioning of E_{11} corresponds to that of $\breve{D}_\alpha^* \breve{D}_\alpha$. It can be verified that, if $B_\gamma = 0$, then (3.3) becomes the standard H^2 DARE associated with $\breve{\mathcal{G}}$. Define the two matrices

$$M_\gamma \doteq \begin{bmatrix} -D'C & -B' & -D'D \\ A & B_\gamma B_\gamma' & B \\ -C'C & -A' & -C'D \end{bmatrix}$$

and

$$N_\gamma \doteq \begin{bmatrix} 0 & 0 & 0 \\ I & 0 & 0 \\ 0 & I & 0 \end{bmatrix}.$$

The relevant result can now be expressed as follows.

LEMMA 3.10. *The pair (M_γ, N_γ) is an extended Hamiltonian pair and the following statements are equivalent:*

(i) *The DARE* (3.3) *has a stabilizing solution X.*

(ii) $(M_\gamma, N_\gamma) \in \mathrm{dom}(Ric_{\mathbb{C}^-})$.

Moreover, if either of these conditions holds, then $(X, F_2) = Ric_{\mathbb{C}^-}(M_\gamma, N_\gamma)$,

$$\begin{bmatrix} \grave{F}_1 \\ \grave{F}_2 \end{bmatrix} \doteq -(\breve{D}_\alpha^* \breve{D}_\alpha - E_{11} + \acute{B}_\alpha^* X \acute{B}_\alpha)^{-1}(\breve{D}_\alpha^* \grave{C} + \acute{B}_\alpha^* X \bar{A})$$

$$= \left(\begin{array}{c|c} A + B_\gamma B_\gamma' X + BF_2 & I \\ \hline B_\gamma' X & 0 \\ F_2 & 0 \end{array} \right) \mathcal{I}_0,$$

and $\bar{A} + \acute{B}_\gamma \grave{F}_1 + \acute{B} \grave{F}_2 = e^{(A + B_\gamma B_\gamma' X + BF_2)h}$.

Proof. The proof is contained in [10]. □

REMARK 3.1. *It follows from Lemma* A.2 *that $(X, F_2) = Ric_{\mathbb{C}^-}(M_\gamma, N_\gamma)$ implies that X is actually the stabilizing solution of the H^∞* continuous-time *algebraic Riccati equation*

$$A'X + XA + C'C + XB_\gamma B_\gamma' X - (D'C + BX)'(D'D)^{-1}(D'C + BX) = 0$$

and $F_2 = -(D'D)^{-1}(D'C + BX)$. Thus, the DARE (3.3) *associated with the system $\breve{\mathcal{G}}$, augmented by $\acute{B}_\gamma$ and $\breve{D}_\gamma$ is, in a sense, equivalent to the CARE associated with $\mathcal{G}$, augmented by B_γ.*

4. System of the form $\mathcal{W}_h \mathcal{G} \mathcal{H}_h$. This section is concerned with systems of the form $\grave{\mathcal{G}} \doteq \mathcal{W}_h \mathcal{G} \mathcal{H}_h$, where $\mathcal{G}$ and $\mathcal{H}_h$ are given by (2.2) and (2.3), respectively. The transfer function of $\grave{\mathcal{G}}$ can be expressed as

$$\grave{G}(z) = \left[\begin{array}{c|c} \bar{A} & \bar{B} \\ \hline \grave{C} & \grave{D} \end{array} \right].$$

Here $\bar{A}$ and $\grave{C}$ are the same as for $\breve{\mathcal{G}}$, while $\bar{B} = \acute{B} \grave{\Phi}_H$ and $\grave{D} = \breve{D} \grave{\Phi}_H$. To ensure that the operator $\grave{D}$ is well defined as a mapping $\mathbb{R} \mapsto L^2[0, h]$, assume that

$$DD_H = 0.$$

The purpose of this section is to study the singularities of $\boldsymbol{S}_{\grave{\mathcal{G}}}(z)$ and $\boldsymbol{C}_{\grave{\mathcal{G}}}(z)$ and the H^∞ DARE associated with $\grave{\mathcal{G}}$. In particular, it is shown that, under the assumption

$\|\breve{D}_\gamma\|_2 < 1$, both the invariant zeros of $\grave{\mathcal{G}}$ and the stabilizing solution of its associated DARE can be characterized in terms of the matrices

$$\Sigma \doteq \exp\left(\begin{bmatrix} -A'_H & -C'_H D'C & -C'_H B' & -C'_H D' D C_H \\ 0 & A & B_\gamma B'_\gamma & BC_H \\ 0 & -C'C & -A' & -C'DC_H \\ 0 & 0 & 0 & A_H \end{bmatrix} h\right) = \begin{bmatrix} \Sigma_{11} & \Sigma_{12} & \Sigma_{13} & \Sigma_{14} \\ 0 & \Sigma_{22} & \Sigma_{23} & \Sigma_{24} \\ 0 & \Sigma_{32} & \Sigma_{33} & \Sigma_{34} \\ 0 & 0 & 0 & \Sigma_{44} \end{bmatrix}, \tag{4.1}$$

where B_γ is as defined in the previous section, and

$$M_H \doteq \begin{bmatrix} I & BD_H \\ 0 & B_H \end{bmatrix}.$$

When B_γ is sufficiently "small," including the case $B_\gamma = 0$, the uncontrollable modes of $\grave{\mathcal{G}}$ can also be expressed in terms of these matrices.

4.1. Preliminary results on $\mathcal{W}_h\mathcal{G}\mathcal{H}_h$. To study the properties of $\grave{\mathcal{G}}$, it is convenient to develop some preliminary results.

PROPOSITION 4.1. $\det \Sigma_{33} \neq 0$ *for any* B_γ *such that* $\|\breve{D}_\gamma\|_2 < 1$.

Proof. This is Theorem 13.5.1 in [3]. □

In order to formulate the next lemma, assume that $\|\breve{D}_\gamma\|_2 < 1$ and define the matrices

$$M_A \doteq \begin{bmatrix} \Sigma_{22} & \Sigma_{24} \end{bmatrix} - \Sigma_{23}\Sigma_{33}^{-1}\begin{bmatrix} \Sigma_{32} & \Sigma_{34} \end{bmatrix}, \tag{4.2a}$$

$$M_{12} \doteq -\begin{bmatrix} \Sigma_{33} & 0 \\ \Sigma_{13} & \Sigma_{11} \end{bmatrix}^{-1}\begin{bmatrix} \Sigma_{32} & \Sigma_{34} \\ \Sigma_{12} & \Sigma_{14} \end{bmatrix}, \tag{4.2b}$$

$$M_{21} \doteq \Sigma_{23}\Sigma_{33}^{-1}. \tag{4.2c}$$

PROPOSITION 4.2. *If* $\|\breve{D}_\gamma\|_2 < 1$, *then*

$$\begin{bmatrix} \bar{A} & \bar{B} \end{bmatrix} + \acute{B}_\gamma \breve{D}_\gamma^*(I - \breve{D}_\gamma\breve{D}_\gamma^*)^{-1}\begin{bmatrix} \grave{C} & \grave{D} \end{bmatrix} = M_A M_H,$$

$$\begin{bmatrix} \grave{C}^* \\ \grave{D}^* \end{bmatrix}(I - \breve{D}_\gamma\breve{D}_\gamma^*)^{-1}\begin{bmatrix} \grave{C} & \grave{D} \end{bmatrix} = M'_H M_{12} M_H,$$

and

$$\acute{B}_\gamma(I - \breve{D}_\gamma^*\breve{D}_\gamma)^{-1}\acute{B}_\gamma^* = M_{21}.$$

Proof. Under the additional assumption that $D_H = 0$, this lemma was established in section IV of [11]. The same arguments used there can be applied to establish the more general result. □

As shown next, manipulations with subblocks are facilitated by the symplectic structure of the matrix Σ.

PROPOSITION 4.3. *If* $\|\breve{D}_\gamma\|_2 < 1$, *then*

$$\begin{bmatrix} \Sigma_{33}^{-1} & 0 \\ 0 & \Sigma_{14} \end{bmatrix}' = \begin{bmatrix} I & 0 \\ -\Sigma'_{34} & \Sigma'_{44} \end{bmatrix}\left(\begin{bmatrix} \Sigma_{22} & \Sigma_{24} \\ \Sigma_{12} & \Sigma_{14} \end{bmatrix} - \begin{bmatrix} \Sigma_{23} \\ \Sigma_{13} \end{bmatrix}\Sigma_{33}^{-1}\begin{bmatrix} \Sigma_{32} & \Sigma_{34} \end{bmatrix}\right)\begin{bmatrix} I & \Sigma'_{13} \\ 0 & \Sigma'_{11} \end{bmatrix}.$$

Proof. From [18, section 21.3], two relevant properties hold for any symplectic matrix

$$S = \begin{bmatrix} S_{11} & S_{12} \\ S_{21} & S_{22} \end{bmatrix}$$

such that $\det S_{22} \neq 0$: *a)* $S_{11} = (S'_{22})^{-1} + S_{12}S_{22}^{-1}S_{21}$ and *b)* $S_{22}^{-1}S_{21}$ is symmetric. On the other hand, it is easy to see that the matrix

$$\begin{bmatrix} \Sigma_{22} & \Sigma_{24} & \Sigma_{23} & 0 \\ 0 & \Sigma_{44} & 0 & 0 \\ \Sigma_{32} & \Sigma_{34} & \Sigma_{33} & 0 \\ \Sigma_{12} & \Sigma_{14} & \Sigma_{13} & \Sigma_{11} \end{bmatrix} = \exp\left(\begin{bmatrix} A & BC_H & B_\gamma B'_\gamma & 0 \\ 0 & A_H & 0 & 0 \\ -C'C & -C'DC_H & -A' & 0 \\ -C'_H D'C & -C'_H D'DC_H & -C'_H B' & -A'_H \end{bmatrix} h \right)$$

is symplectic. Then, from the nonsingularity of Σ_{33} and $\Sigma_{11} = e^{-A'_H h}$, the formula in the proposition can be derived from (a) and (b). ∎

PROPOSITION 4.4. *If* $\|\breve{D}_\gamma\|_2 < 1$, *then*

$$\Upsilon \doteq -(\Sigma_{23}\Sigma_{33}^{-1})^{1/2}\Sigma_{33}\begin{bmatrix} I & 0 \end{bmatrix}\left(-\begin{bmatrix} \Sigma_{33} & 0 \\ \Sigma_{13} & \Sigma_{11} \end{bmatrix}^{-1}\begin{bmatrix} \Sigma_{32} & \Sigma_{34} \\ \Sigma_{12} & \Sigma_{14} \end{bmatrix}\right)^{1/2}$$

is a contraction, i.e., $\bar{\sigma}(\Upsilon) < 1$.

Proof. The matrix Υ is a contraction iff

$$I - \Upsilon\Upsilon' = I + (\Sigma_{23}\Sigma_{33}^{-1})^{1/2}\Sigma_{32}\Sigma'_{33}(\Sigma_{23}\Sigma_{33}^{-1})^{1/2} > 0.$$

Without loss of generality, the pair (A, B_γ) can be assumed to be controllable. (Otherwise the problem can be easily reduced to the one with controllable (A, B_γ) by an appropriate change of basis [3, section 13.5].) Since $\acute{B}_\gamma\acute{B}^*_\gamma = \int_0^h e^{A\tau}B_\gamma B'_\gamma e^{A'\tau}d\tau$, the controllability of (A, B_γ) and Proposition 4.2 yield $M_{21} > 0$. Then

$$\begin{aligned} I - \Upsilon\Upsilon' &= M_{21}^{-1/2}M_{21}(I + \Sigma_{32}\Sigma'_{23})M_{21}^{-1/2} \\ &= M_{21}^{-1/2}\Sigma_{22}(\Sigma_{22}^{-1}\Sigma_{23})\Sigma'_{22}M_{21}^{-1/2}, \end{aligned}$$

where the latter equality is obtained by using $I + \Sigma_{32}\Sigma'_{23} = \Sigma_{33}\Sigma'_{22}$, which, in turn, follows from Proposition 4.3.

Now, to prove the proposition one needs to prove that $\det \Sigma_{23} \neq 0$. To this end, introduce the operators

$$\begin{bmatrix} \grave{C}_a & \breve{D}_a \end{bmatrix} \doteq \left(\begin{array}{c|cc} -A' & I & C' \\ \hline B'_\gamma & 0 & 0 \end{array}\right)\begin{bmatrix} \mathcal{I}_0^* & 0 \\ 0 & I \end{bmatrix}.$$

By applying Proposition 4.2, one gets that $\grave{C}^*_a(I - \breve{D}_a\breve{D}^*_a)^{-1}\grave{C}_a = \Sigma_{22}^{-1}\Sigma_{23}$. Since $\grave{C}^*_a\grave{C}_a = \int_0^h e^{-A\tau}B_\gamma B'_\gamma e^{-A'\tau}d\tau$, the controllability of (A, B_γ) yields $I - \Upsilon\Upsilon' > 0$. ∎

4.2. Uncontrollable modes. The pencil $\boldsymbol{C}_{\grave{\mathcal{G}}}(z)$ is finite dimensional, and consequently its singularities can be characterized by calculating the matrices $\bar{A}$ and $\bar{B}$. This calculation follows from a direct corollary to Proposition 4.2.

LEMMA 4.5. *If* $B_\gamma = 0$, *then the matrix* $\boldsymbol{C}_{\grave{\mathcal{G}}}(z)$ *is right invertible iff* $[\Sigma_{22} - zI \;\; zBD_H + \Sigma_{24}B_H\,]$ *is right invertible.*

Lemma 4.5 gives a complete characterization of the uncontrollable modes of $\grave{\mathcal{G}}$ by calculating Σ under the assumption $B_\gamma = 0$. It turns out that this assumption is associated with the solution of an H^2 optimization, while the solution of the a corresponding H^∞ is based on Σ for a nonzero B_γ. Hence, the verification of stabilizability (and observability) of the system of interest in the H^∞ case might require additional computations. It is then of interest to characterize the stabilizability of $(\bar{A}, \bar{B})$ in terms of Σ for a $B_\gamma \neq 0$. To this end, for a given matrix F define the system

$$\breve{\mathcal{G}}_F \doteq \left[\begin{array}{c|c} \bar{A} + \bar{B}F & \acute{B}_\gamma \\ \hline \grave{C} + \grave{D}F & \breve{D}_\gamma \end{array}\right].$$

Lemma 4.6. *Whenever B_γ is such that $\|\breve{D}_\gamma\|_2 < 1$, the matrix $\boldsymbol{C}_{\hat{\mathcal{G}}}(z)$ is right invertible $\forall |z| \geq 1$ and there exists a matrix F such that $\|\breve{\mathcal{G}}_F\|_{H^\infty} < 1$ only if the matrix $\left[\begin{array}{cc} \Sigma_{22} - zI & zBD_H + \Sigma_{24}B_H \end{array}\right]$ is right invertible $\forall |z| \geq 1$.*

Proof. Denote by $\mathcal{S}_{sc}$ the set of all (discrete-time with either finite- or infinite-dimensional input-output spaces) systems, which are internally stable and have H^∞ norm < 1. Standard loop shifting and orthogonal projection arguments [1] give

$$\breve{\mathcal{G}}_F \in \mathcal{S}_{sc} \iff \mathcal{F}_\ell\big(\breve{\Theta}_\gamma, \breve{\mathcal{G}}_F\big) \in \mathcal{S}_{sc} \iff \bar{\mathcal{G}}_F \in \mathcal{S}_{sc},$$

where the unitary operator

$$\breve{\Theta}_\gamma \doteq \left[\begin{array}{cc} -\breve{D}_\gamma & (I - \breve{D}_\gamma \breve{D}_\gamma^*)^{1/2} \\ (I - \breve{D}_\gamma^* \breve{D}_\gamma)^{1/2} & \breve{D}_\gamma^* \end{array}\right] = (\breve{\Theta}_\gamma^*)^{-1}$$

is well defined and

$$\bar{\mathcal{G}}_F \doteq \left[\begin{array}{c|c} M_A M_H M_F & M_{21}^{1/2} \\ \hline M_{12}^{1/2} M_H M_F & 0 \end{array}\right],$$

where $M_F \doteq \left[\begin{smallmatrix} I \\ F \end{smallmatrix}\right]$. Since the matrix Υ defined in Proposition 4.4 is a contraction, there always exists a unitary dilation of Υ, say

$$\bar{\Theta} \doteq \left[\begin{array}{cc} ? & ? \\ ? & \Upsilon \end{array}\right] = (\bar{\Theta}')^{-1}.$$

Then,

$$\bar{\mathcal{G}}_F \in \mathcal{S}_{sc} \iff \mathcal{F}_\ell\big(\bar{\Theta}, \bar{\mathcal{G}}_F\big) \in \mathcal{S}_{sc}.$$

The corresponding "A"-matrix of $\mathcal{F}_\ell\big(\bar{\Theta}, \bar{\mathcal{G}}_F\big)$ is just

$$\left[\begin{array}{cc} \Sigma_{22} & \Sigma_{24} \end{array}\right] M_H M_F = \Sigma_{22} + \left[\begin{array}{cc} \Sigma_{22} & \Sigma_{24} \end{array}\right] \left[\begin{array}{c} BD_H \\ B_H \end{array}\right] F,$$

which implies that $(\bar{A}, \bar{B})$ is stabilizable and there exists F such that $\|\breve{\mathcal{G}}_F\|_{H^\infty} < 1$ only if the pair $(\Sigma_{22}, \Sigma_{22}BD_H + \Sigma_{24}B_H)$ is stabilizable. The latter, in turn, is equivalent to the right invertibility of the matrix pencil

$$\left[\begin{array}{cc} \Sigma_{22} - zI & \Sigma_{22}BD_H + \Sigma_{24}B_H \end{array}\right] = \left[\begin{array}{cc} \Sigma_{22} - zI & zBD_H + \Sigma_{24}B_H \end{array}\right] \left[\begin{array}{cc} I & BD_H \\ 0 & I \end{array}\right]$$

$\forall |z| \geq 1$, which proves the lemma. □

Lemma 4.6 actually establishes that the absence of the generalized eigenvalues outside the open unit disk of the pencil $\left[\begin{array}{cc} \Sigma_{22}-zI & zBD_H+\Sigma_{24}B_H \end{array}\right]$ is necessary to solve the (state feedback) H^∞ problem for the system

$$\left[\begin{array}{c|cc} \bar{A} & \acute{B}_\gamma & \bar{B} \\ \hline \grave{C} & \breve{D}_\gamma & \grave{D} \\ I & 0 & 0 \end{array}\right].$$

Since the solvability of the latter problem is necessary for the solvability of the corresponding H^∞ output feedback problem with a strictly proper controller, the $\mathbb{D}$-stabilizability of $(\bar{A},\bar{B})$ can be replaced with the condition above without loss of generality. Note also that the pencil above has exactly the same form as the corresponding pencil in Lemma 4.5. Thus, as $B_\gamma \to 0$ the condition of Lemma 4.6 becomes sufficient.

4.3. Invariant zeros. In principle, the invariant zeros of the natural realization of $\grave{\mathcal{G}}$ can be characterized by noting that

$$\ker \boldsymbol{S}_{\grave{\mathcal{G}}}(z) = \ker \left[\begin{array}{cc} \bar{A}-zI & \bar{B} \\ \grave{C}^*\grave{C} & \grave{C}^*\grave{D} \\ \grave{D}^*\grave{C} & \grave{D}^*\grave{D} \end{array}\right].$$

The matrices $\bar{A}$, $\bar{B}$, $\grave{C}^*\grave{C}$, $\grave{C}^*\grave{D}$, and $\grave{D}^*\grave{D}$ are given by Proposition 4.2 when $B_\gamma = 0$. These matrices are precisely the same calculated for the solution of the sampled-data H^2 problems. Unfortunately, as noticed in the discussion after Lemma 4.5, such an approach does not fit well into H^∞ optimization, since the matrices above are not required for the solution. Consequently, it is of interest to express singularities of $\boldsymbol{S}_{\grave{\mathcal{G}}}(z)$ in terms of the H^∞ data, i.e., in terms of Σ when $B_\gamma \neq 0$. Although incorporating a nonzero B_γ into the computations makes the derivations more involved, the final formulae are not more complicated than those obtained in the H^2 case. Moreover, when $B_\gamma = 0$ the formulae reduce to the H^2 ones.

LEMMA 4.7. *Whenever B_γ is such that $\|\breve{D}_\gamma\|_2 < 1$, the operator $\boldsymbol{S}_{\grave{\mathcal{G}}}(z)$ is left invertible iff the matrix*

$$\left[\begin{array}{cc} \Sigma_{22}-zI & zBD_H+\Sigma_{24}B_H \\ B_H'\Sigma_{44}'\Sigma_{12} & B_H'\Sigma_{44}'\Sigma_{14}B_H \\ \Sigma_{32} & \Sigma_{34}B_H \end{array}\right]$$

is left invertible.

Proof. It is clear that if $\|\breve{D}_\gamma\|_2 < 1$ then $\boldsymbol{S}_{\grave{\mathcal{G}}}(z)$ is left invertible iff the matrix

$$\left[\begin{array}{cc} I & \acute{B}_\gamma\breve{D}_\gamma \\ 0 & \grave{C}^* \\ 0 & \grave{D} \end{array}\right]\left[\begin{array}{cc} I & 0 \\ 0 & (I-\breve{D}_\gamma\breve{D}_\gamma^*)^{-1} \end{array}\right]\left[\begin{array}{cc} \bar{A}-zI & \bar{B} \\ \grave{C} & \grave{D} \end{array}\right] = \left[\begin{array}{c} M_AM_H - z\left[\begin{array}{cc} I & 0 \end{array}\right] \\ M_H'M_{12}M_H \end{array}\right]$$

is left invertible. Using (4.2) one can verify that the latter matrix is just the Schur complement of Σ_{33} in

$$M_e \doteq \left[\begin{array}{ccc} \Sigma_{22}-zI & \Sigma_{22}BD_H+\Sigma_{24}B_H & ? \\ 0 & 0 & I \\ -B_H'\Sigma_{44}'\Sigma_{12} & -B_H'\Sigma_{44}'\Sigma_{12}BD_H - B_H'\Sigma_{44}'\Sigma_{14}B_H & ? \\ \Sigma_{32} & \Sigma_{32}BD_H+\Sigma_{34}B_H & \Sigma_{33} \end{array}\right],$$

where ? denotes irrelevant blocks. Therefore, and as $\det \Sigma_{33} \neq 0$, the left invertibility of the operator of interest is equivalent to the left invertibility of M_e. The latter, in turn, is equivalent to the left invertibility of the Schur complement of the identity I in M_e, i.e.,

$$\begin{bmatrix} \Sigma_{22} - zI & zBD_H + \Sigma_{24}B_H \\ -B'_H\Sigma'_{44}\Sigma_{12} & -B'_H\Sigma'_{44}\Sigma_{14}B_H \\ \Sigma_{32} & \Sigma_{34}B_H \end{bmatrix} \begin{bmatrix} I & BD_H \\ 0 & I \end{bmatrix}.$$

This completes the proof. □

REMARK 4.1. *The test in Lemma* 4.7 *can be simplified if additional assumptions are made about the hold function* ϕ_H. *In particular, when* B_H *is square and invertible and* $D_H = 0$, *the pencil reduces to*

$$\begin{bmatrix} \Sigma_{22} - zI & \Sigma_{24} \\ \Sigma_{12} & \Sigma_{14} \\ \Sigma_{32} & \Sigma_{34} \end{bmatrix},$$

while, when $B_H = 0$ *and* $D_H = I$, *the pencil becomes*

$$\begin{bmatrix} \Sigma_{22} - zI & zB \\ \Sigma_{32} & 0 \end{bmatrix}.$$

4.4. Riccati equation. The final derivations of the section are again devoted to the Riccati equation associated with the system. Consider the DARE

$$\begin{aligned} X &= \bar{A}'X\bar{A} + \grave{C}^*\grave{C} \\ &\quad - (\breve{D}^*_\beta\grave{C} + \acute{B}^*_\beta X\bar{A})^*(\breve{D}^*_\beta\breve{D}_\beta - E_{11} + \acute{B}^*_\beta X\acute{B}_\beta)^{-1}(\breve{D}^*_\beta\grave{C} + \acute{B}^*_\beta X\bar{A}), \end{aligned} \tag{4.3}$$

where $\acute{B}_\beta \doteq \begin{bmatrix} \acute{B}_\gamma & \bar{B} \end{bmatrix}$ and $\breve{D}_\beta \doteq \begin{bmatrix} \breve{D}_\gamma & \grave{D} \end{bmatrix}$. This DARE appears when treating the H^∞ problem with fixed hold. Although the equation can be made finite dimensional by computing an LU-decomposition of the operator $\breve{D}^*_\beta\breve{D}_\beta - E_{11} + \acute{B}^*_\beta X\acute{B}_\beta$, the intermediate steps involved in the calculations severely limit any further analysis. Instead, (4.3) can be replaced by an extended symplectic matrix pair, with the advantage that the stabilizing solution can be characterized in terms of the subblocks of the matrix Σ. Moreover, this approach allows the derivation of the formula for the "gain"

$$\grave{F} \doteq -(\breve{D}^*_\beta\breve{D}_\beta - E_{11} + \acute{B}^*_\beta X\acute{B}_\beta)^{-1}(\breve{D}^*_\beta\grave{C} + \acute{B}^*_\beta X\bar{A}) = \begin{bmatrix} \grave{F}_1 \\ \bar{F}_2 \end{bmatrix},$$

required for the lifted solution in [12].

To this end we define the following two matrices:

$$M_\gamma \doteq \begin{bmatrix} B'_H\Sigma'_{44}\Sigma_{12} & B'_H\Sigma'_{44}\Sigma_{13} & B'_H\Sigma'_{44}\Sigma_{14} \\ \Sigma_{22} & \Sigma_{23} & \Sigma_{24} \\ \Sigma_{32} & \Sigma_{33} & \Sigma_{34} \end{bmatrix} \begin{bmatrix} I & 0 & BD_H \\ 0 & I & 0 \\ 0 & 0 & B_H \end{bmatrix} - \begin{bmatrix} 0 & D'_HB' & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}$$

and

$$N_\gamma \doteq \begin{bmatrix} 0 & 0 & 0 \\ I & 0 & 0 \\ 0 & I & 0 \end{bmatrix}.$$

The main results concerning the DARE (4.3) can now be stated.

LEMMA 4.8. *Whenever B_γ is such that $\|\breve{D}_\gamma\|_2 < 1$, the pair (M_γ, N_γ) is an extended symplectic pair, and the following statements are equivalent:*

(i) *The DARE* (4.3) *has a stabilizing solution X.*

(ii) $(M_\gamma, N_\gamma) \in \mathrm{dom}(Ric_{\mathbb{D}})$.

Moreover, if either of these conditions holds, then $(X, \bar{F}_2) = Ric_{\mathbb{D}}(M_\gamma, N_\gamma)$,

$$\grave{F}_1 = \left(\begin{array}{ccc|c} A & B_\gamma B_\gamma' & BC_H & I + BD_H\bar{F}_2 \\ -C'C & -A' & -C'DC_H & X \\ 0 & 0 & A_H & B_H\bar{F}_2 \\ \hline 0 & B_\gamma' & 0 & 0 \end{array}\right)\mathcal{I}_0, \tag{4.4}$$

and $\bar{A} + \acute{B}_\gamma\grave{F}_1 + \bar{B}\bar{F}_2 = \Sigma_{22} + \Sigma_{23}X + (\Sigma_{22}BD_H + \Sigma_{24}B_H)\bar{F}_2$.

Proof. As follows from Lemma A.1, the solvability of the DARE (4.3) is equivalent to the condition $(M_a, N_a) \in \mathrm{dom}(Ric_{\mathbb{D}})$, where

$$M_a \doteq \begin{bmatrix} \bar{A} + \acute{B}_\gamma\breve{D}_\gamma^*(I - \breve{D}_\gamma\breve{D}_\gamma^*)^{-1}\grave{C} & 0 & \bar{B} + \acute{B}_\gamma\breve{D}_\gamma^*(I - \breve{D}_\gamma\breve{D}_\gamma^*)^{-1}\grave{D} \\ \grave{C}^*(I - \breve{D}_\gamma\breve{D}_\gamma^*)^{-1}\grave{C} & -I & \grave{C}^*(I - \breve{D}_\gamma\breve{D}_\gamma^*)^{-1}\grave{D} \\ \grave{D}^*(I - \breve{D}_\gamma\breve{D}_\gamma^*)^{-1}\grave{C} & 0 & \grave{D}^*(I - \breve{D}_\gamma\breve{D}_\gamma^*)^{-1}\grave{D} \end{bmatrix}$$

and

$$N_a \doteq \begin{bmatrix} I & -\acute{B}_\gamma(I - \breve{D}_\gamma^*\breve{D}_\gamma)^{-1}\acute{B}_\gamma^* & 0 \\ 0 & -(\bar{A} + \acute{B}_\gamma\breve{D}_\gamma^*(I - \breve{D}_\gamma\breve{D}_\gamma^*)^{-1}\grave{C})' & 0 \\ 0 & -(\bar{B} + \acute{B}_\gamma\breve{D}_\gamma^*(I - \breve{D}_\gamma\breve{D}_\gamma^*)^{-1}\grave{D})' & 0 \end{bmatrix}.$$

Denote by (M_b, N_b) the matrix pair obtained by the permutation of the second and the third columns of (M_a, N_a). Then, using Proposition 4.2, we have that

$$(M_b, N_b) = \left(\begin{bmatrix} M_A M_H & 0 \\ M_H' M_{12} M_H & -E_1 \end{bmatrix}, \begin{bmatrix} E_1' & -M_{21} \\ 0 & -M_H' M_A' \end{bmatrix}\right).$$

Furthermore, using Proposition 4.3,

$$M_A' = \begin{bmatrix} I \\ -\Sigma_{11}^{-1}\Sigma_{13} \end{bmatrix}\Sigma_{33}^{-1}.$$

Then, taking into account (4.2) and the fact that $\Sigma_{11}^{-1} = \Sigma_{44}'$, one verifies

$$\begin{gathered} M_A - \Sigma_{23}E_1'M_{12} = \begin{bmatrix} \Sigma_{22} & \Sigma_{24} \end{bmatrix}, \\ M_{21} - \Sigma_{23}E_1'M_A' = 0, \\ MM_H'M_{12} = -\begin{bmatrix} I & 0 \\ 0 & B_H'\Sigma_{44}' \end{bmatrix}\begin{bmatrix} \Sigma_{32} & \Sigma_{34} \\ \Sigma_{12} & \Sigma_{14} \end{bmatrix}, \end{gathered}$$

and

$$MM_H'M_A' = E_1,$$

where

$$M \doteq \begin{bmatrix} \Sigma_{33} & 0 \\ B_H'\Sigma_{44}'\Sigma_{13} & I \end{bmatrix}\begin{bmatrix} I & 0 \\ -D_H'B' & I \end{bmatrix}.$$

With these formulae, the equivalence of $Ric_{\mathbb{D}}(M_a, N_a)$ and $Ric_{\mathbb{D}}(M_\gamma, N_\gamma)$ follows by premultiplying the pair (M_b, N_b) by the (nonsingular) matrix $\left[\begin{smallmatrix} I & -\Sigma_{23}E_1' \\ 0 & -M \end{smallmatrix}\right]$ and permuting the second and the third block columns of the resulting pair.

Consider now the operator $\grave{F}_1$ which, by Lemma A.1 in the appendix, is given by

$$\grave{F}_1 = \big(I - \breve{D}_\gamma^* \breve{D}_\gamma - \acute{B}_\gamma^* X \acute{B}_\gamma\big)^{-1}\big(\breve{D}_\gamma^* \grave{C} + \acute{B}_\gamma^* X \bar{A} + (\breve{D}_\gamma^* \grave{D} + \grave{B}_\gamma^* X \bar{B})\bar{F}_2\big),$$

where the operator to be inverted is nonsingular. Note, that $\bar{B} = \acute{B}\grave{\Phi}_H$ and $\grave{D} = \breve{D}\grave{\Phi}_H$ and introduce the following two operators:

$$\begin{aligned} \breve{O}_1 &\doteq \breve{D}_\gamma^* \breve{D}_\gamma + \acute{B}_\gamma^* X \acute{B}_\gamma, \\ \breve{O}_2 &\doteq \breve{D}_\gamma^* \left[\begin{smallmatrix} \grave{C} & \breve{D} \end{smallmatrix}\right] + \acute{B}_\gamma^* X \left[\begin{smallmatrix} \bar{A} & \acute{B} \end{smallmatrix}\right]. \end{aligned}$$

Applying Proposition 2.2 to these operators,

$$\breve{O}_1 = \left(\begin{array}{cc|c|c} A & 0 & \left[\begin{smallmatrix} I & 0 \\ 0 & 0 \end{smallmatrix}\right] \rightleftharpoons \left[\begin{smallmatrix} 0 & 0 \\ -X & I \end{smallmatrix}\right] & B_\gamma \\ -C'C & -A' & & 0 \\ \hline 0 & B_\gamma' & & 0 \end{array}\right),$$

$$\breve{O}_2 = \left(\begin{array}{cc|c|cc} A & 0 & \left[\begin{smallmatrix} I & 0 \\ 0 & 0 \end{smallmatrix}\right] \rightleftharpoons \left[\begin{smallmatrix} 0 & 0 \\ -X & I \end{smallmatrix}\right] & I & B \\ -C'C & -A' & & 0 & -C'D \\ \hline 0 & B_\gamma' & & 0 & 0 \end{array}\right) \left[\begin{smallmatrix} \mathcal{I}_0 & 0 \\ 0 & I \end{smallmatrix}\right]$$

and then, according to Proposition 2.1,

$$(I - \breve{O}_1)^{-1}\breve{O}_2 = \left(\begin{array}{cc|c|cc} A & B_\gamma B_\gamma' & \left[\begin{smallmatrix} I & 0 \\ 0 & 0 \end{smallmatrix}\right] \rightleftharpoons \left[\begin{smallmatrix} 0 & 0 \\ -X & I \end{smallmatrix}\right] & I & B \\ -C'C & -A' & & 0 & -C'D \\ \hline 0 & B_\gamma' & & 0 & 0 \end{array}\right) \left[\begin{smallmatrix} \mathcal{I}_0 & 0 \\ 0 & I \end{smallmatrix}\right]$$

and the nonsingularity of $I - \breve{O}_1$ is equivalent to the nonsingularity of $\Sigma_{33} - X\Sigma_{23}$. Since

$$\left[\begin{smallmatrix} \mathcal{I}_0 & 0 \\ 0 & I \end{smallmatrix}\right] \left[\begin{smallmatrix} I \\ \grave{\Phi}_H \bar{F}_2 \end{smallmatrix}\right] = \left(\begin{array}{c|c} A_H & B_H\bar{F}_2 \\ \hline 0 & I \\ C_H & D_H\bar{F}_2 \end{array}\right)\mathcal{I}_0$$

and $DD_H = 0$, one gets

$$\grave{F}_1 = \left(\begin{array}{ccc|c|c} A & B_\gamma B_\gamma' & BC_H & \left[\begin{smallmatrix} I & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & I \end{smallmatrix}\right] \rightleftharpoons \left[\begin{smallmatrix} 0 & 0 & 0 \\ -X & I & 0 \\ 0 & 0 & 0 \end{smallmatrix}\right] & I + BD_H\bar{F}_2 \\ -C'C & -A' & -C'DC_H & & 0 \\ 0 & 0 & A_H & & B_H\bar{F}_2 \\ \hline 0 & B_\gamma' & 0 & & 0 \end{array}\right)\mathcal{I}_0.$$

Now (4.4) follows from

$$X = -(\Sigma_{33} - X\Sigma_{23})^{-1}\big((\Sigma_{32} - X\Sigma_{22})(I + BD_H\bar{F}_2) + (\Sigma_{34} - X\Sigma_{24})B_H\bar{F}_2\big),$$

which, in turn, can easily be derived form the fact that $\operatorname{Im}\left[\begin{smallmatrix} I & X & \bar{F}_2' \end{smallmatrix}\right]'$ is the stable deflating subspace of the pencil $M_\gamma - \lambda N_\gamma$. The latter fact also leads to the equality for $\bar{A} + \acute{B}_\beta \grave{F}$. □

REMARK 4.2. *As in the cases of the uncontrollable modes and invariant zeros, additional assumptions on the hold function ϕ_H may substantially simplify the form of the matrix M_γ. In the case where $D_H = 0$ and $B_H = I$ it becomes*

$$M_\gamma = \begin{bmatrix} \Sigma_{12} & \Sigma_{13} & \Sigma_{14} \\ \Sigma_{22} & \Sigma_{23} & \Sigma_{24} \\ \Sigma_{32} & \Sigma_{33} & \Sigma_{34} \end{bmatrix},$$

while for $D_H = I$ and $B_H = 0$

$$M_\gamma = \begin{bmatrix} 0 & B' & 0 \\ \Sigma_{22} & \Sigma_{23} & \Sigma_{22}B \\ \Sigma_{32} & \Sigma_{33} & \Sigma_{32}B \end{bmatrix}.$$

5. Coupling condition. Consider the inequality

$$\left\| \begin{bmatrix} X^{1/2} & 0 \\ 0 & I \end{bmatrix} \begin{bmatrix} \bar{A} & \acute{B}_\gamma \\ \grave{C} & \breve{D}_\gamma \end{bmatrix} \begin{bmatrix} Y^{1/2} & 0 \\ 0 & I \end{bmatrix} \right\|_2 < 1, \tag{5.1}$$

where $X = X' \geq 0$ and $Y = Y' \geq 0$ are given matrices and

$$\begin{bmatrix} \bar{A} & \acute{B}_\gamma \\ \grave{C} & \breve{D}_\gamma \end{bmatrix} = \begin{bmatrix} \mathcal{I}_h^* & 0 \\ 0 & I \end{bmatrix} \left(\begin{array}{c|cc} A & I & B_\gamma \\ \hline I & 0 & 0 \\ C & 0 & 0 \end{array} \right) \begin{bmatrix} \mathcal{I}_0 & 0 \\ 0 & I \end{bmatrix}.$$

This is actually the coupling condition for the solutions X and Y of the H^∞ DAREs, which appears in the solution of the sampled-data H^∞ problem (see [12, Theorem 2]). The purpose of this section is to express inequality (5.1) in terms of the matrix

$$\begin{bmatrix} \Sigma_{22} & \Sigma_{23} \\ \Sigma_{32} & \Sigma_{33} \end{bmatrix},$$

defined by (4.1), which can be readily computed.

LEMMA 5.1. *Whenever B_γ is such that $\|\breve{D}_\gamma\|_2 < 1$, inequality* (5.1) *holds iff the following three inequalities hold:*

(a) $\rho(X\,\Sigma_{23}\Sigma_{33}^{-1}) < 1$*;*

(b) $\rho(\Sigma_{33}^{-1}\Sigma_{32}Y) < 1$*;*

(c) $\rho(Y(\Sigma_{33} + \Sigma_{32}Y)^{-1}X(\Sigma'_{33} - \Sigma'_{23}X)^{-1}) < 1$.

Proof. Consider the following two operators: $\acute{O}_X \doteq X^{1/2}\acute{B}_\gamma(I - \breve{D}_\gamma^*\breve{D}_\gamma)^{-1/2}$ and $\grave{O}_Y \doteq (I - \breve{D}_\gamma\breve{D}_\gamma^*)^{-1/2}\grave{C}Y^{1/2}$. By standard dilation theory [18, section 2.11], the inequality (5.1) holds iff the matrix

$$M_a \doteq (I - \acute{O}_X\acute{O}_X^*)^{-1/2}(X^{1/2}\bar{A}Y^{1/2} + \acute{O}_X\breve{D}_\gamma^*\grave{O}_Y)(I - \grave{O}_Y^*\grave{O}_Y)^{-1/2}$$

is well defined and a contraction, i.e., $\bar{\sigma}(M_a) < 1$. By Proposition 4.2,

$$\acute{O}_X\acute{O}_X^* = X^{1/2}\Sigma_{23}\Sigma_{33}^{-1}X^{1/2} \qquad \text{and} \qquad \grave{O}_Y^*\grave{O}_Y = -Y^{1/2}\Sigma_{33}^{-1}\Sigma_{32}Y^{1/2}.$$

Consequently, M_a is well defined iff conditions (a) and (b) hold. Using Propositions 4.2 and 4.3 one gets

$$X^{1/2}\bar{A}Y^{1/2} + \acute{O}_X\breve{D}_\gamma^*\grave{O}_Y = X^{1/2}(\Sigma'_{33})^{-1}Y^{1/2}.$$

Then, since for any matrices M_1 and M_2 of appropriate dimensions $\rho(M_1M_2) = \rho(M_2M_1)$ and the matrix $\Sigma_{23}\Sigma_{33}^{-1}$ is symmetric,

$$\rho(M_aM_a') = \rho\big(Y(\Sigma_{33} + \Sigma_{32}Y)^{-1}X(\Sigma'_{33} - \Sigma'_{23}X)^{-1}\big).$$

The equality $\bar{\sigma}(M_a) = \rho(M_aM_a')$ concludes the proof. ∎

6. Conclusions. In this paper, several properties of systems represented in the lifted domain have been presented. The material considered was originally worked out for solving H^2 and H^∞ optimization problems for sampled-data systems. Eventually, it evolved to include results about systems in the lifted domain of independent general interest.

Appendix. The Riccati operators. The role of extended symplectic and Hamiltonian matrix pairs for solving algebraic Riccati equations has been extensively considered in the literature. See, for instance, [16, 14, 6]. The purpose of this appendix is to present a brief introduction to the results used in the paper.

Recall that the ordered pair of $(2n+m)\times(2n+m)$ matrices (M,N) is called an extended symplectic matrix pair (ESMP) if the associated matrix pencil $M-\lambda N$ verifies the following:

(a) $\det(M-\lambda N)\not\equiv 0$.

(b) If $\lambda\notin\{0,\infty\}$ is a generalized eigenvalue of $M-\lambda N$ of multiplicity r, then so is $\frac{1}{\lambda}$.

(c) If 0 is an eigenvalue of M of multiplicity r, then it is an eigenvalue of N of multiplicity $r+m$.

An ESMP is said to be *dichotomic* if the associated matrix pencil has no generalized eigenvalues on the unit circle. If an ESMP (M,N) is dichotomic, then the pencil $M-\lambda N$ has n eigenvalues in $\mathbb{D}$. Consider the n-dimensional deflating subspace $\mathcal{X}_{\mathbb{D}}(M,N)$ corresponding to eigenvalues in $\mathbb{D}$. It is obvious that

$$\mathcal{X}_{\mathbb{D}}(M,N)=\operatorname{Im}\begin{bmatrix}X_1\\X_2\\X_3\end{bmatrix},$$

where $X_1,X_2\in\mathbb{R}^{n\times n}$, $X_3\in\mathbb{R}^{m\times n}$, and

$$M\begin{bmatrix}X_1\\X_2\\X_3\end{bmatrix}=N\begin{bmatrix}X_1\\X_2\\X_3\end{bmatrix}A_{st},\qquad \sigma(A_{st})\in\mathbb{D}.$$

A dichotomic ESMP is said to be disconjugate if the matrix X_1 is nonsingular. When an ESMP is disconjugate, it is possible to set $X\doteq X_2X_1^{-1}$ and $F\doteq X_3X_1^{-1}$. Since X and F can be uniquely determined from (M,N), it is possible to define the function $Ric_{\mathbb{D}}:(M,N)\to(X,F)$, and set $(X,F)=Ric_{\mathbb{D}}(M,N)$. The domain of $Ric_{\mathbb{D}}$, denoted $\operatorname{dom}(Ric_{\mathbb{D}})$, then consists of all disconjugate ESMP.

Consider now the following DARE, associated with the state feedback H^∞ (or H^2, when $\gamma=\infty$) control problem:

$$X=A'XA+C'C-(D'C+B'XA)'R_\gamma(X)^{-1}(D'C+B'XA), \tag{A.1}$$

where $B\doteq\begin{bmatrix}B_1 & B_2\end{bmatrix}$ and $D\doteq\begin{bmatrix}D_1 & D_2\end{bmatrix}$, and

$$R_\gamma(X)\doteq D'D-\gamma^2E_{11}+B'XB.$$

A solution X_γ of (A.1) is said to be *stabilizing* if $X_\gamma=X_\gamma'$, the matrix $R_\gamma(X_\gamma)$ is nonsingular, and the matrix $A+BF_\gamma$ is Schur, where

$$F_\gamma\doteq-R_\gamma(X_\gamma)^{-1}(D'C+B'X_\gamma A)=\begin{bmatrix}F_{\gamma1}\\F_{\gamma2}\end{bmatrix}.$$

The next lemma, borrowed from [8], establishes a strong equivalence between the stabilizing solution of the DARE (A.1) and an associated ESMP.

LEMMA A.1. *Assume that* $\bar{\sigma}(D_1) < \gamma$ *and form the matrix pair*

$$(M_\gamma, N_\gamma) \doteq \left(\begin{bmatrix} A_\gamma & 0 & B_\gamma \\ C'R_\gamma C & -I & C'R_\gamma D_2 \\ D_2'R_\gamma C & 0 & D_2'R_\gamma D_2 \end{bmatrix}, \begin{bmatrix} I & -B_1\tilde{R}_\gamma B_1' & 0 \\ 0 & -A_\gamma' & 0 \\ 0 & -B_\gamma' & 0 \end{bmatrix} \right),$$

where $R_\gamma \doteq (I - \gamma^{-2}D_1D_1')^{-1}$, $\tilde{R}_\gamma \doteq (\gamma^2 I - D_1'D_1)^{-1}$, *and*

$$\begin{bmatrix} A_\gamma & B_\gamma \end{bmatrix} \doteq \begin{bmatrix} A & B_2 \end{bmatrix} + \gamma^{-2}B_1D_1'R_\gamma \begin{bmatrix} C & D_2 \end{bmatrix}.$$

Then (M_γ, N_γ) *is ESMP and the following two statements are equivalent:*

(i) $(M_\gamma, N_\gamma) \in \mathrm{dom}(Ric_{\mathbb{D}})$.

(ii) *The DARE* (A.1) *has a (unique) stabilizing solution* X_γ.

Moreover, if either of these conditions holds, then $\det(\gamma^2 I - D_1'D_1 - B_1'X_\gamma B_1) \neq 0$,

$$(X_\gamma, F_{\gamma 2}) = Ric_{\mathbb{D}}(M_\gamma, N_\gamma),$$

and

$$F_{\gamma 1} = (\gamma^2 I - D_1'D_1 - B_1'X_\gamma B_1)^{-1}\big(D_1'C + B_1'X_\gamma A + (D_1'D_2 + B_1'X_\gamma B_2)F_{\gamma 2}\big).$$

The ordered pair of $(2n+m) \times (2n+m)$ matrices (M, N) is called an *extended Hamiltonian matrix pair* (EHMP) if the associated matrix pencil $M - \lambda N$ verifies the following.

(a) $\det(M - \lambda N) \not\equiv 0$.

(b) If $\lambda \neq \infty$ is a generalized eigenvalue of $M - \lambda N$ of multiplicity r, then so is $-\lambda$.

(c) 0 is an eigenvalue of N of multiplicity m.

By an analogy with ESMP, an EHMP is said to be *dichotomic* if the associated matrix pencil has no generalized eigenvalues on the imaginary axis. When the EHMP (M, N) is dichotomic, the pencil $M - \lambda N$ has n eigenvalues in $\mathbb{C}^-$. Consider the n-dimensional deflating subspace $\mathcal{X}_{\mathbb{C}^-}(M, N)$ corresponding to eigenvalues in $\mathbb{C}^-$, and write

$$\mathcal{X}_{\mathbb{C}^-}(M, N) = \mathrm{Im} \begin{bmatrix} X_1 \\ X_2 \\ X_3 \end{bmatrix},$$

where $X_1, X_2 \in \mathbb{R}^{n\times n}$, $X_3 \in \mathbb{R}^{m\times n}$, and

$$M \begin{bmatrix} X_1 \\ X_2 \\ X_3 \end{bmatrix} = N \begin{bmatrix} X_1 \\ X_2 \\ X_3 \end{bmatrix} A_{st}, \qquad \sigma(A_{st}) \in \mathbb{C}^-.$$

A dichotomic EHMP is said to be *disconjugate* if the matrix X_1 is nonsingular. For a disconjugate EHMP, set $X \doteq X_2X_1^{-1}$ and $F \doteq X_3X_1^{-1}$, and define the function $Ric_{\mathbb{C}^-} : (M, N) \to (X, F)$; thus $(X, F) = Ric_{\mathbb{C}^-}(M, N)$. The domain of $Ric_{\mathbb{C}^-}$, denoted $\mathrm{dom}(Ric_{\mathbb{C}^-})$, then consists of all disconjugate EHMP.

Hamiltonian pairs play a role in the analysis of CAREs similar to the one ESMPs play in the analysis of DAREs. Consider a CARE in general form:

$$A'X + XA + C'C - (B'X + D'C)'\big(D'D - \gamma^2 E_{11}\big)^{-1}(B'X + D'C) = 0, \tag{A.2}$$

where, again, $B \doteq \begin{bmatrix} B_1 & B_2 \end{bmatrix}$ and $D \doteq \begin{bmatrix} D_1 & D_2 \end{bmatrix}$. For finite γ this equation is the so-called H^∞ CARE, while as $\gamma \to \infty$ (A.2) becomes the H^2 CARE, associated with the state feedback problems. A solution X_γ of (A.2) is said to be *stabilizing* if $X_\gamma = X'_\gamma$, and the matrix $A + BF_\gamma$ is Hurwitz, where

$$F_\gamma \doteq -\big(D'D - \gamma^2 E_{11}\big)^{-1}(D'C + B'X_\gamma) = \begin{bmatrix} F_{\gamma 1} \\ F_{\gamma 2} \end{bmatrix}.$$

The following lemma, which establishes a strong equivalence between the stabilizing solution of the CARE (A.2) and an associated EHMP, can be formulated.

LEMMA A.2. *Assume that $\bar\sigma(D_1) < \gamma$ and form the matrix pair*

$$(M_\gamma, N_\gamma) \doteq \left(\begin{bmatrix} A_\gamma & B_1\tilde R_\gamma B_1' & B_\gamma \\ -C'R_\gamma C & -A'_\gamma & -C'R_\gamma D_2 \\ -D_2'R_\gamma C & -B'_\gamma & -D_2'R_\gamma D_2 \end{bmatrix}, \begin{bmatrix} I & 0 & 0 \\ 0 & I & 0 \\ 0 & 0 & 0 \end{bmatrix} \right),$$

where $R_\gamma \doteq (I - \gamma^{-2}D_1D_1')^{-1}$, $\tilde R_\gamma \doteq (\gamma^2 I - D_1'D_1)^{-1}$, and

$$\begin{bmatrix} A_\gamma & B_\gamma \end{bmatrix} \doteq \begin{bmatrix} A & B_2 \end{bmatrix} + \gamma^{-2}B_1D_1'R_\gamma \begin{bmatrix} C & D_2 \end{bmatrix}.$$

Then (M_γ, N_γ) is EHMP and the following two statements are equivalent:

(i) $(M_\gamma, N_\gamma) \in \mathrm{dom}(Ric_{\mathbb{C}^-})$.

(ii) *D_2 has full column rank and the CARE* (A.2) *has a (unique) stabilizing solution X_γ.*

Moreover, if either of these conditions holds, then

$$(X_\gamma, F_{\gamma 2}) = Ric_{\mathbb{C}^-}(M_\gamma, N_\gamma)$$

and

$$F_{\gamma 1} = (\gamma^2 I - D_1'D_1)^{-1}\big(D_1'(C + D_2F_{\gamma 2}) + B_1'X_\gamma\big).$$

REFERENCES

[1] B. BAMIEH AND J. B. PEARSON, *A general framework for linear periodic systems with applications to H^∞ sampled-data control*, IEEE Trans. Automat. Control, 37 (1992), pp. 418–435.

[2] B. BAMIEH, J. B. PEARSON, B. A. FRANCIS, AND A. TANNENBAUM, *A lifting technique for linear periodic systems with applications to sampled-data control*, Systems Control Lett., 17 (1991), pp. 79–88.

[3] T. CHEN AND B. A. FRANCIS, *Optimal Sampled-Data Control Systems*, Springer-Verlag, London, 1995.

[4] I. GOHBERG AND M. A. KAASHOEK, *Time varying linear systems with boundary conditions and integral operators,* I. *The transfer operator and its properties*, Integral Equations Operator Theory, 7 (1984), pp. 325–391.

[5] I. GOHBERG, P. LANCASTER, AND L. RODMAN, *Invariant Subspaces of Matrices with Applications*, John Wiley & Sons, NY, 1986.

[6] V. IONESCU AND M. WEISS, *Continuous and discrete-time Riccati theory: A Popov-function approach*, Linear Algebra Appl., 193 (1993), pp. 173–209.

[7] A. J. KRENER, *Boundary value linear systems*, Astérisque, 75/76 (1980), pp. 149–165.

[8] L. MIRKIN, *On discrete-time H^∞ problem with a strictly proper controller*, Int. J. Control, 66 (1997), pp. 747–765.

[9] L. MIRKIN AND Z. J. PALMOR, *Mixed discrete/continuous specifications in sampled-data H^2-optimal control*, Automatica J. IFAC, 33 (1997), pp. 1997–2014.

[10] L. Mirkin and Z. J. Palmor, *Algebraic Riccati equations in the lifted domain*, Linear Algebra Appl., submitted.
[11] L. Mirkin and Z. J. Palmor, *A new representation of parameters of lifted systems*, IEEE Trans. Automat. Control, 44 (1999), pp. 833–840.
[12] L. Mirkin, H. Rotstein, and Z. J. Palmor, *H^2 and H^∞ design of sampled-data systems using lifting—Part* I: *General framework and solutions*, SIAM J. Control Optim., 38 (1999), pp. 175–196.
[13] R. M. Redheffer, *On a certain linear fractional transformation*, J. Math. Phys., 39 (1960), pp. 269–286.
[14] H. P. Rotstein, *Constrained H^∞-Optimization for Discrete-Time Control*, Ph.D. thesis, California Institute of Technology, Pasadena, CA, 1992.
[15] G. Tadmor, *H^∞ optimal sampled-data control in continuous-time systems*, Int. J. Control, 56 (1992), pp. 99–141.
[16] P. Van Dooren, *A generalized eigenvalue approach for solving Riccati equations*, SIAM J. Sci. Statist. Comput., 2 (1981), pp. 121–135.
[17] Y. Yamamoto, *A function space approach to sampled data control systems and tracking problems*, IEEE Trans. Automat. Control, 39 (1994), pp. 703–713.
[18] K. Zhou, J. C. Doyle, and K. Glover, *Robust and Optimal Control*, Prentice-Hall, Englewood Cliffs, NJ, 1995.

SIAM J. CONTROL OPTIM.
Vol. 38, No. 1, pp. 219–236

WEAK SHARP MINIMA: CHARACTERIZATIONS AND SUFFICIENT CONDITIONS*

MARCIN STUDNIARSKI† AND DOUG E. WARD‡

Abstract. The problem of identifying weak sharp minima of order m, an important class of (possibly) nonisolated minima, is investigated in this paper. Some sufficient conditions for weak sharp minimality in nonsmooth mathematical programming are presented, and some characterizations of weak sharp minimality are obtained, with special attention given to orders one and two. It is also demonstrated that two of these sufficient conditions guarantee exactness of an l_1 penalty function. A key role in this paper is played by two geometric concepts: the limiting proximal normal cone and a generalization of the contingent cone.

Key words. weak sharp minimizer of order m, normal cone, contingent cone, directional derivative, subdifferential, exact penalty function

AMS subject classifications. Primary, 49J52; Secondary, 90C30

PII. S0363012996301269

1. Introduction.

This paper is devoted to the study of a special type of minimizer for the mathematical program

$$\min\{f(x) \mid x \in C\}, \tag{1.1}$$

where $f : \mathbb{R}^n \to \overline{\mathbb{R}} := [-\infty, +\infty]$ and C is a nonempty subset of $\mathbb{R}^n$.

DEFINITION 1.1 (see [7]). *Let $\|\cdot\|$ be the Euclidean norm on $\mathbb{R}^n$. Suppose that f is finite and constant on the set $S \subset \mathbb{R}^n$, and let $\bar{x} \in S \cap C$ and $m \geq 1$. For $x \in \mathbb{R}^n$, let*

$$\operatorname{dist}^m(x, S) := \inf\big\{ \|y - x\|^m \mid y \in S \big\}.$$

(a) *We say that $\bar{x}$ is a* weak sharp minimizer of order m *for* (1.1) *if there exists $\beta > 0$ such that*

$$f(x) - f(\bar{x}) \geq \beta \operatorname{dist}^m(x, S) \quad \forall x \in C. \tag{1.2}$$

(b) *For $\varepsilon > 0$, let $B(x, \varepsilon) := \{y \in \mathbb{R}^n \mid \|y - x\| \leq \varepsilon\}$. We say that $\bar{x}$ is a* weak sharp local minimizer of order m *for* (1.1) *if there exist $\beta > 0$ and $\varepsilon > 0$ such that*

$$f(x) - f(\bar{x}) \geq \beta \operatorname{dist}^m(x, S) \quad \forall x \in C \cap B(\bar{x}, \varepsilon). \tag{1.3}$$

As an illustration of Definition 1.1, consider $f : \mathbb{R}^2 \to \mathbb{R}$ defined by $f(x, y) = |x|^m$ for $m \geq 1$. Let $C = \mathbb{R}^2$ and $S = \{(0, y) \mid y \in \mathbb{R}\}$. Then each element of S is a weak sharp minimizer of order m for (1.1).

Weak sharp minima occur in many optimization problems. For example, every minimizer of a linear program is a weak sharp minimizer of order one. Burke and

*Received by the editors March 29, 1996; accepted for publication (in revised form) November 24, 1998; published electronically December 15, 1999.
http://www.siam.org/journals/sicon/38-1/30126.html
†Faculty of Mathematics, University of Łodz, ul. Stefana Banacha 22, 90-238 Łodz, Poland (marstud@imul.uni.lodz.pl).
‡Department of Mathematics and Statistics, Miami University, Oxford, OH 45056-1641 (wardde@muohio.edu).

Ferris [7] have shown that solutions of certain convex quadratic programs and linear complementarity problems are weak sharp minima of order one and that many optimization algorithms exhibit finite termination at weak sharp minima. Weak sharp minima of higher order are of great interest in sensitivity analysis in nonlinear programming [4], [5], [14], [15], [23], [24]; in particular, the presence of weak sharp minima in a parametric nonlinear program leads to Hölder continuity properties of the solution set multifunction [15].

In the case where S is the singleton $\{\bar{x}\}$, Definition 1.1 reduces to a familiar concept (see [2], [17], [26], [27], [28], and references therein). In this case, we will call $\bar{x}$ a *strict minimizer of order* m. When $m = 2$, such minimizers are also called strong minimizers, and when $m = 1$, they are called sharp minimizers or strongly unique minimizers.

How can weak sharp minima be identified? It is well known that standard sufficient conditions for minimality (e.g., [10, Theorem 4] for $m = 2$) are sufficient for strict minimality of order m; in fact, they often characterize strict minimality of order m in the presence of a constraint qualification [3], [25], [28]. For the more difficult case of nonisolated minima, this question has been the focus of much recent attention. Burke and Ferris [7] have developed general necessary conditions for weak sharp minima of order one and have shown that those conditions provide characterizations of weak sharp minimality of order one when f, C, and S are convex. Bonnans and Ioffe [4], [5] have derived sufficient conditions and characterizations for weak sharp minimality of order two in the case where f is a pointwise maximum of C^2 functions. In addition, Ward [28] has given some necessary conditions for weak sharp minimality of order m.

In this paper, we present sufficient conditions and characterizations for weak sharp minimality of order m in nonsmooth programming, continuing a line of research begun in [26] and [28]. As in those papers, we start by giving a characterization of weak sharp minimality that requires no differentiability assumptions. We then go on to consider simpler sufficient conditions and to apply our general results in important special cases.

An outline of this paper is as follows. We begin in section 2 by deriving a general characterization of weak sharp minimality of order m in the unconstrained case ($C = \mathbb{R}^n$ in (1.1)). Since our characterization is rather complicated, we then give a simpler sufficient condition for weak sharp minimality and discuss situations in which this sufficient condition provides a characterization. Both the characterization and sufficient condition involve the normal cone of Mordukhovich [18] to the set S at $\bar{x}$, a geometric object that seems to arise naturally in attempts to formulate conditions sufficient for weak sharp minimality. The sufficient condition also involves a special variant of the contingent epiderivative of f [1], a directional derivative that plays a key role in sufficient conditions for strict minimality.

In section 3, we apply the results of section 2 to produce sufficient conditions for weak sharp minimality of orders one and two in constrained problems. We pay particular attention to problems with inequality and/or equality constraints and give a special result for the case of linear constraints. We again mention situations in which sufficient conditions give actual characterizations. In section 4, we show that our sufficient optimality conditions for inequality-constrained problems are also sufficient for exactness of a standard l_1 penalty function. Finally in section 5, we adapt an idea of Auslender [2] to establish another set of sufficient conditions for weak sharp minimality of order one. We then compare these sufficient conditions, which involve

the subdifferential of Mordukhovich [18], [19], [20], [21] rather than a contingent-type epiderivative, with the corresponding conditions from sections 2 and 3.

We conclude this section with a compilation of some notation and definitions that will be useful throughout the paper. For a set $S \subset \mathbb{R}^n$, we denote the *closure* of S by $\operatorname{cl} S$, the *interior* of S by $\operatorname{int} S$, and the *boundary* of S by $\operatorname{bd} S$. The *indicator function* of S is the function i_S defined by $i_S(x) = 0$ for $x \in S$ and $i_S(x) = +\infty$ for $x \in \mathbb{R}^n \backslash S$.

Let $\langle \cdot, \cdot \rangle$ denote the usual inner product on $\mathbb{R}^n$. For a cone $S \subset \mathbb{R}^n$, the *polar* of S is the closed convex cone defined by

$$S^0 := \big\{ y \in R^n \mid \langle x, y \rangle \leq 0 \ \forall x \in S \big\}.$$

For a function $f : \mathbb{R}^n \to \overline{\mathbb{R}}$, the *epigraph* of f is the set

$$\operatorname{epi} f := \big\{ (x, r) \mid f(x) \leq r \big\}.$$

We say that f is *proper* if $\operatorname{epi} f$ is nonempty and f never takes on the value $-\infty$, and that f is *lower semicontinuous* (l.s.c.) if $\operatorname{epi} f$ is closed. If f is finite at x, we say that f is *strictly differentiable* at x [7] if there exists a linear mapping $\nabla f(x) : \mathbb{R}^n \to \mathbb{R}$ such that $\forall y \in \mathbb{R}^n$,

$$\lim_{(v,w,t) \to (x,y,0^+)} \big(f(w + tv) - f(w) \big) / t = \langle \nabla f(x), y \rangle .$$

For $i = 1, 2$, we say that f is C^i at x if f is i times differentiable on a neighborhood of x and the i th derivative function is continuous at x. We will say that f is $C^{1,1}$ at x if f is C^1 at x and the derivative function ∇f is Lipschitzian near x.

2. The unconstrained case. We begin our study of weak sharp minima by examining the case where $C = \mathbb{R}^n$ in (1.1). In this case, we will refer to a weak sharp local minimizer of order m for (1.1) as a *weak sharp local minimizer of order m for f*.

The following concept of normal cone will play a major role in our optimality conditions.

DEFINITION 2.1. *Let $S \subset \mathbb{R}^n$ be nonempty.*

(a) *For $x \in \mathbb{R}^n$, call*

$$P(S, x) := \big\{ w \in \operatorname{cl} S \mid \|x - w\| = \operatorname{dist}(x, S) \big\}.$$

(b) *Let $\bar{x} \in \operatorname{cl} S$. The* normal cone *to S at $\bar{x}$ is defined by*

$$N(S, \bar{x}) := \big\{ y \mid \exists \{y_j\} \to y,\ \{x_j\} \to \bar{x},\ \{t_j\} \subset (0, +\infty),\ \{s_j\} \subset R^n \\ \text{with } s_j \in P(S, x_j) \text{ and } y_j = (x_j - s_j)/t_j \big\}.$$

The normal cone $N(S, \bar{x})$ is often called the Mordukhovich normal cone or limiting proximal normal cone. For information on its properties, see [13], [19], [20], [21], and references therein. In terms of $N(S, \bar{x})$, we can obtain a general characterization of weak sharp local minimality of order m.

THEOREM 2.2. *Let $f : \mathbb{R}^n \to \overline{\mathbb{R}}$ be finite and constant on a closed set $S \subset \mathbb{R}^n$, and let $\bar{x} \in S$ and $m \geq 1$. The following are equivalent:*

(a) *$\bar{x}$ is a weak sharp local minimizer of order m for f.*

(b) *$\forall y \in N(S, \bar{x})$ with $\|y\| = 1$, $\forall$ sequences $\{x_j\} \to \bar{x}$, $\{s_j\}$ with $s_j \in P(S, x_j)$, $\{(x_j - s_j)/\|x_j - s_j\|\} \to y$ and*

$$\liminf_{j \to \infty} \big(f(x_j) - f(s_j) \big) / \|x_j - s_j\|^{m-1} \leq 0,$$

we have

$$\liminf_{j\to\infty} \big(f(x_j) - f(s_j)\big)/\left\|x_j - s_j\right\|^m > 0.$$

(c) $\forall y \in N(S,\bar{x})$ *with* $\|y\| = 1$, $\forall$ *sequences* $\{x_j\} \to \bar{x}$, $\{s_j\}$ *with* $s_j \in P(S,x_j)$ *and* $\{(x_j - s_j)/\left\|x_j - s_j\right\|\} \to y$, *we have*

$$\liminf_{j\to\infty} \big(f(x_j) - f(s_j)\big)/\left\|x_j - s_j\right\|^m > 0.$$

Proof. (a) $\Rightarrow$ (c). Suppose that (a) holds. Let $y \in N(S,\bar{x})$ with $\|y\| = 1$, and let $\{x_j\} \to \bar{x}$, $\{s_j\}$ with $\{(x_j - s_j)/\left\|x_j - s_j\right\|\} \to y$ and $s_j \in P(S,x_j)$. Since S is closed, each $s_j \in S$. By (a), there exists $\beta > 0$ such that for j large enough,

$$f(x_j) - f(s_j) = f(x_j) - f(\bar{x}) \geq \beta \operatorname{dist}^m(x_j,S) = \beta \left\|x_j - s_j\right\|^m.$$

Thus

$$\liminf_{j\to\infty} \big(f(x_j) - f(s_j)\big)/\left\|x_j - s_j\right\|^m \geq \beta,$$

and (c) holds.

(c) $\Rightarrow$ (b). This implication is obvious.

(b) $\Rightarrow$ (a) (by contraposition). Suppose that $\bar{x}$ is not a weak sharp local minimizer of order m for f. Then there exists a sequence $\{x_j\} \to \bar{x}$ such that

$$f(x_j) - f(\bar{x}) < \operatorname{dist}^m(x_j,S)/j.$$

For each j, let $s_j \in P(S,x_j)$. Since S is closed, each $s_j \in S$; and since $\|x_j - s_j\| \leq \|x_j - \bar{x}\|$, we have $\|x_j - s_j\| \to 0$. Taking a subsequence if necessary, we may assume without loss of generality that the sequence $\{(x_j - s_j)/\left\|x_j - s_j\right\|\}$ converges to some y. Then $y \in N(S,\bar{x}), \|y\| = 1$, and

$$f(x_j) - f(s_j) = f(x_j) - f(\bar{x}) < \operatorname{dist}^m(x_j,S)/j = \left\|x_j - s_j\right\|^m/j.$$

Hence

$$\liminf_{j\to\infty} \big(f(x_j) - f(s_j)\big)/\left\|x_j - s_j\right\|^{m-1} \leq 0$$

and

$$\liminf_{j\to\infty} \big(f(x_j) - f(s_j)\big)/\left\|x_j - s_j\right\|^m \leq 0.$$

Therefore (b) does not hold, and the proof is complete. ∎

Theorem 2.2 gives very general, albeit rather complicated, characterizations of weak sharp local minima of order m for f. Are there simpler conditions that imply Theorem 2.2(c) and still provide a characterization in a wide variety of circumstances? Next we propose one answer to this question based on the following concept of directional derivative.

DEFINITION 2.3. *Let S be a nonempty closed subset of $\mathbb{R}^n$, and let $f : \mathbb{R}^n \to \overline{\mathbb{R}}$ be finite on* $\operatorname{bd} S$. *For $x \in \operatorname{bd} S$ and $y \in \mathbb{R}^n$, define*

$$\underline{d}_S^m f(x;y) := \liminf_{\substack{s\to \operatorname{bd} S\, x \\ (t,v)\to(0^+,y)}} \big(f(s+tv) - f(s)\big)/t^m,$$

where "$s \to_{\mathrm{bd}\, S} x$" means that $s \to x$ with each $s \in \mathrm{bd}\, S$. (In particular, (x, y) is an allowable choice of (s, v).)

Definition 2.3 gives a sort of generalization of the directional derivative

$$\underline{d}^m f(x; y) := \liminf_{(t,v)\to(0^+,y)} \big(f(x + tv) - f(x)\big)/t^m,$$

which was used to study strict local minima of order m in [26], [28]. Observe that when $S = \{x\}$, then $\underline{d}^m_S f(x; y)$ reduces to $\underline{d}^m f(x; y)$. It is also worth noting that $\underline{d}^m_S f(x; y)$ and $\underline{d}^m f(x; y)$ may coincide if f is sufficiently smooth.

LEMMA 2.4. *Let $f : \mathbb{R}^n \to \overline{\mathbb{R}}$ be finite on $\mathrm{bd}\, S$, and let $x \in \mathrm{bd}\, S$.*

(a) *If f is strictly differentiable at x with derivative $\nabla f(x)$, then $\forall y \in \mathbb{R}^n$,*

$$\underline{d}^1_S f(x; y) = \langle \nabla f(x), y\rangle = \underline{d}^1 f(x; y).$$

(b) *If f is C^2 at x and there exists $\varepsilon > 0$ with $\nabla f(s) = 0$ $\forall s \in \mathrm{bd}\, S \cap B(x, \varepsilon)$, then $\forall y \in \mathbb{R}^n$,*

$$\underline{d}^2_S f(x; y) = \nabla^2 f(x)(y, y)/2 = \underline{d}^2 f(x; y).$$

Proof. Part (a) follows easily from the definition of strict differentiability. To prove part (b), note that if f is C^2 at x with $\nabla f(s) = 0$ $\forall s \in \mathrm{bd}\, S \cap B(x, \varepsilon)$, then $\forall y \in \mathbb{R}^n$, we have by Taylor's formula that

$$\begin{aligned}\underline{d}^2_S f(x; y) &= \liminf_{\substack{s\to_{\mathrm{bd}\, S} x\\ (t,v)\to(0^+,y)}} \big(f(s + tv) - f(s)\big)/t^2\\ &= \liminf_{\substack{s\to_{\mathrm{bd}\, S} x\\ (t,v)\to(0^+,y)}} \nabla^2 f(s)(v, v)/2 = \nabla^2 f(x)(y, y)/2.\end{aligned}$$

Similarly, $\nabla^2 f(x)(y, y)/2 = \underline{d}^2 f(x; y)$, and so part (b) is true. ∎

The directional derivative $\underline{d}^m_S f(x; \cdot)$ has a number of properties in common with $\underline{d}^m f(x; \cdot)$. In particular, $\underline{d}^m_S f(x; \cdot)$ is l.s.c. as a function of the direction y. From the definition of $\underline{d}^m_S f(x; \cdot)$, it follows easily that if $\underline{d}^m_S f(x; y) > -\infty$ and $\underline{d}^m_S g(x; y) > -\infty$, then

$$\underline{d}^m_S (f + g)(x; y) \geq \underline{d}^m_S f(x; y) + \underline{d}^m_S g(x; y); \tag{2.1}$$

and if in addition there exist $c \in \mathbb{R}$ and $\delta > 0$ such that

$$f(z) = g(z) = c \quad \forall z \in \mathrm{bd}\, S \cap B(\bar{x}, \delta),$$

then for $h(x) := \max\{f(x), g(x)\}$,

$$\underline{d}^m_S h(\bar{x}; y) \geq \max\big\{\underline{d}^m_S f(\bar{x}; y), \underline{d}^m_S g(\bar{x}; y)\big\}. \tag{2.2}$$

We will take advantage of these properties later in this paper.

Making use of Theorem 2.2, we can deduce sufficient conditions for weak sharp minimality in terms of our new directional derivatives.

THEOREM 2.5. *Let $f : \mathbb{R}^n \to \overline{\mathbb{R}}$ be finite and constant on a closed set $S \subset \mathbb{R}^n$, and let $\bar{x} \in \mathrm{bd}\, S$. If $m \geq 1$ and*

$$\underline{d}^m_S f(\bar{x}; y) > 0 \quad \forall y \in N(S, \bar{x})\backslash\{0\}, \tag{2.3}$$

then $\bar{x}$ is a weak sharp local minimizer of order m for f.

Proof. Let $m \geq 1$, and let $y \in N(S, \bar{x})$ with $\|y\| = 1$. Let $\{x_j\} \to \bar{x}$, $\{s_j\}$ with $s_j \in P(S, x_j)$ and $\{(x_j - s_j)/\|x_j - s_j\|\} \to y$. Then $x_j \notin S$, and so $s_j \in \operatorname{bd} S$. Define $t_j := \|x_j - s_j\|$ and $y_j := (x_j - s_j)/t_j$. As noted in the proof of Theorem 2.2, $t_j \to 0^+$. By (2.3), it follows that

$$\begin{aligned}\liminf_{j\to\infty} &\big(f(x_j) - f(s_j)\big)/\|x_j - s_j\|^m \\ &= \liminf_{j\to\infty} \big(f(s_j + t_j y_j) - f(s_j)\big)/t_j^m \geq \underline{d}_S^m f(\bar{x}; y) > 0.\end{aligned}$$

Hence $\bar{x}$ is a weak sharp minimizer of order m for f by Theorem 2.2. ∎

Condition (2.3) is not a general characterization of weak sharp minimality, mainly because Definition 2.3 does not specify that $s \in P(S, s + tv)$. (In the proof that (a) $\Rightarrow$ (c) in Theorem 2.2, the fact that $s_j \in P(S, x_j)$ plays an essential role.) The following example illustrates that the sufficient conditions of Theorem 2.5 are not always necessary conditions.

Example 2.1. Define $f : \mathbb{R}^2 \to \overline{\mathbb{R}}$ by

$$f(x, y) := \begin{cases} x^m & \text{if } x \geq 0,\ 0 \leq y \leq 1, \\ +\infty & \text{otherwise,} \end{cases}$$

and let $S = \{(0, y) \mid 0 \leq y \leq 1\}$. Then each element of S is a weak sharp minimizer of order m for f. When $0 < y < 1$, $N(S, (0, y)) = \{(z, 0) \mid z \in \mathbb{R}\}$. Since

$$\underline{d}_S^m f\big((0, y); (z, 0)\big) = \begin{cases} z^m & \text{if } z \geq 0, \\ +\infty & \text{if } z < 0, \end{cases}$$

(2.3) is satisfied at $\bar{x} = (0, y)$, $0 < y < 1$.

However, $N(S, (0, 0)) = \{(x, y) \mid y \leq 0\}$ and $\underline{d}_S^m\ f((0, 0); (0, -1)) = 0$, so (2.3) does not hold at $\bar{x} = (0, 0)$. Similarly, $N(S, (0, 1)) = \{(x, y) \mid y \geq 0\}$ and $\underline{d}_S^m$ $f((0, 1); (0, 1)) = 0$, so (2.3) is not satisfied at $\bar{x} = (0, 1)$. In general, (2.3) will not hold in situations where S is a convex set that contains more than one point but has empty interior and $\bar{x}$ is not in the relative interior of S.

It is interesting to compare Theorem 2.5 with the necessary conditions for weak sharp minima that were derived in [7], [28]. These necessary conditions are stated in terms of $\underline{d}^m f(\bar{x}; \cdot)$ and the *contingent cone,* which is defined for $C \subset \mathbb{R}^n$ and $\bar{x} \in C$ by

$$K(C, x) := \big\{y \in R^n \mid \exists (t_j, y_j) \to (0^+, y) \text{ with } x + t_j y_j \in C\big\}.$$

In particular, Theorem 2.6 below is a special case of [28, Theorem 4.1] (or, for $m = 1$, [7, Theorem 2.2]).

THEOREM 2.6 (see [28]). *Let $f : \mathbb{R}^n \to \overline{\mathbb{R}}$ be finite and constant on $S \subset \mathbb{R}^n$, and let $\bar{x} \in S$ be a weak sharp local minimizer of order m for f. Then there exists $\beta > 0$ such that*

$$\underline{d}^m f(\bar{x}; y) \geq \beta \operatorname{dist}^m \big(y, K(S, \bar{x})\big) \quad \forall y \in R^n. \tag{2.4}$$

By comparing Theorems 2.5 and 2.6, we can determine assumptions under which condition (2.3) characterizes weak sharp minima.

THEOREM 2.7. *Let $f : \mathbb{R}^n \to \overline{\mathbb{R}}$ be finite and constant on a closed set $S \subset \mathbb{R}^n$, and let $\bar{x} \in \operatorname{bd} S$. If $K(S, \bar{x}) \cap N(S, \bar{x}) = \{0\}$ and*

$$\underline{d}^m f(\bar{x}; y) = \underline{d}_S^m f(\bar{x}; y) \quad \forall y \in N(S, \bar{x}) \backslash \{0\}, \tag{2.5}$$

then $\bar{x}$ is a weak sharp local minimizer of order m for f if and only if (2.3) *holds.*

Proof. Suppose that $\bar{x}$ is a weak sharp minimizer of order m for f, $K(S,\bar{x}) \cap N(S,\bar{x}) = \{0\}$, and $\underline{d}^m f(\bar{x};y) = \underline{d}^m_S f(\bar{x};y)\ \forall y \in N(S,\bar{x})\backslash\{0\}$. Let $y \in N(S,\bar{x})\backslash\{0\}$. Then $y \notin K(S,\bar{x})$, and by (2.4), $\underline{d}^m f(\bar{x};y) > 0$. Hence $\underline{d}^m_S f(\bar{x};y) > 0$ and so (2.3) holds. On the other hand, (2.3) implies that $\bar{x}$ is a weak sharp minimizer of order m for f by Theorem 2.5. □

Remark 2.1. (a) The assumption

$$K(S,\bar{x}) \cap N(S,\bar{x}) = \{0\} \tag{2.6}$$

in Theorem 2.7 holds, in particular, if S is convex. Indeed, if S is convex, then $N(S,\bar{x}) = K(S,\bar{x})^0$ [13], [19], [20], so that if $y \in K(S,\bar{x}) \cap N(S,\bar{x})$, it follows that $\langle y,y\rangle \le 0$. If S is the set of all global minimizers of f, then S will be convex, for example, whenever f is quasi-convex.

However, (2.6) may not hold if S is not convex. For example, if $S = \{(x,y) \mid y = -|x|\}$ and $\bar{x} = (0,0)$, then $S \subset K(S,\bar{x}) \cap N(S,\bar{x})$.

(b) In Example 2.1, we saw that (2.3) does not hold at the points $(0,0)$ and $(0,1)$. Observe that (2.5) is not satisfied at these points since

$$\underline{d}^m_S f\big((0,0);(0,-1)\big) = \underline{d}^m_S f\big((0,1);(0,1)\big) = 0,$$

while

$$\underline{d}^m f\big((0,0);(0,-1)\big) = \underline{d}^m f\big((0,1);(0,1)\big) = +\infty.$$

3. Optimality conditions for constrained problems. In this section, we apply the results of section 2 to deduce optimality conditions for problems with inequality, equality, and abstract set constraints, concentrating on sufficient conditions for weak sharp minima of orders one and two. In stating these conditions, we will use a variation of the contingent cone that takes the set S into account. For $C \subset \mathbb{R}^n$ and $x \in C$, we define this new tangent cone by

$$K_S(C,x) := \big\{y \mid \exists (t_j,y_j) \to (0^+,y),\ \exists x_j \to_{\mathrm{bd}\, S} x \text{ with } x_j + t_j y_j \in C\big\}.$$

Observe that when $S = \{x\}$, then $K_S(C,x) = K(C,x)$. It is also easy to see that $K_S(C,x)$ is defined precisely so that if S is closed and $\bar{x} \in \mathrm{bd}\, S$, then

$$\underline{d}^m_S i_C(\bar{x};\cdot) = i_{K_S(C,\bar{x})}(\cdot) \quad \forall m \ge 1. \tag{3.1}$$

We can use Theorem 2.5 and our new tangent cone to derive sufficient conditions for weak sharp minimality in (1.1).

THEOREM 3.1. *Let $C \subset \mathbb{R}^n$, and let $f : \mathbb{R}^n \to \overline{\mathbb{R}}$ be finite and constant on a closed set S with $\bar{x} \in \mathrm{bd}\, S \cap C$. If $\underline{d}^m_S f(\bar{x};\cdot)$ is proper and*

$$\underline{d}^m_S f(\bar{x};y) > 0 \quad \forall y \in \big(N(S,\bar{x}) \cap K_S(C,\bar{x})\big)\backslash\{0\}, \tag{3.2}$$

then $\bar{x}$ is a weak sharp local minimizer of order m for (1.1).

Proof. Suppose that (3.2) holds, and let $y \in N(S,\bar{x})\backslash\{0\}$. Then by (3.1) and (3.2), we have

$$\underline{d}^m_S (f + i_C)(\bar{x};y) \ge \underline{d}^m_S f(\bar{x};y) + i_{K_S(C,\bar{x})}(y) > 0; \tag{3.3}$$

and so by Theorem 2.5, $\bar{x}$ is a weak sharp local minimizer of order m for $f+i_C$. Hence $\bar{x}$ is a weak sharp local minimizer of order m for (1.1). □

We next present sufficient conditions for weak sharp minimality of order one in the problem

$$\min\big\{f(x) \mid g_i(x) \le 0,\ i \in J,\ h_i(x) = 0,\ i \in L,\ x \in \Omega\big\}, \tag{3.4}$$

where Ω is a nonempty subset of $\mathbb{R}^n$, $J := \{1,\dots,p\}$, $L := \{1,\dots,q\}$, $g_i : \mathbb{R}^n \to \overline{\mathbb{R}}$, $i \in J$, and $h_i : \mathbb{R}^n \to \overline{\mathbb{R}}$, $i \in L$. In (3.4), we define

$$C := \big\{x \in \mathbb{R}^n \mid g_i(x) \le 0,\ i \in J,\ h_i(x) = 0,\ i \in L,\ x \in \Omega\big\}$$

and let Γ be a closed subset of C on which f is finite and constant, with $f(x) = \alpha$ $\forall x \in \Gamma$. For $x \in C$, we single out the set of indices

$$I(x) := \big\{i \in J \mid g_i(x) = 0\big\};$$

and for $\bar{x} \in \Gamma$, we define

$$I^*(\bar{x}) := \big\{i \in I(\bar{x}) \mid \exists \delta > 0 \text{ with } i \in I(y)\ \forall y \in B(\bar{x},\delta) \cap \Gamma\big\}$$

and

$$S(\bar{x}) := \big\{x \mid g_i(x) = 0,\ i \in I^*(\bar{x}),\ h_i(x) = 0,\ i \in L,\ f(x) = \alpha\big\}.$$

Our sufficient conditions can now be stated as follows.

THEOREM 3.2. *Let $\bar{x} \in \Gamma$, and let $S \subset S(\bar{x})$ be closed with $\bar{x} \in \operatorname{bd} S$. Suppose that h_i is strictly differentiable at $\bar{x}$ for each $i \in L$. If $\underline{d}^1_S f(\bar{x};\cdot)$ is proper and*

$$\begin{aligned} &\underline{d}^1_S f(\bar{x};y) > 0 \quad \forall y \in \big(N(S,\bar{x}) \cap K_S(\Omega,\bar{x})\big) \setminus \{0\} \text{ with} \\ &\underline{d}^1_S g_i(\bar{x};y) \le 0 \quad \forall i \in I^*(\bar{x}), \langle \nabla h_i(\bar{x}), y\rangle = 0\ \forall i \in L, \end{aligned} \tag{3.5}$$

then $\bar{x}$ is a weak sharp local minimizer of order one for (3.4)*).*

Proof. The result will follow directly from Theorem 3.1 and the inclusion

$$K_S(C,\bar{x}) \subset \big\{y \in K_S(\Omega,\bar{x}) \mid \underline{d}^1_S g_i(\bar{x};y) \le 0\ \forall i \in I^*(\bar{x}), \langle \nabla h_i(\bar{x}), y\rangle = 0\ \forall i \in L\big\}.$$

To verify this inclusion, let $y \in K_S(C,\bar{x})$. Then there exist sequences $\{x_j\} \subset \operatorname{bd} S$ with $\{x_j\} \to \bar{x}, \{t_j\} \to 0^+$, and $\{y_j\} \to y$ such that $x_j + t_j y_j \in C$. In particular, $x_j + t_j y_j \in \Omega$, and so $y \in K_S(\Omega,\bar{x})$. For $i \in L$, $h_i(x_j + t_j y_j) = 0$ and $h_i(x_j) = 0$ since $S \subset S(\bar{x})$, and because h_i is strictly differentiable at $\bar{x}$, we have $\langle \nabla h_i(\bar{x}), y\rangle = 0$. Finally, for $i \in I^*(\bar{x})$, $g_i(x_j) = 0$ and $g_i(x_j + t_j y_j) \le 0$ $\forall j$, so that $\underline{d}^1_S g_i(\bar{x};y) \le 0$. □

Our next example illustrates Theorem 3.2, highlighting the importance of the choice of S in condition (3.5).

Example 3.1. In problem (3.4), let $n = 2$, $\Omega = \mathbb{R}^2$, $J = \{1,2\}$, $L = \emptyset$, and define $f(x,y) = x$, $g_1(x,y) = -x$, $g_2(x,y) = -y$. Then each element of $S = \{(0,y) \mid y \ge 0\}$ is a weak sharp minimizer of order one for (3.4).

For $y > 0$ and $\bar{x} = (0,y)$, $N(S,\bar{x}) = \{(z,0) \mid z \in \mathbb{R}\}$, $I(\bar{x}) = I^*(\bar{x}) = \{1\}$, and $\underline{d}^1_S g_1(\bar{x};(z,0)) \le 0$ for $z \ge 0$. Since

$$\underline{d}^1_S f\big(\bar{x};(z,0)\big) = z > 0 \quad \forall z > 0,$$

(3.5) holds for $\bar{x} = (0, y)$, $y > 0$. However, for $\bar{x} = (0, 0)$, $N(S, \bar{x}) = \{(x, y) \mid y \le 0\}$, while $I(\bar{x}) = \{1, 2\}$ and $I^*(\bar{x}) = \{1\}$. Since $(0, -1) \in N(S, \bar{x})$ and $\underline{d}^1_S\, g_1(\bar{x}; (0, -1)) = 0$, while $\underline{d}^1_S\, f(\bar{x}; (0, -1)) = 0$, (3.5) is not satisfied at $\bar{x}$.

On the other hand, if we let $S := \{0\} \times \mathbb{R}$, then (3.5) holds with $\bar{x} = (0, y)$ for any $y \ge 0$.

One can also prove sufficient conditions for weak sharp local minima of order two in terms of Lagrangian functions. In the remainder of this section, we present such conditions.

We first set some notation. For $\lambda_i \ge 0$, $i \in J$, $\mu_i \in \mathbb{R}$, $i \in L$, we define the Lagrangian function

$$L(x) := f(x) + \sum_{i \in J} \lambda_i g_i(x) + \sum_{i \in L} \mu_i h_i(x)$$

and single out a set of directions

$$D^*(\bar{x}) := \big\{ v \mid \langle \nabla f(\bar{x}), v \rangle \le 0, \langle \nabla g_i(\bar{x}), v \rangle \le 0 \ \forall i \in I^*(\bar{x}), \langle \nabla h_i(\bar{x}), v \rangle = 0 \ \forall i \in L \big\}$$

and a set of multipliers

$$\Lambda^*(\bar{x}) := \big\{ (\lambda, \mu) \in \mathbb{R}^p \times \mathbb{R}^q \mid \nabla L(\bar{x}) = 0,\ \lambda_i \ge 0,\ i \in J,\ \lambda_i = 0 \ \forall i \notin I^*(\bar{x}) \big\}.$$

In terms of these concepts, we can derive second-order sufficient conditions for problem (3.4) with $\Omega = \mathbb{R}^n$.

THEOREM 3.3. *In problem* (3.4)*, let* $\Omega = \mathbb{R}^n$*. Let* $S \subset \Gamma$ *be closed, and suppose that* f, g_i*,* $i \in J$*, and* h_i*,* $i \in L$*, are strictly differentiable at* $\bar{x} \in \operatorname{bd} S$*. If there exists* $(\lambda, \mu) \in \Lambda(\bar{x})$ *with*

$$\underline{d}^2_S L(\bar{x}; y) > 0 \quad \forall y \in \big(N(S, \bar{x}) \cap D^*(\bar{x})\big) \backslash \{0\}, \tag{3.6}$$

then $\bar{x}$ *is a weak sharp local minimizer of order two for* (3.4)*.*

Proof. We prove the contrapositive. If $\bar{x}$ is not a weak sharp local minimizer of order two for (3.4), there exists a sequence $\{x_j\} \to_C \bar{x}$ such that

$$f(x_j) - f(\bar{x}) < \operatorname{dist}^2(x_j, S)/j.$$

Then each $x_j \notin S$, so there exists $s_j \in \operatorname{bd} S$ with $s_j \in P(S, x_j)$. Call $t_j := \|x_j - s_j\|$ and $y_j := (x_j - s_j)/t_j$. As in the proof of Theorem 2.2, $\{t_j\} \to 0^+$; and we may assume, taking a subsequence if necessary, that $\{y_j\} \to y$ for some $y \ne 0$. Then $y \in N(S, \bar{x}) \backslash \{0\}$. For j large enough, we have

$$\begin{aligned} \big(g_i(x_j) - g_i(s_j)\big)/t_j &\le 0 \quad \forall i \in I^*(\bar{x}), \\ \big(h_i(x_j) - h_i(s_j)\big)/t_j &= 0 \quad \forall i \in L, \end{aligned}$$

and

$$\big(f(x_j) - f(s_j)\big)/t_j < t_j/j.$$

Thus $y \in D^*(\bar{x})$. Now let $\lambda \in \Lambda^*(\bar{x})$. Then for j large enough,

$$\big(L(x_j) - L(s_j)\big)/t_j^2 \le \big(f(x_j) - f(s_j)\big)/t_j^2 \le 1/j.$$

Therefore

$$\underline{d}_S^2 L(\bar{x}; y) \le \liminf_{j\to\infty} \big(L(x_j) - L(s_j)\big)/t_j^2 \le 0. \qquad \square$$

By comparing Theorem 3.3 with the necessary conditions given in [28, section 4], we can formulate a characterization of weak sharp local minimality of order two. To do so, we first define the set of multipliers

$$\Lambda(\bar{x}) := \big\{(\lambda,\mu) \in \mathbb{R}^p \times \mathbb{R}^q \mid \nabla L(\bar{x}) = 0,\ \lambda_i \ge 0,\ \lambda_i g_i(\bar{x}) = 0\ \forall i \in J\big\};$$

and for $(\lambda,\mu) \in \Lambda(\bar{x})$ we single out the sets of indices

$$M(\bar{x}) := \big\{i \in I(\bar{x}) \mid \lambda_i > 0\big\}$$

and

$$N(\bar{x}) := \big\{i \in I(\bar{x}) \mid \lambda_i = 0\big\},$$

and define the set of directions

$$D(\bar{x}) := \big\{v \mid \langle \nabla g_i(\bar{x}), v\rangle = 0\ \forall i \in M(\bar{x}),\ \langle \nabla g_i(\bar{x}), v\rangle \le 0\ \forall i \in N(\bar{x}),\ \langle \nabla h_i(\bar{x}), v\rangle = 0\ \forall i \in L\big\}.$$

Our characterization involves the following constraint qualification.

DEFINITION 3.4. *Let f, g_i, $i \in J$, h_i, $i \in L$, be C^1 at $\bar{x}$, and suppose that $(\lambda,\mu) \in \Lambda(\bar{x})$. We say that the strict Mangasarian–Fromovitz constraint qualification (SMFCQ) is satisfied at $(\bar{x},\lambda,\mu)$ if*

(i) *$\nabla g_i(\bar{x})$, $i \in M(\bar{x})$, $\nabla h_i(\bar{x})$, $i \in L$, are linearly independent;*

(ii) *there exists $z \in \mathbb{R}^n$ such that*

$$\langle \nabla g_i(\bar{x}), z\rangle < 0,\ i \in N(\bar{x});\ \langle \nabla g_i(\bar{x}), z\rangle = 0,\ i \in M(\bar{x});\ \langle \nabla h_i(\bar{x}), z\rangle = 0,\ i \in L.$$

Observe that if $I(\bar{x}) = M(\bar{x})$, then condition SMFCQ coincides with the familiar linear independence constraint qualification. Kyparisis [16] has shown that SMFCQ is necessary and sufficient for $\Lambda(\bar{x})$ to be a singleton. In [28], Ward used this constraint qualification in developing necessary conditions for weak sharp minimality of order two. In particular, Proposition 3.5 below is a special case of [28, Corollary 4.2] (taking into account [28, Lemma 3.1(i)]).

PROPOSITION 3.5. *Let $\Omega = \mathbb{R}^n$ in (3.4), let $S \subset \Gamma$ be closed, and let $\bar{x} \in \operatorname{bd} S$ be a weak sharp local minimizer of order two for (3.4). Suppose that f, g_i, $i \in J$, and h_i, $i \in L$, are $C^{1,1}$ at $\bar{x}$, and assume that SMFCQ is satisfied at $(\bar{x},\lambda,\mu)$. Then there exists $\beta > 0$ such that*

$$\underline{d}^2 L(\bar{x}; y) \ge \beta \operatorname{dist}^2\big(y, K(S,\bar{x})\big) \quad \forall y \in D(\bar{x}). \tag{3.7}$$

Theorem 3.3 and Proposition 3.5 yield the following characterization of local weak sharp minimality of order two for a program with $C^{1,1}$ data.

THEOREM 3.6. *In problem (3.4), let $\Omega = \mathbb{R}^n$, let $S \subset \Gamma$ be closed, and suppose that f, g_i, $i \in J$, and h_i, $i \in L$, are $C^{1,1}$ at $\bar{x} \in \operatorname{bd} S$. Assume that $K(S,\bar{x}) \cap N(S,\bar{x}) = \{0\}$, $D(\bar{x}) = D^*(\bar{x})$, SMFCQ is satisfied at $(\bar{x},\lambda,\mu)$ with $(\lambda,\mu) \in \Lambda^*(\bar{x})$,*

and $\underline{d}^2_S L(\bar{x};\cdot) = \underline{d}^2 L(\bar{x};\cdot)$. *Then* $\bar{x}$ *is a weak sharp local minimizer of order two for* (3.4) *if and only if*

$$\underline{d}^2 L(\bar{x};y) > 0 \quad \forall y \in \big(N(S,\bar{x}) \cap D(\bar{x})\big)\backslash\{0\}. \tag{3.8}$$

Proof. By our assumptions and Theorem 3.3, (3.8) implies that $\bar{x}$ is a weak sharp local minimizer of order two . On the other hand, suppose that $\bar{x}$ is a weak sharp local minimizer of order two for (3.4), and let $y \in (N(S,\bar{x}) \cap D(\bar{x}))\backslash\{0\}$. Since $N(S,\bar{x}) \cap K(S,\bar{x}) = \{0\}$, $y \notin K(S,\bar{x})$, and by (3.7), $\underline{d}^2 L(\bar{x};y) > 0$. Therefore (3.8) holds. □

In the case in which f, g_i, and h_i are C^2, Theorem 3.6 reduces to a slightly simpler result.

COROLLARY 3.7. *In problem* (3.4), *let* $\Omega = \mathbb{R}^n$. *Let* $S \subset \Gamma$ *be closed, and suppose that* f, g_i, $i \in J$, *and* h_i, $i \in L$, *are* C^2 *at* $\bar{x} \in \operatorname{bd} S$. *Assume that* $K(S,\bar{x}) \cap N(S,\bar{x}) = \{0\}$, $D(\bar{x}) = D^*(\bar{x})$, *SMFCQ is satisfied at* $(\bar{x},\lambda,\mu)$ *with* $(\lambda,\mu) \in \Lambda^*(\bar{x})$, *and there exists* $\delta > 0$ *such that* $\nabla L(s) = 0$ $\forall s \in \operatorname{bd} S \cap B(\bar{x},\delta)$. *Then* $\bar{x}$ *is a weak sharp local minimizer of order two for* (3.4) *if and only if*

$$\nabla^2 L(\bar{x})(y,y) > 0 \quad \forall y \in \big(N(S,\bar{x}) \cap D(\bar{x})\big)\backslash\{0\}. \tag{3.9}$$

Proof. This result follows immediately from Theorem 3.6 and Lemma 2.4(b). □

Remark 3.1. In Theorem 3.6 and Corollary 3.7, one condition under which $D(\bar{x}) = D^*(\bar{x})$ and $\Lambda(\bar{x}) = \Lambda^*(\bar{x})$ is the assumption that $I(\bar{x}) = I^*(\bar{x})$, since

$$D(\bar{x}) = \big\{v \mid \langle \nabla f(\bar{x}), v\rangle \le 0, \langle \nabla g_i(\bar{x}), v\rangle \le 0 \ \forall i \in I(\bar{x}), \langle \nabla h_i(\bar{x}), v\rangle = 0 \ \forall i \in L\big\}$$

by [11, Theorem 3.5].

We conclude this section with a discussion of an interesting special case of Theorem 3.2. In problem (3.4), let $\Omega = R^n$, $g_i(x) = \langle a_i, x\rangle - b_i$, $i \in J$, and $h_i(x) = \langle c_i, x\rangle - d_i$, $i \in L$, where each $a_i, c_i \in \mathbb{R}^n$, $b_i, d_i \in \mathbb{R}$; and suppose that f is constant on the set

$$S = \big\{x \in \mathbb{R}^n \mid \langle a_i, x\rangle = b_i,\ i \in J_1, \langle c_i, x\rangle = d_i,\ i \in L\big\} \tag{3.10}$$

for some $J_1 \subset J$. In this case of linear constraints, we have the following result.

PROPOSITION 3.8. *In problem* (3.4), *define* C *as above and suppose* f *is constant on the set* S *in* (3.10). *Let* $\bar{x} \in \Gamma = S \cap C$ *be such that* $I^*(\bar{x}) = J_1$, *and assume that* a_i, $i \in J_1$, c_i, $i \in L$, *are linearly independent. Suppose that* f *is strictly differentiable at* $\bar{x}$ *and there exists* $\lambda \in \Lambda(\bar{x})$ *such that* $M(\bar{x}) = J_1$. *Then* $\bar{x}$ *is a weak sharp local minimizer of order one for* (3.4).

Proof. Let $y \in N(S,\bar{x})$ be such that $\underline{d}^1_S f(\bar{x};y) \le 0$, $\underline{d}^1_S g_i(\bar{x};y) \le 0$ $\forall i \in J_1$, and $\langle \nabla h_i(\bar{x}), y\rangle = 0$ $\forall i \in L$. Since there exists $\lambda \in \Lambda(\bar{x})$ with $M(\bar{x}) = J_1$, it follows (in view of Lemma 2.4) that $\langle a_i, y\rangle = 0$ $\forall i \in J_1$ and $\langle c_i, y\rangle = 0$ $\forall i \in L$. Let A be a matrix whose rows are a_i, $i \in J_1$, and c_i, $i \in L$. Then $Ay = 0$. In addition, it is easy to calculate that $N(S,\bar{x}) = \operatorname{Range} A^T$, and so $y = A^T w$ for some w. Hence

$$0 = Ay = AA^T w,$$

and since A has linearly independent rows, we conclude that $w = 0$. Therefore $y = 0$, and (3.5) holds. By Theorem 3.2, $\bar{x}$ is a weak sharp local minimizer of order one for (3.4). □

We illustrate Proposition 3.8 with an example.

Example 3.2. Consider the problem

$$\min\{x^2 - 2y - 3z \mid x + 2y + 3z \le 6,\ x \ge 0,\ y \ge 0,\ z \ge 0\}.$$

Here the objective function is clearly constant on the set

$$S = \{(x, y, z) \mid x + 2y + 3z = 6,\ x = 0\}.$$

Let $\bar{x} = (0, y, z)$ be an element of $S \cap C$. It is easy to verify that $(1, 1, 0, 0) \in \Lambda(\bar{x})$, giving $M(\bar{x}) = J_1 = \{1, 2\}$. Since

$$A = \begin{bmatrix} 1 & 2 & 3 \\ -1 & 0 & 0 \end{bmatrix}$$

has independent rows, the hypotheses of Proposition 3.8 are satisfied, and we conclude that $\bar{x}$ is a weak sharp local minimizer of order one for this problem.

4. Sufficient conditions and exactness of penalties. One way to approach the solution of constrained optimization problems is via associated unconstrained problems. For example, under fairly general conditions, a minimizer of

$$\min\{f(x) \mid g_i(x) \le 0,\ i \in J,\ x \in \Omega\} \tag{4.1}$$

will also be a minimizer of the problem

$$\min\left\{f(x) + p\sum_{i\in J} \max(g_i(x), 0) \mid x \in \Omega\right\} \tag{4.2}$$

for sufficiently large values of $p > 0$. (See [6], [9], [11], and references therein.) When minimizers of (4.1) and (4.2) are related in this way, the *penalty function*

$$\theta_p(x) := f(x) + p\sum_{i\in J} \max\big(g_i(x), 0\big)$$

is said to be *exact*.

One type of hypothesis that often guarantees exactness of θ_p is a sufficient condition for strict local minimality of order m. (See, for instance, [6], [11], [12], [22], [27].) In this section, we prove similar results for weak sharp minimality. First, we generalize [27, Theorem 4.1] for $m = 1$ and the penalty function θ_p by showing that condition (3.5) of Theorem 3.2 (in the case $L = \emptyset$) implies exactness of θ_p.

In this section, we let $L = \emptyset$ in (3.4) and let $S \subset \Gamma$ be closed. To state our main result, we introduce notation similar to that of [22] and [27], defining

$$K^*(\bar{x}) := \big\{y \in N(S, \bar{x}) \cap K_S(\Omega, \bar{x}) \mid \|y\| = 1,\ \underline{d}^1_S f(\bar{x}; y) \le 0\big\},$$
$$b(\bar{x}) := \min\big\{\underline{d}^1_S f(\bar{x}; y) \mid \|y\| = 1,\ y \in N(S, \bar{x}) \cap K_S(\Omega, \bar{x})\big\},$$

and

$$a(\bar{x}) := \min\left\{\sum_{I^*(\bar{x})} \max\big(0, \underline{d}^1_S g_i(\bar{x}; y)\big) \mid y \in K^*(\bar{x})\right\}.$$

We note that since $\underline{d}^1_S\, f(\bar{x}; \cdot)$ is l.s.c., the set $K^*(\bar{x})$ is compact, although possibly empty. It follows that the minima in the definitions of $b(\bar{x})$ and $a(\bar{x})$ are attained if the constraint sets in those definitions are nonempty.

We can now prove a result on exactness of θ_p.

THEOREM 4.1. *Let $\bar{x} \in \operatorname{bd} S$, and suppose that $\underline{d}^1_S g_i(\bar{x};\cdot)$ is proper for each $i \in I(\bar{x})$. Define*

$$\rho_0 := \begin{cases} -b(\bar{x})/a(\bar{x}) & \text{if } K^*(\bar{x}) \neq \emptyset, \\ 0 & \text{else.} \end{cases}$$

Suppose that $\underline{d}^1_S f(\bar{x};\cdot)$ is proper and

$$\underline{d}^1_S f(\bar{x};y) > 0 \ \forall y \in \big(N(S,\bar{x}) \cap K_S(\Omega,\bar{x})\big)\backslash\{0\} \text{ with } \underline{d}^1_S g_i(\bar{x};y) \leq 0 \ \forall i \in I^*(\bar{x}). \tag{4.3}$$

Then $\forall p > \rho_0$, $\bar{x}$ is a weak sharp local minimizer of order one for (4.2).

Proof. We first observe that if (4.3) is satisfied and $K^*(\bar{x}) \neq \emptyset$, then $a(\bar{x}) > 0$ and $b(\bar{x}) \leq 0$, so that ρ_0 is well defined and nonnegative. To prove our assertion, let $p > \rho_0$, and let $y \in (N(S,\bar{x}) \cap K_S(\Omega,\bar{x}))\backslash\{0\}$. By Theorem 3.1, it suffices to show that $\underline{d}^1_S \theta_p(\bar{x};y) > 0$. Without loss of generality, we may assume that $\|y\| = 1$. Since $\underline{d}^1_S f(\bar{x};\cdot)$ and $\underline{d}^1_S g_i(\bar{x};\cdot)$, $i \in I(\bar{x})$, are proper and

$$\underline{d}^1_S \max\{g_i, 0\}(\bar{x};y) \geq 0 \quad \forall i \in I(\bar{x})\backslash I^*(\bar{x}),$$

it follows from (2.1) and (2.2) that

$$\underline{d}^1_S \theta_p(\bar{x};y) \geq \underline{d}^1_S f(\bar{x};y) + p \sum_{I^*(\bar{x})} \max\big(0, \underline{d}^1_S g_i(\bar{x};y)\big). \tag{4.4}$$

If $K^*(\bar{x}) = \emptyset$, it follows from (4.4) that $\underline{d}^1_S \theta_p(\bar{x};y) > 0$. If $K^*(\bar{x}) \neq \emptyset$, then if $\underline{d}^1_S f(\bar{x};y) > 0$, we again have $\underline{d}^1_S\ \theta\ _p(\bar{x};y) > 0$; while if $y \in K^*(\bar{x})$, the definition of ρ_0 implies that

$$p \sum_{I^*(\bar{x})} \max\big(0, \underline{d}^1_S g_i(\bar{x};y)\big) > -b(\bar{x}),$$

and hence

$$\underline{d}^1_S \theta_p(\bar{x};y) > \underline{d}^1_S f(\bar{x};y) - b(\bar{x}) \geq 0.$$

Therefore $\bar{x}$ is a weak sharp local minimizer of order 1 for (4.2). ∎

Theorem 4.1 generalizes part of [27, Theorem 4.1] to cover the case where S is not necessarily a singleton. We illustrate Theorem 4.1 with an example in which [27, Theorem 4.1] is not applicable.

Example 4.1. In problem (4.1), let $n = 2$, $\Omega = \mathbb{R}^2$, $J = \{1\}$, and define $f(x,y) = x$, $g_1(x,y) = -x + 1$. Then each element of $S = \{(1,y) \mid y \in \mathbb{R}\}$ is a weak sharp minimizer of order 1 for (4.1). For $\bar{x} \in S$, $N(S,\bar{x}) = \{(z,0) \mid z \in \mathbb{R}\}$, $I(\bar{x}) = I^*(\bar{x}) = \{1\}$, and $\underline{d}^1_S g_1(\bar{x};(z,0)) \leq 0$ for $z \geq 0$. Since $\underline{d}^1_S f(\bar{x};(z,0)) = z\ \forall z > 0$, (4.3) holds. It is easy to calculate that $K^*(\bar{x}) = \{(-1,0)\}$, $b(\bar{x}) = -1$, and $a(\bar{x}) = 1$, so by Theorem 4.1, each $\bar{x} \in S$ is a weak sharp local minimizer of order one for (4.2) whenever $p > 1$.

Analogues of Theorem 4.1 can be formulated for other nonsmooth penalty functions of the class considered in [27]. Moreover, it is possible to accommodate equality constraints in Theorem 4.1 by treating them as pairs of inequality constraints.

The second-order sufficient conditions of Theorem 3.3 also imply exactness of θ_p. We now demonstrate this for problem (4.1) with $\Omega = \mathbb{R}^n$.

THEOREM 4.2. *In problem* (4.1)*, let* $\Omega = \mathbb{R}^n$*. Let* $S \subset \Gamma$ *be closed, and suppose that* f *and* g_i*,* $i \in J$*, are strictly differentiable at* $\bar{x} \in \operatorname{bd} S$ *and* $\lambda \in \Lambda^*(\bar{x})$*. If*

$$\underline{d}^2_S L(\bar{x}; y) > 0 \quad \forall y \in \big(N(S, \bar{x}) \cap D^*(\bar{x})\big) \backslash \{0\}, \tag{4.5}$$

then $\bar{x}$ *is a weak sharp local minimizer of order two for* θ_p $\forall p > \max_{i \in J} \lambda_i$*.*

Proof. Let $p > \max_{i \in J} \lambda_i$ and $y \in N(S, \bar{x}) \backslash \{0\}$. By Theorem 2.5, it suffices to show that $\underline{d}^2_S \theta_p(\bar{x}; y) > 0$. Since $\underline{d}^2_S \theta_p(\bar{x}; y) \geq \underline{d}^1_S \theta_p(\bar{x}; y)$, we may assume that $\underline{d}^1_S \theta_p(\bar{x}; y) \leq 0$. Then, as in (4.4), we have

$$0 \geq \underline{d}^1_S \theta_p(\bar{x}; y) \geq \langle \nabla f(\bar{x}), y \rangle + p \sum_{I^*(\bar{x})} \max\big(0, \langle \nabla g_i(\bar{x}), y \rangle\big),$$

and so $\langle \nabla f(\bar{x}), y \rangle \leq 0$. Since $\nabla L(\bar{x}) = 0$, it follows that

$$\sum_{I^*(\bar{x})} \lambda_i \langle \nabla g_i(\bar{x}), y \rangle \geq p \sum_{I^*(\bar{x})} \max\big(0, \langle \nabla g_i(\bar{x}), y \rangle\big). \tag{4.6}$$

Observe that (4.6) and the fact that $p > \max_{i \in J} \lambda_i$ imply that $y \in D^*(\bar{x})$. By (4.5), $\underline{d}^2_S L(\bar{x}; y) > 0$. Finally, note that since $\lambda \in \Lambda^*(\bar{x})$, we have $L(s) = \theta_p(s) = f(\bar{x})$ for $s \in \operatorname{bd} S$ sufficiently close to $\bar{x}$. Hence

$$\underline{d}^2_S \theta_p(\bar{x}; y) \geq \underline{d}^2_S L(\bar{x}; y) > 0,$$

and the proof is complete. □

Theorem 4.2 is a weak sharp analogue of results for strict local minima of order two like [6, Theorem 4.7] and the related theorems cited on page 986 of [6].

5. More sufficient conditions for weak sharp minima of order one. In this section, we present sufficient conditions for weak sharp minimality of order one which are different from those of sections 2 and 3. These conditions involve the Mordukhovich subdifferential [18], [19], [20], [21], [13] instead of the directional derivative $\underline{d}^1_S f(x; \cdot)$. We begin by reviewing some essential definitions.

DEFINITION 5.1. *Let* $f : \mathbb{R}^n \to \overline{\mathbb{R}}$ *be finite at* $x \in \mathbb{R}^n$*.*

(a) *The* Mordukhovich subdifferential *(or* approximate subdifferential*) of* f *at* x *is the set*

$$\partial f(x) := \big\{x^* \in \mathbb{R}^n \mid (x^*, -1) \in N\big(\operatorname{epi} f, (x, f(x))\big)\big\}.$$

(b) *The* singular Mordukhovich subdifferential *of* f *at* x *is the set*

$$\partial^\infty f(x) := \big\{x^* \in \mathbb{R}^n \mid (x^*, 0) \in N\big(\operatorname{epi} f, (x, f(x))\big)\big\}.$$

Much information about ∂f and $\partial^\infty f$ is given in [18], [19], [20], [21], [13], and references therein. We mention here that if f is convex, then ∂f coincides with the subdifferential of convex analysis, and if f is strictly differentiable at x, then $\partial f(x) = \{\nabla f(x)\}$. The singular subdifferential $\partial^\infty f$ is used to describe conditions under which calculus rules for ∂f are valid; an important thing to remember about $\partial^\infty f$ is that $\partial^\infty f(x) = \{0\}$ if and only if f is Lipschitzian near x.

We now give sufficient conditions for weak sharp minimality of order one in the unconstrained case. Our proof makes use of a technique of [2, Theorem 1.1].

THEOREM 5.2. *Let $f : \mathbb{R}^n \to \mathbb{R}$ be continuous, and let $\bar{x} \in \mathbb{R}^n$. Suppose that there exists $\delta > 0$ such that $f(\bar{x}) \le f(x)\ \forall x \in B(\bar{x}, \delta)$. Define $S := \{x \in B(\bar{x}, \delta) \mid f(x) = f(\bar{x})\}$. Suppose that there exists $c > 0$ such that*

$$\|z\| \ge c \quad \forall z \in \partial f(x),\ \forall x \in B(\bar{x}, \delta) \backslash S. \tag{5.1}$$

Then there exists $\mu \in (0, \delta)$ such that

$$f(x) \ge f(\bar{x}) + c\,\mathrm{dist}(x, S) \quad \forall x \in B(\bar{x}, \mu). \tag{5.2}$$

Proof. Let $F(x) := f(x) - f(\bar{x})$. For $\rho \in [0, \delta]$, define

$$h(\rho) := \max\{F(x) \mid x \in B(\bar{x}, \rho)\}/c.$$

Since F is continuous, there exists $\rho^* \in (0, \delta)$ such that

$$F(x) \le F(\bar{x}) + c\delta/2 \quad \forall x \in B(\bar{x}, \rho^*),$$

and so

$$h(\rho) \le \delta/2 \quad \forall \rho \in [0, \rho^*]. \tag{5.3}$$

We claim that (5.2) is true with $\mu := \rho^*/4$. Indeed, if (5.2) does not hold with $\mu := \rho^*/4$, there exists $u \in B(\bar{x}, \rho^*/4)$ such that

$$\mathrm{dist}(u, S) > F(u)/c,$$

and so for some $t \in (1, 1.5)$,

$$\mathrm{dist}(u, S) > \gamma := tF(u)/c.$$

Since $F(u) \ge 0$, we have $\mathrm{dist}(u, S) > 0$, and it follows that $u \notin S$. Hence $\gamma > 0$. Moreover, (5.3) and the definition of h imply that

$$\gamma \le th(\rho^*/4) \le t\delta/2 < .75\delta. \tag{5.4}$$

Now since $\gamma c/t = F(u)$ and $F(x) \ge 0\ \forall x \in B(\bar{x}, \delta)$, we have

$$F(u) \le \inf\{F(x) \mid x \in B(\bar{x}, \delta)\} + \gamma c/t.$$

Applying Ekeland's variational principle [8, Theorem 7.5.1] to the l.s.c. function F with $V := B(\bar{x}, \delta), \varepsilon := \gamma c/t$ and $\lambda := \gamma$, we deduce that there exists $u_\gamma \in B(\bar{x}, \delta)$ such that

$$\|u - u_\gamma\| \le \gamma, \tag{5.5}$$

$$F(w) + c\,\|w - u_\gamma\|/t > F(u_\gamma) \quad \forall w \in V \backslash \{u_\gamma\}. \tag{5.6}$$

By (5.6), u_γ minimizes the function $\psi(w) := F(w) + c\,\|w - u_\gamma\|/t$ on V. Conditions (5.4) and (5.5) imply that

$$\|u_\gamma - \bar{x}\| \le \|u_\gamma - u\| + \|u - \bar{x}\| \le .75\delta + .25\rho^* < \delta,$$

which means that u_γ lies in the interior of V, and so u_γ is a local minimizer for ψ. By the sum formula for subdifferentials (e.g., [20, Corollary 4.6]),

$$0 \in \partial\psi(u_\gamma) \subset \partial F(u_\gamma) + \frac{c}{t}\partial(\|\cdot - u_\gamma\|)(u_\gamma).$$

Hence there exists $z_\gamma \in \partial f(u_\gamma)$ such that $z_\gamma = cy_\gamma/t$ for some y_γ with $\|y_\gamma\| \leq 1$. Finally, since $\operatorname{dist}(u, S) > \gamma$, (5.5) implies that $u_\gamma \notin S$. Thus by (5.1),

$$c \leq \|z_\gamma\| = c\,\|y_\gamma\|\,/t \leq c/t < c,$$

which is a contradiction. □

We observe that any subdifferential with a calculus rule like [20, Corollary 4.6] could be used in place of ∂f in (5.1). One reason to choose the Mordukhovich subdifferential in (5.1) is that ∂f is the smallest possible subdifferential having a sum rule and some other desirable properties [13].

We illustrate Theorem 5.2 with a simple example.

Example 5.1. Define $f : \mathbb{R}^2 \to \mathbb{R}$ by $f(x, y) := |y|$, and let $\bar{x} := (0, 0)$ and $\delta > 0$. Then $S = \{(0, y) \mid -\delta \leq y \leq \delta\}$ and $\|z\| = 1\ \forall z \in \partial f(x, y)$ with $(x, y) \in B(\bar{x}, \delta)\backslash S$. Hence (5.1) holds with $c = 1$, and $\bar{x}$ is a weak sharp minimum of order one for f.

The following example shows that (5.1) is not necessary for weak sharp minimality of order one. It also shows that (2.3) may hold even when (5.1) is not satisfied.

Example 5.2. Define $f : \mathbb{R} \to \mathbb{R}$ by

$$f(x) := \begin{cases} 0 & \text{if } x \leq 0, \\ 2^{-n} & \text{if } 2^{-n-1} \leq x \leq 2^{-n},\ n \text{ an odd integer}, \\ 3x - 2^{-n} & \text{if } 2^{-n-1} \leq x \leq 2^{-n},\ n \text{ an even integer}, \end{cases}$$

and let $\bar{x} = 0$. The function f is continuous (Lipschitzian, in fact), and for $\delta > 0$, $S = \{x \mid -\delta \leq x \leq 0\}$. Since $f(x) \geq x\ \forall x \geq 0$, (5.2) holds with $c = 1$ and any $\mu \in (0, \delta)$. However, $0 \in \partial f(x)$ for $x = 3/2^{2n+1}$, $n = 0, 1, 2, \ldots$, so (5.1) does not hold for any $c > 0$.

On the other hand, $N(S, \bar{x}) = \{y \mid y \geq 0\}$ and

$$\underline{d}^1_S f(\bar{x}; y) = \underline{d}^1 f(\bar{x}; y) = y \quad \forall y \geq 0,$$

so (2.3) holds at $\bar{x} = 0$. In this example, condition (2.3) identifies a weak sharp minimizer of order one that is not detected by (5.1).

Since Theorem 5.2 specifies that f be continuous, it is not possible to use indicator functions, as in section 3, to derive further optimality conditions for constrained problems. However, other standard techniques can be used to extend Theorem 5.2 to the constrained case. We illustrate one such technique for an inequality-constrained problem in Corollary 5.3.

COROLLARY 5.3. *Let $g_i : \mathbb{R}^n \to \mathbb{R}$, $i = 0, 1, \ldots, m$, be continuous, and define $C := \{x \in \mathbb{R}^n \mid g_i(x) \leq 0,\ i = 1, \ldots, m\}$. Suppose that $\bar{x} \in C$ and $g_0(\bar{x}) \leq g_0(x)$ $\forall x \in C \cap B(\bar{x}, \delta)$. Call $S := \{x \in C \cap B(\bar{x}, \delta) \mid g_0(x) = g_0(\bar{x})\}$. Assume that the following conditions are satisfied $\forall x \in B(\bar{x}, \delta)\backslash S$:*

(i) *If $x_i^* \in \partial^\infty g_i(x)$, $i \in \{0\} \cup I(x)$, and $\sum x_i^* = 0$, then $x_i^* = 0\ \forall i$.*

(ii) *There exists $c > 0$ such that*

$$\bigcup \left\{ \sum_{i \in 0 \cup I(x)} \lambda_i \circ \partial g_i(x) \mid \lambda_i \geq 0,\ \sum \lambda_i = 1 \right\} \cap \operatorname{int} B(0, c) = \emptyset,$$

where we define $\lambda \circ \partial g(x)$ *to be* $\lambda \partial g(x)$ *for* $\lambda > 0$ *and* $\partial^\infty g(x)$ *for* $\lambda = 0$.

Then there exists $\mu \in (0, \delta)$ *such that*

$$g_0(x) \geq g_0(\bar{x}) + c\,\mathrm{dist}(x, S) \quad \forall x \in B(\bar{x}, \mu) \cap C. \tag{5.7}$$

Proof. Define $f(x) := \max\{g_0(x) - g_0(\bar{x}), g_1(x), \ldots, g_m(x)\}$. Then $S = \{x \in B(\bar{x}, \delta) \mid f(x) = f(\bar{x})\}$. Since (i) is satisfied, Theorem 7.5 of [21] implies that

$$\partial f(x) \subset \bigcup \left\{ \sum_{i \in 0 \cup I(x)} \lambda_i \circ \partial g_i(x) \mid \lambda_i \geq 0, \sum \lambda_i = 1 \right\}$$

$\forall x \in B(\bar{x}, \delta) \backslash S$. Condition (ii) then guarantees that (5.1) holds. By Theorem 5.2, we conclude that (5.7) holds. □

It is possible for Corollary 5.3 to identify weak sharp minima of order one that are not recognized by Theorem 3.2. We revisit Example 3.1 to demonstrate this fact.

Example 5.3. Let $n = 2$, $m = 2$ in Corollary 5.3. For $(x, y) \in \mathbb{R}^2$, define $g_0(x, y) = x$, $g_1(x, y) = -x$, $g_2(x, y) = -y$. As noted in Example 3.1, each element of $S = \{(0, y) \mid y \geq 0\}$ is a weak sharp minimizer of order one for this problem, but $(0, 0)$ is not identified by Theorem 3.2. We observe that (i) of Corollary 5.3 is satisfied $\forall (x, y) \in \mathbb{R}^2$ since each g_i is Lipschitzian. For $\bar{x} = (0, y)$ and $y > 0$, $I(z) = \emptyset$ $\forall z \in \mathbb{R}^2 \backslash S$ near $\bar{x}$, and (ii) is satisfied with $c = 1$ since $\partial g_0(z) = \{(1, 0)\}$. For $\bar{x} = (0, 0)$ and $z \in \mathbb{R}^2 \backslash S$ near $\bar{x}$, either $I(z) = \emptyset$ or $I(z) = \{2\}$. In either case, (ii) is satisfied. Hence Corollary 5.3 identifies each element of S as a weak sharp local minimizer of order one.

Examples 5.2 and 5.3 show that the optimality conditions of Theorem 5.2 and Corollary 5.3 are, in general, neither weaker nor stronger than the corresponding parts of Theorems 2.5 and 3.2.

6. Conclusion. In this paper, we have developed a number of sufficient conditions and characterizations of weak sharp minimality in nonsmooth programming, in many cases generalizing known results for strict minimality. Several questions merit further investigation. For example, can more be said about the relationship between $\underline{d}^m_S f(x; \cdot)$ and $\underline{d}^m f(x; \cdot)$? Can our results be applied fruitfully to the study of sensitivity analysis in nonlinear programming? We hope to address such questions in future work.

REFERENCES

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

[2] A. Auslender, *Stability in mathematical programming with nondifferentiable data*, SIAM J. Control Optim., 22 (1984), pp. 239–254.

[3] A. Ben-Tal, *Second-order and related extremality conditions in nonlinear programming*, J. Optim. Theory Appl., 31 (1980), pp. 143–165.

[4] J. F. Bonnans and A. Ioffe, *Second-order sufficiency and quadratic growth for nonisolated minima*, Math. Oper. Res., 20 (1995), pp. 801–817.

[5] J. F. Bonnans and A. Ioffe, *Quadratic growth and stability in convex programming problems with multiple solutions*, J. Convex Anal., 2 (1995), pp. 41–57.

[6] J. V. Burke, *An exact penalization viewpoint of constrained optimization*, SIAM J. Control Optim., 29 (1991), pp. 968–998.

[7] J. V. Burke and M. C. Ferris, *Weak sharp minima in mathematical programming*, SIAM J. Control Optim., 31 (1993), pp. 1340–1359.

[8] F. H. Clarke, *Optimization and Nonsmooth Analysis*, Wiley, New York, 1983.

[9] G. Di Pillo and F. Facchinei, *Exact barrier function methods for Lipschitz programs*, Appl. Math. Optim., 32 (1995), pp. 1–31.

[10] A. V. Fiacco and G. P. McCormick, *Nonlinear Programming: Sequential Unconstrained Minimization Techniques*, Wiley, New York, 1968.
[11] S.-P. Han and O. L. Mangasarian, *Exact penalty functions in nonlinear programming*, Math. Programming, 17 (1979), pp. 251–269.
[12] S. Howe, *New conditions for exactness of a simple penalty function*, SIAM J. Control, 11 (1973), pp. 378–381.
[13] A. D. Ioffe, *Approximate subdifferentials and applications* I*: The finite-dimensional theory*, Trans. Amer. Math. Soc., 281 (1984), pp. 389–416.
[14] A. D. Ioffe, *On sensitivity analysis of nonlinear programs in Banach spaces: The approach via composite unconstrained optimization*, SIAM J. Optim., 4 (1994), pp. 1–43.
[15] D. Klatte, *On quantitative stability for non-isolated minima*, Control Cybernet., 23 (1994), pp. 183–200.
[16] J. Kyparisis, *On uniqueness of Kuhn–Tucker multipliers in nonlinear programing*, Math. Programing, 32 (1985), pp. 242–246.
[17] E. S. Levitin, A. A. Milyutin, and N. P. Osmolovskii, *Conditions of high order for a local minimum in problems with constraints*, Russian Math. Surveys, 33 (1978), pp. 97–168.
[18] B. S. Mordukhovich, *Maximum principle in the problem of time optimal response with nonsmooth constraints*, J. Appl. Math. Mech., 40 (1976), pp. 960–969.
[19] B. S. Mordukhovich, *Complete characterization of openness, metri regularity, and Lipschitzian properties of multifunctions*, Trans. Amer. Math. Soc., 340 (1993), pp. 1–35.
[20] B. S. Mordukhovich, *Generalized differential calculus for nonsmooth and set-valued mappings*, J. Math. Anal. Appl., 183 (1994), pp. 250–288.
[21] B. S. Mordukhovich and Yongheng Shao, *Nonsmooth sequential analysis in Asplund spaces*, Trans. Amer. Math. Soc., 348 (1996), pp. 1235–1280.
[22] E. Rosenberg, *Exact penalty functions and stability in locally Lipschitz programming*, Math. Programming, 30 (1984), pp. 340–356.
[23] A. Shapiro, *Perturbation theory of nonlinear programs when the set of solutions is not a singleton*, Appl. Math. Optim., 18 (1988), pp. 215–229.
[24] A. Shapiro, *Perturbation analysis of optimization problems in Banach spaces*, Numer. Funct. Anal. Optim., 13 (1992), pp. 97–116.
[25] G. Still and M. Streng, *Optimality conditions in smooth nonlinear programming*, J. Optim. Theory Appl., 90 (1996), pp. 483–515.
[26] M. Studniarski, *Necessary and sufficient conditions for isolated local minima of nonsmooth functions*, SIAM J. Control Optim., 24 (1986), pp. 1044–1049.
[27] D. E. Ward, *Exact penalties and sufficient conditions for optimality in nonsmooth optimization*, J. Optim. Theory Appl., 57 (1988), pp. 485–499.
[28] D. E. Ward, *Characterizations of strict local minima and necessary conditions for weak sharp minima*, J. Optim. Theory Appl., 80 (1994), pp. 551–571.

SIAM J. CONTROL OPTIM.
Vol. 38, No. 1, pp. 237–253

ON THE MINIMUM PROBLEM FOR A CLASS OF NONCOERCIVE NONCONVEX FUNCTIONALS*

GRAZIANO CRASTA†

Abstract. We are concerned with the problem of existence of solutions to the variational problem

$$\min\left\{\int_0^R g(t,v'(t))\,dt;\ v\in AC([0,R]),\ v(R)=0\right\},$$

with only one fixed endpoint prescribed. The map $g\colon [0,R]\times\mathbb{R}\to\overline{\mathbb{R}}$ is a normal integrand, for which neither convexity nor superlinear growth conditions are assumed.

As an application, we give an existence result for the radially symmetric variational problem

$$\min_{u\in W_0^{1,1}(B_R)}\int_{B_R}\left[f\left(|x|,|\nabla u(x)|\right)+a(|x|)\,u(x)\right]dx,$$

where B_R is the ball of $\mathbb{R}^n$ centered at the origin and with radius $R>0$, the map $f\colon [0,R]\times[0,+\infty[\to\overline{\mathbb{R}}$ is a normal integrand, and $a\in L^1(0,R)$. Again, neither convexity nor superlinear growth conditions are made on f.

These kinds of problems, with nonconvex Lagrangians with respect to ∇u, arise in different fields of mathematical physics such as optimal design and nonlinear elasticity.

Key words. calculus of variations, existence, radially symmetric solutions, nonconvex problems, noncoercive problems

AMS subject classifications. Primary, 49J05; Secondary, 49J10, 49J30

PII. S0363012997330701

1. Introduction.

We study the existence of solutions to the variational problem

$$\min_{v\in V}\Gamma(v),\quad \Gamma(v)\doteq\int_0^R g(t,v'(t))\,dt, \tag{1}$$

where $V\doteq\{v\in AC([0,R]);\ v(R)=0\}$. The map $g\colon [0,R]\times\mathbb{R}\to\overline{\mathbb{R}}$ is a normal integrand for which neither convexity nor superlinear growth conditions are assumed. Furthermore, we remark that only one fixed endpoint is prescribed.

As customary, we consider the relaxed problem associated with (1), namely,

$$\min_{v\in V}\overline{\Gamma}(v),\quad \overline{\Gamma}(v)\doteq\int_0^R g^{**}(t,v'(t))\,dt, \tag{2}$$

where $g^{**}(t,\cdot)$ is the bipolar of the map $\xi\mapsto g(t,\xi)$.

If g is assumed to be superlinear, then a solution to (2) can be found using the direct method of the calculus of variations (see, for example, [14]). From the properties of this solution, one can often deduce the existence (or nonexistence) of a solution to the nonconvex problem (1).

*Received by the editors November 26, 1997; accepted for publication (in revised form) December 29, 1998; published electronically December 15, 1999.
http://www.siam.org/journals/sicon/38-1/33070.html

†Dipartimento di Matematica Pura ed Applicata, Università di Modena, Via Campi 213/B, I-41100 Modena, Italy (crasta@unimo.it).

In our case we follow a different approach based on the indirect method described below. As a first step we prove that there exists a nonempty closed interval $J \subseteq \mathbb{R}$ such that the fixed endpoints variational problem

$$\min_{v \in V(\alpha)} \overline{\Gamma}(v), \qquad V(\alpha) \doteq \{v \in V;\, v(0) = -\alpha\}, \tag{3}$$

has a solution for every $\alpha \in J$. Next, we show that the extended real-valued function $\overline{\phi}(\alpha) \doteq \inf\left\{\overline{\Gamma}(v);\ v \in V(\alpha)\right\}$ attains its minimum in a point $\overline{a} \in J$. Using the properties of the set of integrals of a decomposable family of integrable functions, we prove that the map $\phi(\alpha) \doteq \inf\{\Gamma(v);\ v \in V(\alpha)\}$ coincides with $\overline{\phi}$, and that the nonconvex fixed endpoints variational problem

$$\min_{v \in V(\alpha)} \Gamma(v) \tag{4}$$

has a solution for $\alpha = \overline{a}$ (see section 2 for the definition of decomposable family and Theorem 2.1 for the properties of the set of integrals of such a family). Finally, from the analysis above, we conclude that problem (1) has a solution.

The issue of the existence of solutions to fixed endpoints variational problems, either with nonconvex or noncoercive integrands, was also considered in [11, 12, 13, 15, 18, 19, 20].

As an application, we prove that there exists a solution to the radially symmetric variational problem

$$\min_{u \in W_0^{1,1}(B_R)} \int_{B_R} \left[f\left(|x|, |\nabla u(x)|\right) + a(|x|)\, u(x)\right] dx, \tag{5}$$

where B_R is the ball of $\mathbb{R}^n$ centered at the origin and with radius $R > 0$, the map $f\colon [0,R] \times [0,+\infty[\to \overline{\mathbb{R}}$ is a normal integrand, and $a \in L^1(0,R)$. We emphasize again the fact that f is not required to be convex nor superlinear with respect to ∇u. Nonconvex radially symmetric functionals were considered in [7] for a more general framework on the integrand, assuming superlinear growth with respect to ∇u. For related results we mention also [6, 24].

These kinds of problems arise in different fields of mathematical physics. Among others, we cite [16, 17] and [2, 3, 21, 24], respectively, for problems of optimal design and nonlinear elasticity.

The plan of the paper is the following. In section 3 we state the main result concerning problem (1), and we give some examples. Then, in sections 4 and 6, respectively, we give applications to the existence of solutions to problem (5) and to a minimization problem involving the Laplacian. The proofs of the results stated in sections 3 and 4 are given, respectively, in sections 7 and 5.

2. Preliminaries. We shall denote by $\operatorname{int} A$, $\overline{A}$, $\operatorname{co} A$, $\overline{\operatorname{co}}\, A$, respectively, the interior, the closure, the convex hull, and the closure of the convex hull of a set A. Given a function $\psi\colon X \to \overline{\mathbb{R}} \doteq]-\infty, +\infty]$, we denote by $\operatorname{epi}\psi$ and $\operatorname{graph}\psi$, respectively, its epigraph and its graph, and by ψ^* and ψ^{**}, respectively, its polar and bipolar functions. If ψ admits an affine minorant, then we have that $\operatorname{epi}\psi^{**} = \overline{\operatorname{co}}\operatorname{epi}\psi$ (see [14, Prop. 3.2]).

We say that $\phi\colon [0,R] \times \mathbb{R} \to \overline{\mathbb{R}}$ is a normal integrand if $\phi(t,\cdot)$ is lower semicontinuous for almost every (a.e.) $t \in [0,R]$, and there exists a Borel function $\hat{\phi}\colon [0,R] \times [0,+\infty[\to \overline{\mathbb{R}}$ such that $\hat{\phi}(t,\cdot) = \phi(t,\cdot)$ for a.e. $t \in [0,R]$ (see [14, Def. VIII.1.1]). We

shall denote by ϕ^* and ϕ^{**}, respectively, the polar and the bipolar of the function $\phi(t,\cdot)$.

A convex function $\psi\colon \mathbb{R} \to \overline{\mathbb{R}}$ is proper if its essential domain $\operatorname{Dom}\psi \doteq \{\xi \in \mathbb{R};\ \psi(\xi) < +\infty\}$ is not empty. As customary, the subgradient of ψ at ξ will be denoted by $\partial\psi(\xi)$. For the definition and the main properties of the subgradient we refer to [14, 22]. We recall that, if $\xi \in \operatorname{Dom}\psi$, then $\partial\psi(\xi) = [\psi'_-(\xi), \psi'_+(\xi)]$, where $\psi'_-(\xi)$ and $\psi'_+(\xi)$ denote, respectively, the left and right derivatives of ψ at ξ. If we extend $\psi'_\pm$ beyond the interval $\operatorname{Dom}\psi$ by setting both $+\infty$ for points lying to the right of $\operatorname{Dom}\psi$ and both $-\infty$ for points lying to the left, then $\psi'_\pm$ are monotone nondecreasing functions on $\mathbb{R}$, finite in $\operatorname{int}\operatorname{Dom}\psi$, and for every ξ one has

$$\begin{aligned} &\lim_{\eta\searrow\xi} \psi'_+(\eta) = \psi'_+(\xi), \quad \lim_{\eta\nearrow\xi} \psi'_+(\eta) = \psi'_-(\xi), \\ &\lim_{\eta\searrow\xi} \psi'_-(\eta) = \psi'_+(\xi), \quad \lim_{\eta\nearrow\xi} \psi'_-(\eta) = \psi'_-(\xi) \end{aligned} \tag{6}$$

(see [22, Thm. 24.1]). For notational convenience, if ψ depends on two variables (t,ξ), we denote by $\partial\psi(t,\xi)$ the subgradient of the function $\xi \mapsto \psi(t,\xi)$ at (t,ξ).

We say that a convex function ψ is strictly convex at infinity if $\operatorname{Dom}\psi = \mathbb{R}$ and its graph contains no halflines. It is easy to check that ψ is strictly convex at infinity if and only if for every $r_1 > 0$ there exists $r_2 > 0$ such that, for all $p \in \mathbb{R}$, $\partial\psi^*(p) \cap \overline{B}_{r_1} \neq \emptyset$ implies $\partial\psi^*(p) \subseteq B_{r_2}$ (see [10, Def. 2.1]). We recall that, if ϕ is any function such that ϕ^{**} is strictly convex at infinity, then the convex hull of the epigraph of ϕ is closed (see [13, Thm. 3.6]). As a consequence, ϕ^{**} coincides with the convexification of ϕ, that is, $\phi^{**} = \sup\{\psi \le \phi;\ \psi \text{ convex}\}$.

Another result which will be used in the following concerns the set of integrals of a decomposable family of integrable functions. More precisely, given a set $K \subset L^1([0,R],\mathbb{R}^m)$, we shall denote by $I(K)$ the set of integrals of the elements of K, that is,

$$I(K) \doteq \left\{ \int_0^R w(t)dt;\ w \in K \right\}. \tag{7}$$

We say that K is decomposable if, for every measurable set $E \subset [0,R]$ and all $u, v \in K$, one has $u \cdot \chi_E + v \cdot \chi_{[0,R]\setminus E} \in K$, where χ_E denotes the characteristic function of the set E.

Now we recall an extension of the classical Lyapunov theorem on the range of nonatomic vector-valued measures (see [19, Thm. 3]).

THEOREM 2.1. *If $K \subset L^1([0,R],\mathbb{R}^m)$ is decomposable, then $I(K)$ is convex and $I(K) = I(\operatorname{co} K)$. If, in addition, K is strongly closed in the L^1 topology, then $I(K)$ contains the compact extremal faces of $\overline{I(K)}$.*

3. The main result. In this section we state the main theorem of the paper, and we give some examples. The proof will be postponed to section 7.

We collect here the hypotheses that will be used in the rest of the paper. Conditions (G1)–(G3) will be enough to prove the existence of a solution to the convex problem (2). In order to obtain a solution to (1) we also need the hypotheses (G4) and (G5).

(G1) The map $g\colon [0,R]\times[0,+\infty[\to \overline{\mathbb{R}}$ is a normal integrand, and $\operatorname{Dom} g^{**}(t,\cdot) = \mathbb{R}$ for a.e. $t \in [0,R]$.

(G2) The map $t \mapsto g^{**}(t,0)$ is integrable on $[0,R]$.

(G3) There exist a positive constant m and a function $b \in L^1(0,R)$ such that

$$g^{**}(t,\xi) \geq m|\xi| - b(t) \quad \text{a.e. } t \in [0,R] \ \forall \xi \in \mathbb{R}. \tag{8}$$

(G4) For every $r_1 > 0$ there exists $r_2 > 0$ such that, for a.e. $t \in [0,R]$ and every $p \in \mathbb{R}$, if $\partial g^*(t,p) \cap \overline{B}_{r_1} \neq \emptyset$ then $\partial g^*(t,p) \subseteq B_{r_2}$.

(G5) The map $t \mapsto g(t,\xi)$ is integrable on $[0,R]$ for every $\xi \in \mathbb{R}$.

Remark 3.1. 1. It is easy to prove that, if (G2) holds, then (G3) is equivalent to

(G3′) There exists a positive constant m such that $g^*(\cdot,p) \in L^1(0,R)$ for every $p \in [-m,m]$.

Indeed, let us assume that (G3) holds, and let $p \in [-m,m]$. From (G3) and the definition of g^* we deduce that

$$-g^{**}(t,0) \leq g^*(t,p) \leq b(t) \quad \text{a.e. } t \in [0,R]; \tag{9}$$

hence, from (G2), we conclude that $g^*(\cdot,p) \in L^1(0,R)$. Conversely, assume that the condition (G3′) holds. Then it is easy to check that (G3) holds with $b(t) \doteq \max\{g^*(t,m),\, g^*(t,-m)\}$.

2. Assumption (G3) looks similar to other growth conditions found in literature, suitable to minimize fixed endpoint variational problems (see, for example, [8, §10.4, §11.1]). The main difference is that these conditions require that for every $m \in \mathbb{R}$ (and not only for a fixed one) there exists an integrable function $b = b_m$ such that (8) holds. For example, let us consider a function g satisfying the bounds $a_1|\xi| - a_2 \leq g(t,\xi) \leq c_1|\xi| + c_2$ for some positive constants a_j, c_j, $j = 1,2$. It is clear that (G3) holds with $m = a_1$, whereas (8) cannot hold if $m > c_1$. It should be noted that the weaker growth condition (G3) is sufficient since only one fixed endpoint is prescribed.

3. It is easy to see that, if (G4) holds, then $g^{**}(t,\xi)$ is strictly convex at infinity with respect to ξ for almost every fixed $t \in [0,R]$. Furthermore, from Proposition 2.2 in [10], we have that (G4) holds if, for example, the function $g^{**}(t,\xi)$ is continuous on $[0,R] \times \mathbb{R}$ and strictly convex at infinity with respect to ξ for every fixed $t \in [0,R]$.

4. If g is coercive, that is, if $g(t,\xi) \geq \psi(|\xi|)$ where $\psi\colon [0,+\infty[\to \mathbb{R}$ is a continuous increasing function satisfying $\lim_{s\to+\infty} \psi(s)/s = +\infty$, then g satisfies (G3) and (G4).

5. From (G3) it follows that the functional $\overline{\Gamma}$ defined in (2) is bounded from below.

6. It is worth rewriting conditions (G1)–(G5) when $g \equiv g(\xi)$ does not depend on t. It is easy to see that assumptions (G1)–(G3) and (G4)–(G5) can be replaced, respectively, by (g1)–(g2) and (g3) below.

(g1) $g\colon \mathbb{R} \to \overline{\mathbb{R}}$ is a lower semicontinuous function, and $\operatorname{Dom} g^{**} = \mathbb{R}$.

(g2) There exist positive constants m and b such that $g^{**}(\xi) \geq m|\xi| - b$ for every $\xi \in \mathbb{R}$.

(g3) g^{**} is strictly convex at infinity, and $g(\xi) \in \mathbb{R}$ for every $\xi \in \mathbb{R}$.

THEOREM 3.2. *Assume that* (G1)–(G3) *hold. Then the convex problem* (2) *admits a solution* $\bar{v} \in V$ *satisfying the Euler–Lagrange inclusion*

$$0 \in \partial g^{**}(t, \bar{v}'(t)) \quad \text{a.e. } t \in [0,R]. \tag{10}$$

If, in addition, (G4) *and* (G5) *hold, then* $\bar{v}$ *satisfies*

$$g(t,\bar{v}'(t)) = g^{**}(t,\bar{v}'(t)) \qquad \text{a.e. } t \in [0,R], \tag{11}$$

thus providing a solution to the nonconvex problem (1).

Remark 3.3. Notice that, if assumption (G5) holds, then for every $\alpha \in \mathbb{R}$ the set $\{v \in V(\alpha);\ \overline{\Gamma}(v) < +\infty\}$ is not empty (we recall that $V(\alpha)$ is the set defined in (4)). Hence, if (G1)–(G3) and (G5) hold, from Corollary 1 in [23] we can infer that $\bar{v}$ is a solution to problem (2) if and only if it satisfies the Euler–Lagrange inclusion (10).

Example 3.4. Let $n = 1$, and consider the function $g(t, \xi) \doteq g_0(|\xi|) - t\,\xi$, where

$$g_0(s) \doteq \begin{cases} s - \sqrt{s} & \text{if } s > 1/4, \\ -1/4 & \text{if } 0 \le s \le 1/4. \end{cases}$$

It is easy to show that the assumptions (G1), (G2), (G4), and (G5) are satisfied, whereas (G3) is satisfied if and only if $R < 1$ (it is enough to choose $m = 1 - R$).

We are going to show that, if $R \ge 1$, then the variational problem (1) has no solution. Let $R \ge 1$, and assume by contradiction that $u \in V$ is a solution to (1). It is easy to check that $\xi \mapsto g_0(|\xi|)$ is continuously differentiable and $g_0'(s) = 1 - s^{-1/2}/2$ if $s > 1/4$, while $g_0'(s) = 0$ if $0 \le s \le 1/4$. From the Euler–Lagrange necessary conditions (10) we deduce that $g_0'(|u'(t)|)\,\operatorname{sign} u'(t) = t$ for a.e. $t \in [0, R]$. Since $|g_0'(s)| < 1$ for every s, this relation cannot be satisfied a.e. if $R > 1$. In the case $R = 1$ we deduce that $u'(t) = \left[4(1-t)^2\right]^{-1}$ for a.e. $t \in [0, 1]$. But now $u' \notin L^1(0, 1)$, giving a contradiction.

Example 3.5. Let $n = 1$, $g(t, \xi) \doteq \sqrt{1 + \xi^2} - t\,\xi$. Again, (G1), (G2), (G4), and (G5) are satisfied, whereas (G3) holds if and only if $R < 1$, choosing $m = 1 - R$. It is easy to see that, in this case, for every $R > 1$ problem (1) does not have a solution. Indeed, if u is a solution, then from the Euler–Lagrange necessary conditions (10) we have that $u'(t)/\sqrt{1 + u'(t)^2} = t$ for a.e. $t \in [0, R]$. Since the left-hand side is strictly less than 1 for every t, this condition cannot be satisfied if $R > 1$.

4. Application to radially symmetric problems. In this section we give an application of Theorem 3.2 to radially symmetric variational problems of the form (5).

We collect here the assumptions on f and a. Conditions (H1)–(H4) will be enough to prove the existence of a solution to the relaxed problem

$$\min_{u \in W_0^{1,1}(B_R)} \int_{B_R} \left[f^{**}\left(|x|, |\nabla u(x)|\right) + a(|x|)\, u(x) \right] dx. \tag{12}$$

In order to obtain a solution to the nonconvex problem (5) we also need the hypotheses (H5) and (H6).

(H1) The map $f\colon [0, R] \times [0, +\infty[\to \overline{\mathbb{R}}$ is a normal integrand, and $\operatorname{Dom} f^{**}(t, \cdot) = \mathbb{R}$ for a.e. $t \in [0, R]$.

(H2) The map $x \mapsto f^{**}(|x|, 0)$ is integrable on B_R; that is, $t \mapsto t^{n-1} f^{**}(t, 0)$ is integrable on $[0, R]$.

(H3) The function $A\colon [0, R] \to \mathbb{R}$ defined by $A(0) \doteq 0$, $A(t) \doteq t^{1-n} \int_0^t s^{n-1} a(s)\, ds$, $t \in]0, R]$, belongs to $L^\infty(0, R)$.

(H4) There exists a constant $M > \|A\|_{L^\infty}$ such that for every $\rho \in [0, M[$ there exists a function $c_\rho \in L^1(0, R)$ such that

$$f^{**}(t, \xi) \ge \rho\,|\xi| - t^{1-n} c_\rho(t) \quad \text{a.e. } t \in [0, R],\ \xi \in \mathbb{R}.$$

(H5) For every $r_1 > 0$ there exists $r_2 > 0$ such that, for every $(t, p) \in [0, T] \times \mathbb{R}$, if $\partial f^*(t, p) \cap \overline{B}_{r_1} \ne \emptyset$, then $\partial f^*(t, p) \subseteq B_{r_2}$.

(H6) The map $t \mapsto t^{n-1} f(t, s)$ is integrable on $[0, R]$ for every $s \in [0, +\infty[$.

Remark 4.1. From (H4) we deduce that $M \le \operatorname{ess\,inf}_{t\in[0,R]} M(t)$, where $M(t) \doteq \lim_{s\to+\infty} f^{**}(t,s)/s$.

We are now in a position to state the main theorem concerning radially symmetric problems. The proof will be postponed to section 5.

THEOREM 4.2. *Assume that* (H1)–(H4) *hold. Then the convex problem* (12) *has at least one radially symmetric solution* $\tilde u(x) = \bar u(|x|)$, *and* $\bar u$ *satisfies the Euler–Lagrange inclusion*

$$\bar u'(t) \in \partial f^*(t, A(t)) \quad \text{a.e. } t \in [0,R]. \tag{13}$$

Furthermore, if either $a(t) > 0$ *(resp.,* < 0*) for a.e.* $t \in [0,R]$ *or* $f^{**}(t,\cdot)$ *has a strict minimum at* 0 *for a.e.* $t \in [0,R]$, *then every solution to* (12) *is radially symmetric.*

If, in addition, (H5) *and* (H6) *hold, then* $\tilde u$ *satisfies*

$$f(|x|, |\nabla \tilde u(x)|) = f^{**}(|x|, |\nabla \tilde u(x)|) \qquad \text{for a.e. } x \in B_R, \tag{14}$$

thus providing a solution to the nonconvex problem (5).

For the reader's convenience, we state Theorem 4.2 in the case $f = f(|\xi|)$ does not depend explicitly on x.

THEOREM 4.3. *Assume the following:*

(i) $f\colon [0,+\infty[\to \overline{\mathbb{R}}$ *is a lower semicontinuous function, and* $\operatorname{Dom} f^{**} = \mathbb{R}$;

(ii) $M \doteq \lim_{s\to+\infty} f^{**}(s)/s > 0$;

(iii) $a \in L^1(0,R)$ *and* $\sup_{t\in]0,R]} \left| t^{1-n} \int_0^t s^{n-1}\, a(s)\, ds \right| < M$.

Then the variational problem

$$\min_{u\in W_0^{1,1}(B_R)} \int_{B_R} \left[f^{**}\left(|\nabla u(x)|\right) + a(|x|)\, u(x) \right] dx \tag{15}$$

has a radially symmetric solution. If, in addition,

(iv) f^{**} *is strictly convex at infinity,*

then the nonconvex problem

$$\min_{u\in W_0^{1,1}(B_R)} \int_{B_R} \left[f\left(|\nabla u(x)|\right) + a(|x|)\, u(x) \right] dx \tag{16}$$

has a radially symmetric solution.

Remark 4.4. In the particular case $a(t) \equiv 1$, condition (iii) becomes

(iii′) $R < nM$.

In this case the bound $R < nM$ is optimal, in the sense explained in Example 3.4.

Example 4.5. We discuss here an example treated in [5, 25, 4] without the symmetry assumption on the domain. Let $a(t) \equiv 1$, and let $f\colon [0,+\infty[\to \overline{\mathbb{R}}$ be a lower semicontinuous function satisfying $f(L) = 0$ for some $L > 0$, and

$$f(s) \ge \max\{0, \lambda(s-L)\} \quad \forall s \ge 0$$

for some $\lambda > 0$. It easy to see that $f^{**}(\xi) = 0$ if $|\xi| \le L$, while $f^{**}(\xi) \ge \lambda(|\xi| - L)$ if $|\xi| \ge L$, so that $M \ge \lambda$.

Now we want to apply Theorem 4.3. Conditions (i) and (ii) are satisfied by f^{**}, while (iii′) is certainly satisfied if $R < n\lambda$. Thus, if $R < n\lambda$, there exists a solution $\tilde u(x) = \bar u(|x|)$ to (15), and $\bar u(t)$ satisfies (13), that is,

$$\frac{t}{n} \in \partial f^{**}(\bar u'(t)) \quad \text{a.e. } t \in [0,R]. \tag{17}$$

This implies that $\bar{u}'(t) \geq 0$ for a.e. t. It is easy to check that

$$\partial f^{**}(s) \begin{cases} = \{0\} & \text{if } 0 \leq s < L, \\ \supseteq [0, \lambda] & \text{if } s = L, \\ \subseteq [\lambda, +\infty[& \text{if } s > L. \end{cases} \tag{18}$$

Furthermore, the condition $R < n\lambda$ implies that $0 < t/n < \lambda$ for every $t \in]0, R]$, so that from (17) and (18) we necessarily have $\bar{u}'(t) = L$ a.e. $t \in [0, R]$; that is, $\bar{u}(t) = -L(R - t)$, $t \in [0, R]$. Thus a solution $\tilde{u}$ to (15) is given by $\tilde{u}(x) \doteq -L \operatorname{dist}(x, \partial B_R)$. Since $f^{**}(L) = f(L)$, $\tilde{u}$ provides a solution to (16) even if f is not convex. In [5, 25] it was proved that there exists a solution whenever $R \leq n\lambda$.

5. Proof of Theorem 4.2. We collect here some results that will be used in order to reduce (12) to a one-dimensional problem (see [7, Thms. 1 and 2]).

PROPOSITION 5.1. *Problem* (12) *has a solution if and only if it admits at least one radially symmetric solution. Furthermore, if either* $a(t) > 0$ *(resp.,* < 0*) for a.e.* $t \in [0, R]$ *or* $f^{**}(t, \cdot)$ *has a strict minimum at* 0 *for a.e.* $t \in [0, R]$*, then every solution to* (12) *is radially symmetric.*

Proposition 5.1 implies that problem (12) is equivalent to

$$\min_{u \in W} \int_0^R t^{n-1} \left[f^{**}(t, u'(t)) + a(t)\, u(t) \right] dt, \tag{19}$$

where

$$W \doteq \left\{ u \in AC_{loc}(]0, R]);\ u(R) = 0,\ t \mapsto t^{n-1} u'(t) \in L^1(0, R) \right\}, \tag{20}$$

in the sense that (12) has a solution if and only if (19) has a solution.

In order to treat problem (19), we need a further reduction. We begin by proving an integration-by-parts lemma involving functions in the set W defined in (20).

LEMMA 5.2. *Let* $a \in L^1(0, R)$*,* $u \in W$*. Then the maps* $t \mapsto t^{n-1} a(t) u(t)$ *and* $t \mapsto B(t) u'(t)$*,* $B(t) \doteq \int_0^t s^{n-1} a(s)\, ds$*, are integrable on* $[0, R]$*, and the following integration-by-parts formula holds:*

$$\int_0^R t^{n-1} a(t) u(t)\, dt = -\int_0^R B(t) u'(t)\, dt. \tag{21}$$

Proof. Since $u \in W$, there exists $k \in L^1(0, R)$ such that $u'(t) = t^{1-n} k(t)$ a.e. $t \in [0, R]$; that is, $u(t) = -\int_t^R s^{1-n} k(s)\, ds$. Clearly we have

$$|u(t)| \leq t^{1-n} \|k\|_{L^1} \qquad \text{a.e. } t \in [0, R], \tag{22}$$

so that $|t^{n-1}\, a(t)\, u(t)| \leq \|k\|_{L^1} |a(t)| \in L^1(0, R)$. Furthermore, from the trivial estimate

$$|B(t)| \leq t^{n-1} \int_0^t |a(s)|\, ds, \qquad t \in [0, R], \tag{23}$$

we deduce that $|B(t) u'(t)| \leq \|a\|_{L^1} |k(t)| \in L^1(0, R)$. In order to prove (21), from the properties of the Lebesgue integral it suffices to show that

$$\lim_{\varepsilon \to 0+} \int_\varepsilon^R t^{n-1} a(t) u(t)\, dt = -\lim_{\varepsilon \to 0+} \int_\varepsilon^R B(t) u'(t)\, dt.$$

Integrating by parts we have that

$$\int_\varepsilon^R t^{n-1}a(t)u(t)\,dt = -B(\varepsilon)u(\varepsilon) - \int_\varepsilon^R B(t)u'(t)\,dt.$$

It remains to prove that $\lim_{\varepsilon\to 0+} B(\varepsilon)u(\varepsilon) = 0$. From (22) and (23) we infer that $|B(\varepsilon)u(\varepsilon)| \le \|k\|_{L^1}\int_0^\varepsilon |a(s)|\,ds$, and from the absolute continuity of the integral the last term tends to zero as $\varepsilon \to 0^+$. □

Using Lemma 5.2, we can integrate by parts the term $a(t)\,u(t)$ in (19), so that the integral in (19) is equal to $\int_0^R t^{n-1}\left[f^{**}(t,u'(t)) - A(t)\,u'(t)\right]\,dt$, where A is the function defined in (H3). With every $u \in W$ we associate a function $v \in V$ in the following way:

$$v(t) \doteq -\int_t^R s^{n-1}u'(s)\,ds.$$

It is easy to see that the map $u \in W \mapsto v \in V$ is a bijection; hence problem (19) is equivalent to problem (2), with

$$g(t,\xi) \doteq t^{n-1} f\left(t, \frac{|\xi|}{t^{n-1}}\right) - A(t)\,\xi, \tag{24}$$

in the sense that (19) has a solution if and only if (2) has a solution.

It remains to prove that the assumptions of Theorem 3.2 are satisfied. Clearly, (G1), (G2), and (G4), respectively, are equivalent to (H1), (H2), and (H5). It is also easy to verify that (H3) and (H6) imply (G5). Furthermore, from (H3) and (H4) we deduce that for every $\rho \in [0, M[$ there exists $c_\rho \in L^1(0,R)$ such that

$$g^{**}(t,\xi) \ge \rho|\xi| - c_\rho(t) - A(t)\xi \quad \text{a.e. } t \in [0,R]\ \forall \xi \in \mathbb{R}.$$

Now, from (H4), it is clear that (G3) holds choosing $m \in]0, M - \|A\|_{L^\infty}[$.

6. A minimization problem involving the Laplacian. Consider the minimization problem

$$\min_{u\in W_*} \int_{B_R} \left[f(|x|, \triangle u(x)) + a(|x|)\,u(x)\right]\,dx, \tag{25}$$

where $W_* \doteq \{u \in W^{2,1}(B_R) \cap W_0^{1,1}(B_R);\ \frac{\partial u}{\partial n} = 0 \text{ on } \partial B_R\}$. This kind of problem was considered in [6] for superlinear Lagrangians. In that paper it was proved that the corresponding relaxed problem

$$\min_{u\in W_*} \int_{B_R} \left[f^{**}(|x|, \triangle u(x)) + a(|x|)\,u(x)\right]\,dx \tag{26}$$

has a solution if and only if it admits at least one radially symmetric solution.

We recall that, if $\tilde u \in W_*$ is radially symmetric, that is, $\tilde u(x) = \bar u(|x|)$ for some $\bar u \in W_S \doteq \{u \in W^{2,1}_{loc}(]0,R]);\ \exists v \in W_* \text{ s.t. } u(|x|) = v(x)\}$, then, for $x \neq 0$, one has

$$\triangle\tilde u(x) = |x|^{1-n}\frac{d}{dt}\left[t^{n-1}\bar u'(t)\right]_{t=|x|}. \tag{27}$$

Let us define $V_* \doteq \left\{w \in W^{2,1}([0,R]);\ w(R) = 0,\ w'(R) = 0\right\}$. To every $u \in W_S$ we can associate a $w \in V_*$ by setting $w'(t) = t^{n-1}u'(t)$, $t \in]0,R]$.

Now, let $B(t) \doteq \int_0^t s^{n-1}a(s)\,ds$, $A(t) \doteq -\int_0^t s^{1-n}B(s)\,ds$, $t \in [0,R]$. Let $u \in W_S$, and let w be the corresponding function in V_*. Integrating twice by parts (see Lemma 5.2) we deduce that

$$\int_0^R t^{n-1}a(t)\,u(t)\,dt = -\int_0^R B(t)\,u'(t)\,dt$$
$$= -\int_0^R t^{1-n}w'(t)\,B(t)\,dt = -\int_0^R A(t)\,w''(t)\,dt.$$

Recalling (27), this implies that (26) is equivalent to

$$\min_{w\in V_*}\int_0^R \left[t^{n-1}f^{**}(t,t^{1-n}w''(t)) - A(t)\,w''(t)\right]\,dt, \tag{28}$$

in the sense that (26) has a solution if and only if (28) has a solution.

We can make a further reduction, associating with every $w \in V_*$ a function $v \in V$ such that $w' = v$, obtaining the minimization problem (2). Thus we can conclude that the following existence theorem for (25) holds.

THEOREM 6.1. *Assume that* (H1) *and* (H2) *hold, that the function*

$$A(t) \doteq \int_0^t s^{1-n}\left(\int_0^s \sigma^{n-1}a(\sigma)\,d\sigma\right)\,ds, \qquad t \in [0,R],$$

belongs to $L^\infty(0,R)$, *and* (H4) *is satisfied. Then problem* (26) *has at least one radially symmetric solution. If, in addition,* (H5) *and* (H6) *hold, then the nonconvex problem* (25) *has a radially symmetric solution.*

7. Proof of Theorem 3.2. In the first part of this section we shall study the convex problem (2), so that only conditions (G1)–(G3) will be assumed to hold. Finally, starting from Lemma 7.13 below, we shall prove that, under the additional conditions (G4) and (G5), the nonconvex problem (1) has a solution.

LEMMA 7.1. *Assume that* (G1)–(G3) *hold. Then the set*

$$H \doteq \left\{p \in \mathbb{R} :\ g^*(\cdot,p) \in L^1(0,R)\right\} \tag{29}$$

is an interval, and $[-m,m] \subseteq H$.

Proof. Let $p,q \in H$ and define $p_\lambda \doteq \lambda p + (1-\lambda)q$, $\lambda \in [0,1]$. From (G2) and the inequality

$$-g^{**}(t,0) \le g^*(t,p_\lambda) \le \lambda g^*(t,p) + (1-\lambda)g^*(t,q) \quad \text{a.e. } t \in [0,R]\ \forall \lambda \in [0,1],$$

it follows that $p_\lambda \in H$ for every $\lambda \in [0,1]$, and this proves that H is an interval. The inclusion $[-m,m] \subseteq H$ follows from (G3′). □

LEMMA 7.2. *Assume that* (G1)–(G3) *hold. Then, for every* $p \in]-m,m[$, *the set* S *of all measurable selections of the multifunction* $t \mapsto \partial g^*(t,p)$ *is integrably bounded.*

Proof. Choose $\varepsilon > 0$ such that $\overline{B}_\varepsilon(p) \subset]-m,m[$. Let $h \in S$. From the inequality

$$g^*(t,q) - g^*(t,p) \ge h(t)(q-p) \quad \forall q \in \mathbb{R} \text{ a.e. } t \in [0,R],$$

we get

$$-\varepsilon^{-1}\left[g^*(t,p-\varepsilon) - g^*(t,p)\right] \le h(t) \le \varepsilon^{-1}\left[g^*(t,p+\varepsilon) - g^*(t,p)\right],$$

and the conclusion now follows from Lemma 7.1. □

The first part of Theorem 3.2 can be already proved at this stage.

THEOREM 7.3. *Assume that* (G1)–(G3) *hold. Then problem* (2) *has a solution. More precisely, if* ξ *is a measurable selection of* $\partial g^*(\cdot,0)$*, then the function* $\bar v(t) \doteq -\int_t^R \xi(s)\,ds$ *provides a solution to* (2).

Proof. Since $0 \in]-m,m[$, from Lemma 7.2 we have that $\xi \in L^1(0,R)$, so that the function $\bar v(t) \doteq -\int_t^R \xi(s)\,ds$, $t \in]0,R]$, belongs to V. It is easy to see that $\bar v$ is a solution to (2). Indeed, since $0 \in \partial g^{**}(t,\bar v'(t))$ a.e. $t \in [0,R]$, we have that $\overline\Gamma(v) - \overline\Gamma(\bar v) \geq 0$ for every $v \in V$. □

In order to proceed with the analysis of problem (1), we need a further reduction. More precisely, for every $\alpha \in \mathbb{R}$ let us consider the functions

$$\phi(\alpha) \doteq \inf_{v\in V(\alpha)} \Gamma(v), \qquad \overline\phi(\alpha) \doteq \inf_{v\in V(\alpha)} \overline\Gamma(v), \tag{30}$$

where $V(\alpha)$ is the set defined in (3), with the convention $\inf\emptyset = +\infty$ (we recall that, at this stage, we are not assuming (G5), so that the maps ϕ and $\overline\phi$ may well assume the value $+\infty$ at some point). Now it is easy to see that (1) has a solution if and only if there exists $\overline a \in \mathbb{R}$ such that $\phi(\overline a) = \min_\alpha \phi(\alpha)$, and the fixed endpoints variational problem (4) has a solution for $\alpha = \overline a$. The same conclusion holds for the functional $\overline\Gamma$ and the associated function $\overline\phi$.

First we shall prove that the problem (3) has a solution for every $\alpha \in J$, where J is a nonempty (possibly unbounded) closed interval of $\mathbb{R}$. Then we shall prove that $\overline\phi$ attains its minimum in a point $\overline a \in J$, so that (2) has a solution. In order to prove that (1) has a solution, we shall show that $\phi \equiv \overline\phi$, and that (4) has a solution for $\alpha = \overline a$. For the analysis of noncoercive nonconvex problems like (4) see also [11, 12]. Necessary and sufficient conditions for the existence of solutions to the convex problem (3) can be found in [18], in the case $g^{**}(t,\cdot)$ continuously differentiable for a.e. t.

We recall a weak form of the Euler–Lagrange inclusion that will be used in the sequel (see [1, Thm. 3.1]).

THEOREM 7.4. *If* v *is a solution to* (3)*, then there exists* $p \in \mathbb{R}$ *such that*

$$p \in \partial g^{**}(t,v'(t)) \qquad \text{a.e. } t \in [0,R]. \tag{31}$$

Remark 7.5. If $h \in L^1(0,R)$ is an integrable selection of $\partial g^*(\cdot,p)$ for some p, then the function $v(t) \doteq -\int_t^R h(s)\,ds$ solves (2) for $\alpha = \int_0^R h(s)\,ds$. Namely,

$$\overline\Gamma(w) - \overline\Gamma(v) = \int_0^R \left[g^{**}(t,w'(t)) - g^{**}(t,v'(t))\right]\,dt \geq p\int_0^R \left[w'(t) - v'(t)\right]\,dt = 0$$

for every $w \in V(\alpha)$.

Let $h_1, h_2 \colon [0,R] \to [-\infty,+\infty]$ be the measurable functions defined by

$$h_1(t) \doteq \lim_{p\to -m+} (g^*)'_-(t,p), \quad h_2(t) \doteq \lim_{p\to m-} (g^*)'_+(t,p). \tag{32}$$

We remark that these limits exist, finite or infinite, for every $t \in [0,R]$, since the maps $p \mapsto (g^*)'_\pm(t,p)$ are monotone nondecreasing. Let us define $\alpha_j \doteq \int_0^R h_j(t)\,dt$, $j = 1,2$. Clearly $\alpha_1 \leq \alpha_2$, and, by Lemma 7.2, $\alpha_1 \in [-\infty,+\infty[$, $\alpha_2 \in]-\infty,+\infty]$. Let us define the closed interval

$$J \doteq \{\alpha \in \mathbb{R};\ \alpha_1 \leq \alpha \leq \alpha_2\}. \tag{33}$$

Remark 7.6. If $\alpha_1 \in \mathbb{R}$, then by definition $h_1 \in L^1(0,R)$. Furthermore, from (6), $h_1(t) = (g^*)'_+(t,-m)$ for a.e. $t \in [0,R]$. Now, from Remark 7.5, we can conclude that problem (3) has a solution for $\alpha = \alpha_1$. The same argument works for α_2 if $\alpha_2 \in \mathbb{R}$.

THEOREM 7.7. *Assume that* (G1)–(G3) *hold. Then problem* (3) *has a solution for every* $\alpha \in J$.

Proof. If $\alpha_j \in \mathbb{R}$ for some $j \in \{1,2\}$ and $\alpha = \alpha_j$, then the proof follows from Remark 7.6. Now let us assume that $\alpha \in]\alpha_1, \alpha_2[$. The maps $I^{\pm}\colon]-m,m[\to \mathbb{R}$,

$$I^{\pm}(p) \doteq \int_0^R (g^*)'_{\pm}(t,p)\,dt,$$

are monotone nondecreasing and $I^-(p) \le I^+(p)$ for every $p \in]-m,m[$. Furthermore, from the monotone convergence theorem and the fact that

$$\lim_{p\to -m+} (g^*)'_-(t,p) = \lim_{p\to -m+} (g^*)'_+(t,p),$$
$$\lim_{p\to m-} (g^*)'_-(t,p) = \lim_{p\to m-} (g^*)'_+(t,p),$$

one gets

$$\lim_{p\to -m+} I^{\pm}(p) = \alpha_1, \qquad \lim_{p\to m-} I^{\pm}(p) = \alpha_2.$$

Hence there exists $p \in]-m,m[$ such that $I^-(p) \le \alpha \le I^+(p)$. Let S be the set of all measurable selections of $\partial g^*(\cdot,p)$. Since $p \in]-m,m[$, from Lemma 7.2 S is integrably bounded; hence by Aumann's theorem (see [9, Thm. 7.2.1]) we obtain $\{\int_0^R h(t)\,dt;\ h \in S\} = [I^-(p), I^+(p)]$. Thus we can conclude that there exists $h \in S$ such that $\int_0^R h(t)\,dt = \alpha$, so that, recalling Remark 7.5, the proof is complete. ∎

Now we prove some properties of the maps ϕ and $\overline{\phi}$ defined in (30). We recall that, from Remark 3.1(5), these maps are bounded from below.

LEMMA 7.8. *The maps* $\overline{\phi}$ *and* ϕ *are convex.*

Proof. Let us define

$$K_1 \doteq \left\{(u,v) \in L^1(0,R)\times L^1(0,R);\ v(t) \ge g^{**}(t,u(t)) \text{ a.e. } t\right\}, \tag{34}$$

$$K_2 \doteq \left\{(u,v) \in L^1(0,R)\times L^1(0,R);\ v(t) \ge g(t,u(t)) \text{ a.e. } t\right\}. \tag{35}$$

Since the sets K_j are decomposable, from Theorem 2.1 we deduce that the sets $I(K_j)$, $j = 1,2$, defined in (7), are convex subsets of $\mathbb{R}^2$. From the very definition of $\overline{\phi}$ and ϕ we infer that

$$\begin{aligned} &I(K_1) \subseteq \operatorname{epi}\overline{\phi} \subseteq \overline{I(K_1)}, \quad I(K_2) \subseteq \operatorname{epi}\phi \subseteq \overline{I(K_2)},\\ &\operatorname{epi}\overline{\phi} \setminus \operatorname{graph}\overline{\phi} \subset I(K_1), \quad \operatorname{epi}\phi \setminus \operatorname{graph}\phi \subset I(K_2). \end{aligned} \tag{36}$$

Let us prove that $\overline{\phi}$ is convex. Assume by contradiction that there exist $\alpha,\beta \in \mathbb{R}$, $\lambda \in]0,1[$, and $\varepsilon > 0$ such that $\overline{\phi}(\lambda\alpha + (1-\lambda)\beta) = \lambda\overline{\phi}(\alpha) + (1-\lambda)\overline{\phi}(\beta) + 2\varepsilon$. Since $(\alpha, \overline{\phi}(\alpha)+\varepsilon)$, $(\beta, \overline{\phi}(\beta)+\varepsilon) \in I(K_1)$, from the convexity of $I(K_1)$ we deduce that $(\lambda\alpha + (1-\lambda)\beta, \lambda\overline{\phi}(\alpha) + (1-\lambda)\overline{\phi}(\beta) + \varepsilon) \in I(K_1)$. By the definition of $I(K_1)$ and $\overline{\phi}$, this implies that $\overline{\phi}(\lambda\alpha + (1-\lambda)\beta) \le \lambda\overline{\phi}(\alpha) + (1-\lambda)\overline{\phi}(\beta) + \varepsilon$, a contradiction. The convexity of ϕ can be proved in the same way. ∎

LEMMA 7.9. *There exists a constant* $c \in \mathbb{R}$ *such that* $\overline{\phi}(\alpha) \ge m|\alpha| - c$ *for every* $\alpha \in \mathbb{R}$. *In particular*

$$\lim_{|\alpha|\to+\infty} \overline{\phi}(\alpha) = +\infty. \tag{37}$$

Proof. Let us define $c \doteq \|b\|_{L^1} + 1$, where $b \in L^1(0,R)$ is the function given in (G3). By the very definition of $\overline{\phi}$ we have that, for every fixed $\alpha \in \operatorname{Dom}\overline{\phi}$, there exists $v \in V(\alpha)$ such that $\overline{\Gamma}(v) \le \overline{\phi}(\alpha) + 1$; hence, from (G3),

$$\overline{\phi}(\alpha) \ge \overline{\Gamma}(v) - 1 \ge m\,\|v'\|_{L^1} - c \ge m\left|\int_0^R v'(t)\,dt\right| - c = m|\alpha| - c. \tag{38}$$

Since this inequality is trivially satisfied if $\alpha \notin \operatorname{Dom}\overline{\phi}$, the lemma is proved. □

Remark 7.10. If (G5) holds, then from Lemma 7.8 we deduce that ϕ and $\overline{\phi}$ are finite convex maps on $\mathbb{R}$; hence they are continuous on $\mathbb{R}$. In general, $\operatorname{Dom}\phi$ and $\operatorname{Dom}\overline{\phi}$ can be proper subsets of $\mathbb{R}$.

LEMMA 7.11. *The map $\overline{\phi}$ is continuous relative to J.*

Proof. Since $\overline{\phi}$ is a convex function and, by Theorem 7.7, $J \subseteq \operatorname{Dom}\overline{\phi}$, we have that $\overline{\phi}$ is upper semicontinuous relative to J (see [22, Thm. 10.2]); hence it is enough to prove that it is lower semicontinuous relative to the same interval.

Let $(a_k)_k \subset \operatorname{Dom}\overline{\phi}$, $\lim_k a_k = \alpha \in J$. From Theorem 7.7 there exist $v \in V(\alpha)$, $v_k \in V(a_k)$, $k \in \mathbb{N}$, such that $\overline{\Gamma}(v) = \overline{\phi}(\alpha)$, $\overline{\Gamma}(v_k) \le \overline{\phi}(a_k) + 1/k$, $k \in \mathbb{N}$. From Theorem 7.4 there exists $p \in \mathbb{R}$ such that $p \in \partial g^{**}(t, v'(t))$ a.e. $t \in [0,R]$ so that, for every $k \in \mathbb{N}$,

$$g^{**}(t, v_k'(t)) - g^{**}(t, v(t)) \ge p(v_k'(t) - v'(t)) \quad \text{a.e. } t \in [0,R].$$

Integrating this inequality we get, for every $k \in \mathbb{N}$,

$$\overline{\phi}(a_k) - \overline{\phi}(\alpha) + \frac{1}{k} \ge \overline{\Gamma}(v_k) - \overline{\Gamma}(v) \ge p\int_0^R [v_k'(t) - v'(t)]\,dt = p(a_k - \alpha).$$

Taking the $\liminf$ for $k \to +\infty$ we obtain $\overline{\phi}(\alpha) \le \liminf_k \overline{\phi}(a_k)$, which concludes the proof. □

PROPOSITION 7.12. *The map $\overline{\phi}$ attains its minimum in J; that is,*

$$\inf_{\alpha\in\mathbb{R}} \overline{\phi}(\alpha) = \min_{\alpha\in J} \overline{\phi}(\alpha). \tag{39}$$

Proof. Let us assume that $\alpha_1 \in \mathbb{R}$. Let $v(t) \doteq -\int_t^R h_1(s)\,ds$, $t \in [0,R]$, where h_1 is the function defined in (32). From Remark 7.6 we easily infer that $v \in V(\alpha_1)$, $\overline{\Gamma}(v) = \overline{\phi}(\alpha_1)$, and $h_1(t) \in \partial g^*(t,-m)$ for a.e. $t \in [0,R]$.

Let us fix $\alpha \in \operatorname{Dom}\overline{\phi}$. For every $\varepsilon > 0$ there exists $v_\varepsilon \in V(\alpha)$ such that $\overline{\Gamma}(v_\varepsilon) \le \overline{\phi}(\alpha) + \varepsilon$; hence, recalling that $v'(t) = h_1(t) \in \partial g^*(t,-m)$,

$$\begin{aligned}\overline{\phi}(\alpha) - \overline{\phi}(\alpha_1) &\ge \overline{\Gamma}(v_\varepsilon) - \overline{\Gamma}(v) - \varepsilon \\ &\ge -m\int_0^R [v_\varepsilon'(t) - v'(t)]\,dt - \varepsilon = -m(\alpha - \alpha_1) - \varepsilon.\end{aligned}$$

Since $\varepsilon > 0$ is arbitrary we deduce that

$$\overline{\phi}(\alpha) - \overline{\phi}(\alpha_1) \ge -m(\alpha - \alpha_1) \qquad \forall \alpha \in \mathbb{R}; \tag{40}$$

hence

$$\overline{\phi}(\alpha) - \overline{\phi}(\alpha_1) \ge 0 \qquad \forall \alpha \le \alpha_1. \tag{41}$$

Similarly, if $\alpha_2 \in \mathbb{R}$, with the same argument given above one can prove that

$$\overline{\phi}(\alpha) - \overline{\phi}(\alpha_2) \geq m(\alpha - \alpha_2) \qquad \forall \alpha \in \mathbb{R}; \tag{42}$$

hence

$$\overline{\phi}(\alpha) - \overline{\phi}(\alpha_2) \geq 0 \qquad \forall \alpha \geq \alpha_2. \tag{43}$$

Now (39) follows from (41), (43) and Lemmas 7.9 and 7.11. □

Notice that the analysis above provides an alternative proof of Theorem 7.3. Indeed, from Proposition 7.12, $\overline{\phi}$ reaches its minimum value in a point $\overline{a} \in J$. Let $\bar{v} \in V(\overline{a})$ be such that $\overline{\Gamma}(\bar{v}) = \overline{\phi}(\overline{a})$. For every $v \in V$, setting $\alpha = -v(0)$, we have that $\overline{\Gamma}(\bar{v}) = \overline{\phi}(\overline{a}) \leq \overline{\phi}(\alpha) \leq \overline{\Gamma}(v)$; hence $\bar{v}$ provides a solution to (2).

Now we start studying the nonconvex problem (1). In the following technical lemmas we shall denote by $\overline{G}$ the functional defined by $\overline{G}(u) \doteq \int_0^R g^{**}(t, u(t))\, dt$, $u \in L^1(0, R)$.

LEMMA 7.13. *Assume that* (G1)–(G3) *hold. Let $u \in L^1(0, R)$, and assume that $\overline{G}(u) < +\infty$. Then for every $\varepsilon > 0$ there exists $u_\varepsilon \in L^\infty(0, R)$ such that*

$$\|u_\varepsilon - u\|_{L^1} < \varepsilon, \qquad |\overline{G}(u_\varepsilon) - \overline{G}(u)| < \varepsilon.$$

Proof. For every $k \in \mathbb{N}$ let us define the truncated function $u_k \doteq (u \wedge k) \vee (-k)$. Since $u \in L^1(0, R)$, from the dominated convergence theorem we deduce that $\lim_k \|u_k - u\|_{L^1} = 0$. Furthermore, from (G3) it follows that the convex map $\xi \mapsto g^{**}(t, \xi) + b(t)$ is nonnegative for a.e. $t \in [0, R]$; hence, for every $k \in \mathbb{N}$,

$$0 \leq g^{**}(t, u_k(t)) + b(t) \leq g^{**}(t, u(t)) + g^{**}(t, 0) + 2\, b(t) \quad \text{a.e. } t \in [0, R]. \tag{44}$$

From (G2) and the facts that $\overline{G}(u) < +\infty$ and $b \in L^1(0, R)$, we deduce that the right-hand side of (44) is an integrable function on $[0, R]$. Hence, from the dominated convergence theorem, we conclude that $\lim_k \overline{G}(u_k) = \overline{G}(u)$, completing the proof. □

LEMMA 7.14. *Assume that* (G1) *and* (G4) *hold. Then for every $u_0 \in L^\infty(0, R)$ there exist $u_1, u_2 \in L^\infty(0, R)$ and a measurable function $\lambda\colon [0, R] \to [0, 1]$ such that*

$$u_0(t) = \lambda(t) u_1(t) + (1 - \lambda(t)) u_2(t) \quad \text{a.e. } t \in [0, R], \tag{45}$$

$$g^{**}(t, u_0(t)) = \lambda(t) g(t, u_1(t)) + (1 - \lambda(t)) g(t, u_2(t)) \quad \text{a.e. } t \in [0, R]. \tag{46}$$

Proof. From (G4) and Corollary 3.8 in [13] we have that there exist measurable functions u_1, u_2, and λ satisfying (45) and (46). It remains to prove that $u_1, u_2 \in L^\infty(0, R)$.

From (45) and (46) we have that, for a.e. $t \in [0, R]$, the points $u_j(t)$, $j = 0, 1, 2$, belong to the same face of $\operatorname{epi} g^{**}(t, \cdot)$. In particular, for a.e. $t \in [0, R]$ there exists $p = p(t) \in \mathbb{R}$ such that

$$u_j(t) \in \partial g^*(t, p), \quad j = 0, 1, 2. \tag{47}$$

Choosing $r_1 = \|u_0\|_{L^\infty}$, from (47) and (G4) there exists $r_2 > 0$ such that $\|u_j\|_{L^\infty} \leq r_2$, $j = 1, 2$, concluding the proof. □

PROPOSITION 7.15. *Assume that* (G1)–(G5) *hold. Then $\phi = \overline{\phi}$, and $\overline{I(K_1)} = \overline{I(K_2)}$.*

Proof. The second part of the theorem is a direct consequence of the first one. Since $g \geq g^{**}$ we have the trivial inequality $\phi \geq \overline{\phi}$. Assume by contradiction that there exists a point $\alpha \in \mathbb{R}$ such that $\phi(\alpha) > \overline{\phi}(\alpha)$. Since, from (G5), ϕ and $\overline{\phi}$ are continuous on $\mathbb{R}$, there exists $\varepsilon > 0$ such that

$$\phi(x) \geq \overline{\phi}(\alpha) + 3\varepsilon \quad \forall x \in [\alpha - \varepsilon, \alpha + \varepsilon]. \tag{48}$$

From the very definition of $\overline{\phi}$ there exists $u \in L^1(0,R)$ such that $\int_0^R u(t)\,dt = \alpha$, and $\overline{G}(u) \leq \overline{\phi}(\alpha) + \varepsilon$. From Lemma 7.13 there exists $u_0 \in L^\infty(0,R)$ such that, if $x \doteq \int_0^R u_0(t)\,dt$, then $|x - \alpha| < \varepsilon$ and

$$\overline{G}(u_0) \leq \overline{G}(u) + \varepsilon \leq \overline{\phi}(\alpha) + 2\varepsilon. \tag{49}$$

From Lemma 7.14 there exist functions $u_1, u_2 \in L^\infty(0,R)$ and a measurable map $\lambda\colon [0,R] \to [0,1]$ such that (45) and (46) hold. Since $u_j \in L^\infty(0,R)$, $j = 0,1,2$, it is easy to check that (G3) and (G5) imply that the maps $t \mapsto g^{**}(t, u_j(t))$, $j = 0,1,2$, belong to $L^1(0,R)$. Furthermore, from (46) we deduce that $g(t,u_j(t)) = g^{**}(t,u_j(t))$, $j = 1,2$, for a.e. $t \in [0,R]$; hence the maps $t \mapsto g(t,u_j(t))$, $j = 1,2$, belong to $L^1(0,R)$ too.

From Lyapunov's theorem on the range of nonatomic vector-valued measures (see, for example, [8, Thm. 16.1.v]), there exists a measurable set $E \subseteq [0,R]$ such that, denoting $E^C \doteq [0,R] \setminus E$,

$$\int_0^R [\lambda(t)u_1(t) + (1-\lambda(t))u_2(t)]\,dt = \int_E u_1(t)\,dt + \int_{E^C} u_2(t)\,dt, \tag{50}$$

$$\begin{aligned}\int_0^R [\lambda(t)g(t,u_1(t)) + (1-\lambda(t))g(t,u_2(t))]\,dt \\ = \int_E g(t,u_1(t))\,dt + \int_{E^C} g(t,u_2(t))\,dt.\end{aligned} \tag{51}$$

Let us define $\tilde{u} \doteq u_1\chi_E + u_2\chi_{E^C}$. From (45) and (50) we have that

$$\int_0^R \tilde{u}(t)\,dt = \int_0^R u_0(t)\,dt = x, \tag{52}$$

while from (51), (46), and (49) we deduce that

$$\int_0^R g(t,\tilde{u}(t))\,dt = \int_0^R g^{**}(t,u_0(t))\,dt = \overline{G}(u_0) \leq \overline{\phi}(\alpha) + 2\varepsilon. \tag{53}$$

From (52) and (53) we thus conclude that $\phi(x) \leq \overline{\phi}(\alpha) + 2\varepsilon$. Since $|x - \alpha| < \varepsilon$, this is in contradiction with (48). ∎

We are now in a position to conclude the proof of Theorem 3.2. Let us assume that (G1)–(G5) hold. Since K_2 is decomposable and strongly closed in the L^1 topology, from Theorem 2.1 it follows that $I(K_2)$ is a convex subset of $\mathbb{R}^2$ that contains every extremal compact face of $\overline{I(K_2)}$. From Propositions 7.12 and 7.15 there exists a point $\overline{a} \in J$ where ϕ attains its minimum value. Furthermore, from Lemma 7.9 we have that $\lim_{|\alpha|\to\infty} \phi(\alpha) = +\infty$, so that the point $(\overline{a}, \phi(\overline{a}))$ must belong to a compact extremal face of $\operatorname{epi}\phi = \overline{I(K_2)}$. Henceforth $(\overline{a}, \phi(\overline{a})) \in I(K_2)$; that is, there exists $\bar{v} \in V(\overline{a})$ such that $\Gamma(\bar{v}) = \phi(\overline{a})$, and $\bar{v}$ provides a solution to (1). Finally, the relation (11) follows from the equality $\Gamma(\bar{v}) = \overline{\Gamma}(\bar{v})$. ∎

Remark 7.16. The necessity of the approximation process developed in Lemmas 7.13 and 7.14, and used in Proposition 7.15, depends on the fact that, given $u_0 \in L^1(0,R)$, conditions (G1)–(G5) are not sufficient to provide the existence of integrable functions u_1 and u_2 satisfying (45) and (46).

Indeed, we shall give below an example of a function $g\colon \mathbb{R} \to \mathbb{R}$ satisfying (G1)–(G5) for which the following holds: There exists a function $u_0 \in L^1(0,1)$ such that, for every choice of functions u_1, u_2 satisfying (45) and (46), then at least one of these functions is not integrable on $[0,1]$.

Let us define $a_0 = 0$, $a_k = 2^{k^2}/k^2$ $(k = 1, 2, \dots)$, and let $(b_k)_k$ be the sequence defined by the recurrence relations $b_0 = 0$, $b_{k+1} = b_k + [1 - 1/(k+1)](a_{k+1} - a_k)$, $k = 0, 1, \dots$. It is easy to verify that $(a_k)_k$ is a strictly monotone increasing sequence, and $\lim_k a_k = +\infty$. Let $g\colon \mathbb{R} \to \mathbb{R}$ be a continuous function satisfying the following properties: $g(-\xi) = g(\xi)$ for every ξ, $g(a_k) = b_k$ for every $k \in \mathbb{N}$, and

$$g(\lambda a_k + (1-\lambda)a_{k+1}) > \lambda b_k + (1-\lambda)b_{k+1} \quad \forall \lambda \in]0,1[\ \forall k \in \mathbb{N}. \tag{54}$$

It is easy to show that g^{**} is a piecewise affine function such that

$$\partial g^{**}(\xi) = \begin{cases} \left\{1 - \dfrac{1}{k}\right\} & \text{if } |\xi| \in]a_{k-1}, a_k[,\ k \in \mathbb{N}, \\ \left[1 - \dfrac{1}{k}, 1 - \dfrac{1}{k+1}\right] & \text{if } |\xi| = a_k,\ k = 1, 2, \dots, \\ \{0\} & \text{if } \xi = 0, \end{cases}$$

and, from (54),

$$g^{**}(\xi) = g(\xi) \quad \text{if and only if} \quad |\xi| = a_k \text{ for some } k \in \mathbb{N}. \tag{55}$$

Furthermore, g satisfies (G1)–(G5): concerning (G3), it is enough to choose $m = 1/2$ and $b = 1$.

Since $a_{k+1} > a_k$ for every $k \in \mathbb{N}$ and $\lim_k (a_{k+1} - a_k) = +\infty$, there exists $\varepsilon > 0$ such that $a_k + \varepsilon < a_{k+1}$ for all $k \in \mathbb{N}$. Let us define the function

$$u_0 \doteq \sum_{k=0}^{\infty} (a_k + \varepsilon)\chi_{I_k}, \qquad I_k \doteq \left]2^{-(k+1)^2},\ 2^{-k^2}\right].$$

We claim that $u_0 \in L^1(0,1)$. Indeed, $u_0 \geq 0$, and

$$\begin{aligned} \int_0^1 u_0(t)\,dt &= \sum_{k=0}^{\infty} (a_k + \varepsilon) meas(I_k) \\ &= \sum_{k=0}^{\infty} \left(\frac{2^{k^2}}{k^2} + \varepsilon\right)\left[2^{-k^2} - 2^{-(k+1)^2}\right] < +\infty. \end{aligned}$$

Furthermore, $0 \leq g(\xi) \leq |\xi|$, so that $\overline{G}(u_0) < +\infty$. Now let u_1 and u_2 be two functions satisfying (45) and (46). Since $u_0(t) \in]a_k, a_{k+1}[$ for every $t \in I_k$, from (55) we deduce that, for a.e. $t \in I_k$, either $u_1(t) = a_k$, $u_2(t) = a_{k+1}$ or $u_1(t) = a_{k+1}$,

$u_2(t) = a_k$. Hence

$$\int_0^1 [u_1(t) + u_2(t)]\, dt = \sum_{k=0}^{\infty} (a_k + a_{k+1})\ meas(I_k) \geq \sum_{k=0}^{\infty} a_{k+1}\, meas(I_k)$$
$$= \sum_{k=0}^{\infty} \frac{2^{(k+1)^2}}{(k+1)^2} \left[2^{-k^2} - 2^{-(k+1)^2}\right] = \sum_{k=0}^{\infty} \frac{2^{2k+1} - 1}{(k+1)^2} = +\infty.$$

Thus we can conclude that at least one of the functions u_1, u_2 is not integrable on $[0, 1]$.

Acknowledgments. The author wishes to thank Annalisa Malusa for the helpful discussions during the preparation of the manuscript and the anonymous referees for making suggestions that improved the presentation of the paper.

REFERENCES

[1] L. Ambrosio, O. Ascenzi, and G. Buttazzo, *Lipschitz regularity for minimizers of integral functionals with highly discontinuous integrands*, J. Math. Anal. Appl., 142 (1989), pp. 301–316.
[2] J. M. Ball, *Convexity conditions and existence theorems in nonlinear elasticity*, Arch. Rational Mech. Anal., 63 (1977), pp. 337–403.
[3] P. Bauman and D. Phillips, *A non-convex variational problem related to change of phase*, Appl. Math. Optim., 21 (1990), pp. 113–138.
[4] P. Celada, S. Perrotta, and G. Treu, *Existence of solutions for a class of non convex minimum problems*, Math. Z., 228 (1998), pp. 177–199.
[5] A. Cellina, *Minimizing a functional depending on ∇u and on u*, Ann. Inst. H. Poincaré Anal. Non Linéaire, 14 (1997), pp. 339–352.
[6] A. Cellina and F. Flores, *Radially symmetric solutions of a class of problems of the calculus of variations without convexity assumptions*, Ann. Inst. H. Poincaré Anal. Non Linéaire, 9 (1992), pp. 465–478.
[7] A. Cellina and S. Perrotta, *On minima of radially symmetric functionals of the gradient*, Nonlinear Anal., 23 (1994), pp. 239–249.
[8] L. Cesari, *Optimization—Theory and Applications*, Springer-Verlag, New York, 1983.
[9] F. H. Clarke, *Optimization and Nonsmooth Analysis*, Wiley Interscience, New York, 1983.
[10] F. H. Clarke and P. D. Loewen, *An intermediate existence theory in the calculus of variations*, Ann. Scuola Norm. Sup. Pisa Cl. Sci. (4), 16 (1989), pp. 487–526.
[11] G. Crasta, *An existence result for non-coercive non-convex problems in the calculus of variations*, Nonlinear Anal., 26 (1996), pp. 1527–1533.
[12] G. Crasta, *Existence of minimizers for non-convex variational problems with slow growth*, J. Optim. Theory Appl., 99 (1998), pp. 381–401.
[13] G. Crasta and A. Malusa, *Existence results for noncoercive variational problems*, SIAM J. Control Optim., 34 (1996), pp. 2064–2076.
[14] I. Ekeland and R. Temam, *Convex Analysis and Variational Problems*, North–Holland, Amsterdam, 1977.
[15] N. Fusco, P. Marcellini, and A. Ornelas, *Existence of minimizers for some nonconvex one-dimensional integrals*. Portugal. Math., 55 (1998), pp. 167–185.
[16] B. Kawohl, J. Stara, and G. Wittum, *Analysis and numerical studies of a problem of shape design*, Arch. Rational Mech. Anal., 114 (1991), pp. 349–363.
[17] R. Kohn and G. Strang, *Optimal design and relaxation of variational problems*, I, II *and* III, Comm. Pure Appl. Math., 39 (1976), pp. 113–137, 139–182, 353–377.
[18] C. Marcelli, *One-dimensional non-coercive problems of the calculus of variations*, Ann. Mat. Pura Appl. (4), 173 (1997), pp. 145–161.
[19] C. Olech, *The Lyapunov theorem: Its extensions and applications*, in Methods of Non-convex Analysis, A. Cellina ed., Springer-Verlag, New York, 1991, pp. 84–103.
[20] A. Ornelas, *Existence of scalar minimizers for nonconvex simple integrals of sum type*, J. Math. Anal. Appl., 221 (1998), pp. 559–573.
[21] J. P. Raymond, *Existence and uniqueness results for minimization problems with non-convex functionals*, J. Optim. Theory Appl., 82 (1994), pp. 571–592.

[22] R. T. Rockafellar, *Convex Analysis*, Princeton University Press, Princeton, NJ, 1970.
[23] R. T. Rockafellar, *Existence and duality theorems for convex problems of Bolza*, Trans. Amer. Math. Soc., 159 (1971), pp. 1–40.
[24] R. Tahraoui, *Sur une classe de fonctionnelles non convexes et applications*, SIAM J. Math. Anal., 21 (1990), pp. 37–52.
[25] M. Vornicescu, *A variational problem on subsets of* $\mathbb{R}^n$, Proc. Roy. Soc. Edinburgh Sect. A, 127 (1997), pp. 1089–1101.

SIAM J. CONTROL OPTIM.
Vol. 38, No. 1, pp. 254–271

MINIMAX CONTROL OF PARABOLIC SYSTEMS WITH STATE CONSTRAINTS*

NADIR ARADA† AND JEAN-PIERRE RAYMOND†

Abstract. In this paper we study a minimax control problem for parabolic equations in the presence of pointwise state constraints. The terminology minimax here refers to a cost functional defined with a L^∞-norm. The directional derivatives of the L^∞-norm are elements of $(L^\infty)'$. Therefore, the adjoint equation may involve finitely additive measures in place of Radon measures. To overcome this difficulty, we introduce a compactification (of Stone–Čech type). We prove necessary optimality conditions which are new, both in the case with no state constraints and in the case with state constraints. Under some convexity conditions, these optimality conditions are also sufficient.

Key words. control problems, minimax, supremum norm, pointwise state constraints, compactification

AMS subject classifications. 49K35, 49K20

PII. S0363012998341411

1. Introduction. In this paper, we consider an optimal control problem with a cost functional of the form

$$J(y,u) = ||F(y,u)||_{\infty,Q} = \text{ess-sup}_{(x,t)\in Q}|F(y(x,t),u(x,t))|.$$

The pair $(y,u) \in C(\overline{Q}) \times L^\infty(Q)$ (the state and control variables) satisfies a semilinear parabolic equation

$$\frac{\partial y}{\partial t} + Ay + \Phi(\cdot,y,u) = 0 \quad \text{in } Q, \quad y = \Psi|_\Sigma \quad \text{on } \Sigma, \quad y(0) = \Psi|_{\Omega\times\{0\}} \quad \text{in } \Omega, \tag{1}$$

where Ω is a bounded domain in $\mathbb{R}^N$, $Q = \Omega\times]0,T[$, $\Sigma = \Gamma\times]0,T[$, Γ is the boundary of Ω, $T > 0$, A is a second-order elliptic operator, $\Psi \in C(\overline{\Sigma} \cup (\Omega \times \{0\}))$. The set U_{ad} of admissible controls is a convex subset of $L^\infty(Q)$. Pointwise constraints of the form

$$g(x,t,y(x,t)) \le 0 \qquad \text{for all } (x,t) \in \overline{Q} \tag{2}$$

are imposed on the state variable. In this setting g is a function from $\overline{Q} \times \mathbb{R}$ into $\mathbb{R}$. The paper is concerned with the following control problem:

$$(\mathcal{P}) \qquad \inf\{J(y,u) \mid (y,u) \in C(\overline{Q}) \times U_{ad}, \ (y,u) \text{ satisfies (1) and (2)}\}.$$

As in [5], [13], we here use the terminology minimax because the cost functional is defined by an L^∞-norm. Let us mention that the same terminology is often used in the literature in the game-theoretical sense, that is, minimization with respect to control variables and maximization with respect to disturbances [11]. There is a growing interest in studying problems for functionals defined by a supremum norm [5], [9], [13], [14]. A survey of the existing literature for control problems involving supremum norm functionals, and governed by ordinary differential equations, is given

*Received by the editors July 6, 1998; accepted for publication (in revised form) June 7, 1999; published electronically December 15, 1999.
http://www.siam.org/journals/sicon/38-1/34141.html
†UMR CNRS MIP, Université Paul Sabatier, 31062 Toulouse cedex 4, France (arada@mip.ups-tlse.fr, raymond@mip.ups-tlse.fr).

in the introduction of [5]. For problems governed by partial differential equations, the interest is more recent [9]. To understand the difficulties in deriving optimality conditions, consider the following control problem with no state constraints:

$$(\mathcal{P}_c) \qquad \inf\{I(y,u) = ||G(y,u)||_{C(\overline{Q})} \;\mid\; (y,u) \in C(\overline{Q}) \times U_c, \quad (y,u) \text{ satisfies (1)}\},$$

where U_c is a closed convex set in $C(\overline{Q})$. In this case, for a regular function G, we can prove that if $(\bar{y},\bar{u})$ is an optimal solution of $(\mathcal{P}_c)$, then there exists a Radon measure $\bar{\eta}$ on $\overline{Q}$, that is, a regular Borel measure such that

$$\int_Q \bar{p}\ \Phi'_u(\cdot,\bar{y},\bar{u})(u-\bar{u})\,dx\,dt + \Big\langle \bar{\eta}, G'_u(\bar{y},\bar{u})(u-\bar{u})\Big\rangle_{\mathcal{M}(\overline{Q}),C(\overline{Q})} \geq 0$$

for all $u \in U_c$, where the adjoint state $\bar{p} \in L^1(0,T;W_0^{1,1}(\Omega))$ satisfies

$$(3) \qquad \begin{cases} -\dfrac{\partial \bar{p}}{\partial t} + A^*\bar{p} + \Phi'_y(\cdot,\bar{y},\bar{u})\bar{p} + G'_y(\bar{y},\bar{u})^*\bar{\eta}|_Q = 0 & \text{in } Q, \\ \\ \bar{p}(\cdot,T) + G'_y(\bar{y},\bar{u})^*\bar{\eta}|_{\Omega_T} = 0 & \text{in } \Omega, \end{cases}$$

and

$$\bar{\eta} \neq 0, \qquad \Big\langle \bar{\eta}, z\Big\rangle_{\mathcal{M}(\overline{Q}),C(\overline{Q})} \leq 0 \quad \text{for all } z \in \Big\{ y \in C(\overline{Q}) \mid\; [G(\bar{y},\bar{u});y] < 0\Big\},$$

where $[G(\bar{y},\bar{u});y]$ stands for the derivative of the mapping $z \longrightarrow ||z||_{C(\overline{Q})}$ at $G(\bar{y},\bar{u})$ in the direction y (see section 2). Similar results are obtained by W. Yu for some identification problems [14]. Now, if the set of admissible controls is a subset of $L^\infty(Q)$, it is natural to consider functionals of the form of J. In this case, the main difficulty to derive optimality conditions comes from that the directional derivatives of the mapping $z \longrightarrow ||z||_{\infty,Q}$ are elements of $(L^\infty(Q))'$. Therefore, the corresponding adjoint equation may involve finitely additive measures in place of Radon measures. Contrary to (3), such an equation cannot be studied with the classical tools of distribution theory. To overcome this difficulty, we suggest following the method developed in [1], [3]. By introducing a compactification of the domain Q, we are able to prove a representation theorem for finitely additive measures belonging to $(L^\infty(Q))'$ (Theorem 4.1). Roughly speaking, with each element of $(L^\infty(Q))'$, we associate a Radon measure on $\overline{Q}$ and a bounded linear transformation. With these tools, we obtain new optimality conditions for minimax control problems with (or without) pointwise state constraints. Due to the representation theorem of elements of $(L^\infty(Q))'$, the adjoint equation is written in a classical form (with Radon measures as source terms). The bounded linear transformation intervenes only in optimality conditions for the distributed control.

In Theorem 4.2, optimality conditions are obtained in a Lagrangian form and not in the form of a Pontryagin's principle. Let us explain why. To prove a Pontryagin's principle, we need continuity properties of the cost functional with respect to diffuse perturbations of controls (see [9], [2], [6]). These continuity properties are not satisfied for J (because J is defined by a L^∞-norm). A classical way to overcome this difficulty is to approximate the L^∞-norm by a L^r-norm and to write approximate optimality conditions. But the passage to the limit when r tends to infinity can only be carried out for a subclass of admissible controls [5], [9], [13]. In this case (see [9, p. 203]), it is not obvious to verify if optimality conditions are trivial or not.

By using Lagrangian perturbations, we recover continuity properties for the cost functional, and optimality conditions are stated in Lagrangian form, but they give nontrivial information. In particular, we are able to derive some interesting corollaries (Corollaries 5.4–5.7) from the general result stated in Theorem 4.2.

In the case where $U_{ad} \subset C(\overline{Q})$, we recover classical optimality conditions (Corollary 5.4). For problems with no state constraints, optimality conditions are obtained in qualified form (Corollary 5.1). Moreover, if F is convex, necessary optimality conditions are also sufficient (Corollary 5.2). For problems with state constraints, optimality conditions in qualified form are also sufficient if F is convex (Theorem 4.5).

2. Examples and assumptions. For simplicity, we have considered only an equation with a distributed control. But the results of our paper may be extended to problems with controls on the boundary and in the initial condition. Let us give elementary examples for which optimality conditions of the paper may be applied.

EXAMPLE 1. *Let y_d be a given function in $C(\overline{Q})$. Set*

$$J(y) = ||y - y_d||_{C(\overline{Q})},$$

$$U_{ad} = \{u \in L^\infty(Q) \mid a \le u \le b\}, \qquad a \in L^\infty(Q), \qquad b \in L^\infty(Q).$$

Let y_u be the solution of

$$\frac{\partial y}{\partial t} - \Delta y = u + f \quad \text{in } Q, \quad y = 0 \quad \text{on } \Sigma, \quad y(0) = \Psi \quad \text{in } \Omega,$$

with $f \in L^q(Q)$ $(q > \frac{N}{2} + 1)$, and $\Psi \in C_o(\Omega)$. Suppose that there exists $u \in U_{ad}$ such that $z_1 \le y_u \le z_2$, where z_1, z_2 are two functions in $C(\overline{Q})$. The problem

$$\inf\{J(y_u) \ \mid \ u \in U_{ad}, \quad z_1 \le y_u \le z_2\}$$

admits solutions. Optimality conditions may be deduced from Corollary 5.7.

EXAMPLE 2. *Let y_d be a given function in $L^\infty(Q)$. Define U_{ad} as in Example* 1. *Let y_u be the solution to*

$$\frac{\partial y}{\partial t} - \Delta y + uy = f \quad \text{in } Q, \quad y = 0 \quad \text{on } \Sigma, \quad y(0) = \Psi \quad \text{in } \Omega,$$

where f and Ψ are as in Example 1. *The identification problem*

$$\textit{Minimize} \quad ||y_u - y_d||_{\infty,Q}, \quad u \in U_{ad}$$

admits solutions. Optimality conditions may be deduced from Corollary 5.1.

EXAMPLE 3. *Let us give an example with two controls $u_1 \in L^\infty(Q)$ and $u_2 \in L^\infty(Q)$. Consider the problem*

$$\textit{Minimize} \quad ||y + u_1 - y_d||_{\infty,Q},$$

subject to

$$\frac{\partial y}{\partial t} - \Delta y = u_2 + f \quad \text{in } Q, \quad y = 0 \quad \text{on } \Sigma, \quad y(0) = \Psi \quad \text{in } \Omega,$$

$$-\varepsilon \le u_1 \le \varepsilon, \qquad a \le u_2 \le b,$$

where ε is a positive number, f, Ψ, a, b are as in Example 1. *The control u_1 may play the role of a noise with level ε. The above problem admits solutions. Optimality conditions for this problem may be obtained by adapting the result of Corollary* 5.1 *to the case of two control variables.*

Assumptions and notation. Throughout the paper, Ω denotes a bounded, open, and connected subset in $\mathbb{R}^N$ ($N \geq 2$) of class $C^{2+\gamma}$ for some $0 < \gamma \leq 1$. We denote by Ω_0 (resp., Ω_T) the set $\Omega \times \{0\}$ (resp., $\Omega \times \{T\}$). The second-order elliptic operator A is of the form $Ay(x) = -\sum_{i,j=1}^N D_i(a_{ij}(x)D_j y(x))$. ($D_i$ denotes the partial derivative with respect to x_i.) The coefficients a_{ij} belong to $C^{1+\gamma}(\overline{\Omega})$ and satisfy the conditions

$$a_{ij}(x) = a_{ji}(x) \qquad \text{for every } \ i,j \in \{1,\dots,N\}, \qquad m_0|\chi|^2 \leq \sum_{i,j=1}^N a_{ij}(x)\chi_i\chi_j$$

for all $\chi \in \mathbb{R}^N$ and for all $x \in \overline{\Omega}$, with $m_0 > 0$. For every $1 \leq \ell \leq \infty$, the usual norms in the spaces $L^\ell(Q)$, $L^\ell(\Sigma)$, will be denoted by $||\cdot||_{\ell,Q}$, $||\cdot||_{\ell,\Sigma}$. Throughout the paper $\langle\cdot,\cdot\rangle_{*,Q}$ denotes the duality pairing between the spaces $(L^\infty(Q))'$ and $L^\infty(Q)$. If E is a locally compact subset of $\overline{Q}$, $\mu \in \mathcal{M}_b(E)$ (the space of bounded Radon measures on E) and $y \in C_b(E)$, we set $\langle \mu, y\rangle_{b,E} = \int_E y(x,t)d\mu(x,t)$. We denote by int $\mathcal{C}$, the interior of $\mathcal{C}$ for the usual topology of $C(\overline{Q})$ and by $\mathbf{cl}_\infty S$, the closure of $S \subset L^\infty(Q)$ for the usual topology of $L^\infty(Q)$. For z and χ in $L^\infty(Q)$, we set

$$[z;\chi] = \lim_{\rho \searrow 0} \frac{||z+\rho\chi||_{\infty,Q} - ||z||_{\infty,Q}}{\rho}.$$

The previous limit exists for all z and for all χ, because the mapping $||\cdot||_{\infty,Q}$ is convex and the function $\rho \longrightarrow \frac{||z+\rho\chi||_{\infty,Q}-||z||_{\infty,Q}}{\rho}$ is nondecreasing on $\mathbb{R}_+^*$. Moreover, we can easily see that the mapping $\chi \longrightarrow [z;\chi]$ is Lipschitz of rank 1 from $L^\infty(Q)$ into $\mathbb{R}$ (see [8, Chapter 7]).

A1. Φ is a Carathéodory function from $Q \times \mathbb{R}^2$ into $\mathbb{R}$ (i.e., $\Phi(\cdot,y,u)$ is measurable for all $(y,u) \in \mathbb{R}^2$ and $\Phi(x,t,\cdot)$ is continuous for almost all $(x,t) \in Q$). For almost all $(x,t) \in Q$, $\Phi(x,t,\cdot)$ is of class C^1. Moreover, the following estimates hold:

$$|\Phi(x,t,0,u)| \leq \eta(|u|), \quad 0 \leq \Phi'_y(x,t,y,u) \leq \eta(|y|)\eta(|u|), \quad |\Phi'_u(x,t,y,u)| \leq \eta(|y|)\eta(|u|),$$

where η is a nondecreasing function from $\mathbb{R}^+$ into $\mathbb{R}^+$.

A2. F is of class C^1 on $\mathbb{R}^2$. Moreover, the following estimate holds:

$$|F(y,u)| + |F'_y(y,u)| + |F'_u(y,u)| \leq \eta(|u|)\eta(|y|).$$

A3. U_{ad} is a bounded closed convex subset in $L^\infty(Q)$.

A4. g is continuous on $\overline{Q} \times R$ and, for every $(x,t) \in \overline{Q}$, $g(x,t,\cdot)$ is differentiable and g'_y is continuous on $\overline{Q} \times R$. Moreover, we suppose that there exists $\theta_o > 0$ such that $g(x,t,\Psi(x,t)) \leq -\theta_o$ for all $(x,t) \in \overline{\Sigma} \cup \Omega_0$. ($\Psi$ is the function appearing in (1).)

3. State equation and adjoint equation.

3.1. State equation.

The following result is proved in [2, Proposition 3.9].

PROPOSITION 3.1. *Let a be a nonnegative function in $L^q(Q)$ ($q > \frac{N}{2} + 1$), let ϕ be in $L^q(Q)$, and ψ be in $C(\overline{\Sigma} \cup \Omega_0)$. The solution y of*

$$\frac{\partial y}{\partial t} + Ay + ay = \phi \quad \text{in } Q, \quad y = \psi|_\Sigma \quad \text{on } \Sigma, \quad y(\cdot,0) = \psi(0) \quad \text{in } \Omega, \tag{4}$$

belongs to $C(\overline{Q})$ and satisfies the following estimate:

$$||y||_{C(\overline{Q})} \leq C\Big(||\phi||_{q,Q} + ||\psi||_{C(\overline{\Sigma}\cup\Omega_0)}\Big) \qquad \text{for all } q > \frac{N}{2} + 1,$$

where $C \equiv C(M, \Omega, T, N, q)$.

THEOREM 3.2 (see [2, Theorem 3.11]). *Let* u *be in* $L^\infty(Q)$. *Equation* (1) *admits a unique solution* $y_u \in C(\overline{Q})$ *satisfying*

$$||y_u||_{C(\overline{Q})} \leq C\Big(||u||_{q,Q} + ||\Psi||_{C(\overline{\Sigma}\cup\Omega_0)} + 1\Big) \qquad \text{for all } q > \frac{N}{2} + 1,$$

where $C \equiv C(T, \Omega, N, q)$. *Moreover, the mapping* $u \longrightarrow y_u$ *is continuous from* $L^\infty(Q)$ *into* $C(\overline{Q})$.

3.2. Adjoint equation. Let a be a nonnegative function in $L^\infty(Q)$. Consider the equation

$$-\frac{\partial p}{\partial t} + Ap + ap = \mu|_Q \quad \text{in } Q, \quad p = 0 \quad \text{on } \Sigma, \quad p(T) = \mu|_{\Omega_T} \quad \text{in } \Omega, \tag{5}$$

where $\mu = \mu|_Q + \mu|_{\Omega_T}$ is a bounded Radon measure on $Q \cup \Omega_T$, $\mu|_Q$ is the restriction of μ to Q, and $\mu|_{\Omega_T}$ the restriction of μ to Ω_T.

DEFINITION 3.3. *A function* $p \in L^1(0, T; W_0^{1,1}(\Omega))$ *is a weak solution to* (5) *if and only if*

$$\int_Q \left(-p\frac{\partial z}{\partial t} + \sum_{i,j=1}^N a_{ij} D_j p D_i z + azp \right) dx\, dt = \langle \mu, z \rangle_{b, Q\cup\Omega_T}$$

for all $z \in C^1(\overline{Q}) \cap C_0(Q \cup \Omega_T)$.

THEOREM 3.4 (see [2, Theorem 4.2]). *Let* a *be a nonnegative function in* $L^\infty(Q)$ *and let* $\mu \in \mathcal{M}_b(Q \cup \Omega_T)$. *Equation* (5) *admits a unique weak solution* p *in* $L^1(0, T; W_0^{1,1}(\Omega))$. *The function* p *belongs to* $L^\delta(0, T; W_0^{1,d}(\Omega))$, *for every* (δ, d) *satisfying* $\delta \geq 1$, $d \geq 1$, $\frac{N}{2d} + \frac{1}{\delta} > \frac{N+1}{2}$ *and satisfies*

$$||p||_{L^\delta(0,T;W^{1,d}(\Omega))} \leq C||\mu||_{\mathcal{M}_b(Q\cup\Omega_T)},$$

where $C \equiv C(\Omega, T, \delta, d)$ *is independent of* a. *Moreover, there exist a function* $\frac{\partial p}{\partial n_A} \in L^1(\Sigma)$ *and a function* $p(0) \in L^1(\Omega)$ *such that*

$$\int_Q \left(\frac{\partial z}{\partial t} + Az + az \right) p\, dx\, dt = \langle \mu, z \rangle_{b, Q\cup\Omega_T} + \int_\Sigma z \frac{\partial p}{\partial n_A}\, ds\, dt - \int_\Omega z(0)p(0)\, dx$$

for all $z \in Y_q = \{y \in C_b(Q\cup\Omega_T) \mid \frac{\partial y}{\partial t} + Ay \in L^q(Q),\ y|_\Sigma \in L^\infty(\Sigma),\ \text{and } y(0) \in C(\overline{\Omega})\}$.

4. Statement of the main results. Every $\zeta \in (L^\infty(Q))'$ is identified with a measure $\hat{\hat{\zeta}} \in \mathcal{M}(\overline{Q} \times Q^\#)$, where $Q^\#$ is a compactification of Q (see section 6). For notational simplicity, $\hat{\hat{\zeta}}$ will still be denoted by ζ. We denote by $\mathcal{B}$ the canonical projection from $\mathcal{M}(\overline{Q} \times Q^\#)$ onto $\mathcal{M}(\overline{Q})$ defined by

$$\mathcal{B} : \zeta \in \mathcal{M}(\overline{Q} \times Q^\#) \longrightarrow \mathcal{B}(\zeta) \in \mathcal{M}(\overline{Q}),$$

$$\Big\langle \mathcal{B}(\zeta), \phi \Big\rangle_{\mathcal{M}(\overline{Q}), C(\overline{Q})} = \Big\langle \zeta, \phi \Big\rangle_{\mathcal{M}(\overline{Q}\times Q^\#), C(\overline{Q}\times Q^\#)} \qquad \text{for all } \phi \in C(\overline{Q}).$$

Throughout the paper, for any $\zeta \in (L^\infty(Q))'$, $|\zeta|$ stands for the total variation of ζ, $\mathcal{B}(|\zeta|)$ is the canonical projection of $|\zeta|$ onto $\mathcal{M}(\overline{Q})$, and $L^\infty_{\mathcal{B}(|\zeta|)}(\overline{Q})$ denotes the space of $\mathcal{B}(|\zeta|)$-essentially bounded $\mathcal{B}(|\zeta|)$-measurable functions on $\overline{Q}$. The proofs of the following theorems are given in sections 6 and 7.

THEOREM 4.1. *Let $\zeta \in (L^\infty(Q))'$. There exists a bounded linear transformation $\Lambda_\zeta : L^\infty(Q) \longrightarrow L^\infty_{\mathcal{B}(|\zeta|)}(\overline{Q})$ such that*

$$(6)\quad \Big\langle \zeta, h\phi \Big\rangle_{*,Q} = \int_{\overline{Q}} \Lambda_\zeta(h)\phi \; d\mathcal{B}(|\zeta|) \qquad \textit{for all } h \in L^\infty(Q) \;\; \textit{and for all } \phi \in C(\overline{Q}).$$

If $\mathcal{A}$ is an open subset of $\overline{Q}$, then we have

$$(7)\qquad \int_{\overline{Q}} \Lambda_\zeta(\tilde{h}) \; d\mathcal{B}(|\zeta|) = \Big\langle \mathcal{B}(\zeta), \tilde{h} \Big\rangle_{b,\mathcal{A}} + \int_{\overline{Q}\setminus\mathcal{A}} \Lambda_\zeta(\tilde{h}) \; d\mathcal{B}(|\zeta|)$$

for all $\tilde{h} \in C_b(\mathcal{A}) \cap L^\infty(Q)$. Moreover,

$$(8)\quad \Big\langle \ell^*\zeta, \phi \Big\rangle_{*,Q} = \Big\langle \mathcal{B}(\zeta), \ell\phi \Big\rangle_{b,\mathcal{A}} + \Big\langle \mathcal{B}(\ell^*\zeta), \phi \Big\rangle_{b,\overline{Q}\setminus\mathcal{A}}$$

for all $\phi \in C(\overline{Q})$ and for all $\ell \in (C_b(\mathcal{A}) \cap L^\infty(Q))$, where $\ell^*\zeta$ is the measure defined by

$$\Big\langle \ell^*\zeta, h \Big\rangle_{*,Q} = \Big\langle \zeta, \ell h \Big\rangle_{*,Q} \qquad \textit{for all } h \in L^\infty(Q).$$

We shall say that $(\bar{y}, \bar{u})$ satisfies the regularity condition **(R)** if there exists (z_o, u_o) in $C(\overline{Q}) \times U_{ad}$ such that

$$g(\cdot, \bar{y}) + g'_y(\cdot, \bar{y}) \; z_o \in \operatorname{int} \mathcal{C},$$

$$\frac{\partial z_o}{\partial t} + Az_o + \Phi'_y(\cdot, \bar{y}, \bar{u})z_o = -\Phi'_u(\cdot, \bar{y}, \bar{u})(u_o - \bar{u}) \; \textit{in } Q,$$

$$z_o = 0 \qquad \textit{on } \Sigma, \qquad z_o(0) = 0 \qquad \textit{in } \Omega.$$

Throughout the paper, we suppose that $(\bar{y}, \bar{u})$ is an admissibe solution for $(\mathcal{P})$, satisfying $||F(\bar{y}, \bar{u})||_{\infty,Q} > 0$.

THEOREM 4.2. *If* A1–A4 *are fulfilled, and if $(\bar{y}, \bar{u})$ is a solution of $(\mathcal{P})$, then there exist $\bar{\mu} \in \mathcal{M}(\overline{Q})$, $\bar{\zeta} \in (L^\infty(Q))'$, and $\bar{p} \in L^1(0, T; W_0^{1,1}(\Omega))$ such that the following conditions hold*:

- *Nontriviality condition*:

$$(9)\qquad (\bar{\zeta}, \bar{\mu}|_{Q\cup\Omega_T}) \neq (0, 0), \qquad \bar{\mu} \geq 0.$$

- *Complementary conditions:*

$$(10)\qquad \Big\langle \bar{\mu}, z - g(\cdot, \bar{y}) \Big\rangle_{\mathcal{M}(\overline{Q}), C(\overline{Q})} \leq 0 \quad \textit{for all } z \in \mathcal{C},$$

$$(11)\qquad \Big\langle \bar{\zeta}, \chi \Big\rangle_{*,Q} \leq 0 \quad \textit{for all } \chi \in \mathbf{cl}_\infty \Big\{ y \in L^\infty(Q) \mid [F(\bar{y}, \bar{u}); y] < 0 \Big\}.$$

$$(12)\;\; \textit{If } \bar{\mu}|_{Q\cup\Omega_T} = 0, \textit{ then } \langle \bar{\zeta}, \chi \rangle_{*,Q} < 0 \quad \textit{for all } \chi \in \Big\{ y \in L^\infty(Q) \mid [F(\bar{y}, \bar{u}); y] < 0 \Big\}.$$

- *Adjoint equation:*

$$(13)\qquad \begin{cases} -\dfrac{\partial \bar p}{\partial t} + A\bar p + \Phi'_y(\cdot,\bar y,\bar u)\bar p + g'_y(\cdot,\bar y)^*\bar\mu|_Q + \mathcal{B}(F'_y(\bar y,\bar u)^*\bar\zeta)|_Q = 0 & \text{in } Q,\\ \bar p(\cdot,T) + g'_y(\cdot,\bar y)^*\bar\mu|_{\Omega_T} + \mathcal{B}(F'_y(\bar y,\bar u)^*\bar\zeta)|_{\Omega_T} = 0 & \text{in } \Omega. \end{cases}$$

- *Optimality condition for $\bar u$:*

$$(14)\qquad \int_Q \bar p\ \Phi'_u(x,t,\bar y,\bar u)(u-\bar u)\,dx\,dt + \Big\langle \bar\zeta, F'_u(\bar y,\bar u)(u-\bar u)\Big\rangle_{*,Q} \ge 0 \qquad \text{for all } u\in U_{ad},$$

where $g'_y(\cdot,\bar y)^*\bar\mu\,|_Q$ is the restriction of $g'_y(\cdot,\bar y)^*\bar\mu$ to Q, $g'_y(\cdot,\bar y)^*\bar\mu\,|_{\Omega_T}$ is the restriction of $g'_y(\cdot,\bar y)^*\bar\mu$ to Ω_T, $\mathcal{B}(F'_y(\bar y,\bar u)^*\bar\zeta)\,|_Q$ is the restriction of $\mathcal{B}(F'_y(\bar y,\bar u)^*\bar\zeta)$ to Q, $\mathcal{B}(F'_y(\bar y,\bar u)^*\bar\zeta)|_{\Omega_T}$ is the restriction of $\mathcal{B}(F'_y(\bar y,\bar u)^*\bar\zeta)$ to Ω_T, $g'_y(\cdot,\bar y)^*\bar\mu$ and $F'_y(\bar y,\bar u)^*\bar\zeta$ are defined by

$$\Big\langle g'_y(\cdot,\bar y)^*\bar\mu, z\Big\rangle_{\mathcal{M}(\overline{Q}),C(\overline{Q})} = \Big\langle \bar\mu, g'_y(\cdot,\bar y)z\Big\rangle_{\mathcal{M}(\overline{Q}),C(\overline{Q})} \quad \text{for all } z\in C(\overline{Q}),$$

$$\Big\langle F'_y(\bar y,\bar u)^*\bar\zeta, h\Big\rangle_{*,Q} = \Big\langle \bar\zeta, F'_y(\bar y,\bar u)h\Big\rangle_{*,Q} \quad \text{for all } h\in L^\infty(Q).$$

Moreover, if $(\bar y,\bar u)$ satisfies the regularity condition (**R**)*, then $\bar\zeta\neq 0$.*

REMARK 4.3. (i) Notice that for every $\varepsilon\in]0,1[$, we have

$$\Big[F(\bar y,\bar u); -\varepsilon F(\bar y,\bar u)\Big] = -\varepsilon\ ||F(\bar y,\bar u)||_{\infty,Q} < 0.$$

It follows that 0 belongs to $\mathbf{cl}_\infty\{\chi\in L^\infty(Q)\mid [F(\bar y,\bar u);\chi]<0\}$.

(ii) In the same way, if $z\in L^\infty(Q)$ satisfies $z\neq 0$ and $||z||_{\infty,Q}\le ||F(\bar y,\bar u)||_{\infty,Q}$, then there exists $(z_\varepsilon)_\varepsilon\subset L^\infty(Q)$ such that

$$||z_\varepsilon||_{\infty,Q} < ||z||_{\infty,Q}\le ||F(\bar y,\bar u)||_{\infty,Q} \quad \text{and} \quad ||z_\varepsilon - z||_{\infty,Q}\longrightarrow 0 \ \text{ when } \varepsilon\to 0.$$

It is easy to see that $(z_\epsilon - F(\bar y,\bar u))_\varepsilon\subset\{\chi\in L^\infty(Q)\mid [F(\bar y,\bar u);\chi]<0\}$. Therefore, $z-F(\bar y,\bar u)$ belongs to $\mathbf{cl}_\infty\{\chi\in L^\infty(Q)\mid [F(\bar y,\bar u);\chi]<0\}$. Due to (11), we have

$$(15)\qquad \Big\langle \bar\zeta, z-F(\bar y,\bar u)\Big\rangle_{*,Q}\le 0 \quad \text{for all } z \text{ satisfying } ||z||_{\infty,Q}\le ||F(\bar y,\bar u)||_{\infty,Q}.$$

REMARK 4.4. Notice that $\bar\mu$ is a measure on $\overline{Q}$, but the nontriviality condition is stated with the restriction of μ to $Q\cup\Omega_T$.

THEOREM 4.5 (sufficient optimality conditions). *Suppose that* A1–A4 *are fulfilled. Suppose in addition that Φ is of the form $\Phi(\cdot,y,u)=\beta(\cdot)\,y-u$, that $g(\cdot,y)=y$, and that F is convex. If $(\bar y,\bar u,\bar p)\in C(\overline{Q})\times U_{ad}\times L^1(0,T;W^{1,1}(\Omega))$ satisfies* (1)–(2) *together with* (9)–(14) *for $\bar\zeta\neq 0$, then $(\bar y,\bar u)$ is an optimal pair for problem $(\mathcal{P})$.*

REMARK 4.6. This result is still true with obvious modifications for constraints of the form $y\in\mathcal{C}$, where $\mathcal{C}$ is a closed convex subset of $C(\overline{Q})$, with nonempty interior in $C(\overline{Q})$. The proof of Theorem 4.5 is similar to the one given in Corollary 5.2.

5. Some applications.

5.1. Problems with no state constraints. In this section, we are interested in the following control problem:

$$(\widetilde{\mathcal{P}}) \qquad \inf\{J(y,u) \ \mid \ (y,u) \in C(\overline{Q}) \times U_{ad}, \ \ (y,u) \text{ satisfies (1)}\}.$$

Since there is no state constraints stated in $(\widetilde{\mathcal{P}})$, we are able to give necessary optimality conditions in qualified form.

COROLLARY 5.1. *If* A1–A3 *are fulfilled and if* $(\bar{y},\bar{u})$ *is a solution of* $(\widetilde{\mathcal{P}})$*, then there exist* $\bar{\zeta} \in (L^\infty(Q))'$ *and* $\bar{p} \in L^1(0,T;W_0^{1,1}(\Omega))$ *such that the following conditions hold*:

$$\bar{\zeta} \neq 0, \tag{16}$$

$$\left\langle \bar{\zeta}, y \right\rangle_{*,Q} < 0 \quad \textit{for all } y \in \left\{ \chi \in L^\infty(Q) \mid [F(\bar{y},\bar{u});\chi] < 0 \right\}, \tag{17}$$

$$\begin{cases} -\dfrac{\partial \bar{p}}{\partial t} + A\bar{p} + \Phi'_y(x,t,\bar{y},\bar{u})\bar{p} + \mathcal{B}(F'_y(\bar{y},\bar{u})^*\zeta)|_Q = 0 & \textit{in } Q, \\ \bar{p}(x,T) + \mathcal{B}(F'_y(\bar{y},\bar{u})^*\zeta)|_{\Omega_T} = 0 & \textit{in } \Omega, \end{cases} \tag{18}$$

$$\int_Q \bar{p}\ \Phi'_u(x,t,\bar{y},\bar{u})(u-\bar{u})\,dx\,dt + \left\langle \bar{\zeta}, F'_u(\bar{y},\bar{u})(u-\bar{u}) \right\rangle_{*,Q} \geq 0 \tag{19}$$

for all $u \in U_{ad}$.

Proof. Due to Theorem 4.2, there exist $\bar{\mu} \in \mathcal{M}(\overline{Q})$, $\bar{\zeta} \in (L^\infty(Q))'$, and $\bar{p} \in L^1(0,T;W_0^{1,1}(\Omega))$ such that (9)–(14) hold. From (10), we deduce

$$\left\langle \bar{\mu}, z - g(\cdot,\bar{y}) \right\rangle_{\mathcal{M}(\overline{Q}),C(\overline{Q})} \leq 0 \quad \text{for all } z \in C(\overline{Q}).$$

Therefore, we obtain

$$\langle \bar{\mu}, z \rangle_{\mathcal{M}(\overline{Q}),C(\overline{Q})} \leq 0 \quad \text{for all } z \in C(\overline{Q}),$$

and thus $\bar{\mu} \equiv 0$. Taking (9) and (12) into account, we deduce (16) and (17). ∎

COROLLARY 5.2 (sufficient optimality conditions). *Suppose that* A1–A3 *are fulfilled. Suppose in addition that* Φ *is of the form* $\Phi(\cdot,y,u) = \beta(\cdot)\ y - u$ *and that* F *is convex. If* $(\bar{y},\bar{u},\bar{p},\bar{\zeta}) \in C(\overline{Q}) \times U_{ad} \times L^1(0,T;W_0^{1,1}(\Omega)) \times (L^\infty(Q))'$ *satisfies* (1) *together with* (16)–(19)*, then* $(\bar{y},\bar{u})$ *is an optimal pair for the problem* $(\widetilde{\mathcal{P}})$.

Proof. Let $(\bar{y},\bar{u},\bar{p},\bar{\zeta})$ be in $C(\overline{Q}) \times U_{ad} \times L^1(0,T;W_0^{1,1}(\Omega)) \times (L^\infty(Q))'$, satisfying (1) together with (16)–(19). Due to Theorem 3.4, $\bar{p}$ satisfies

$$\int_Q \bar{p}\left(\frac{\partial z}{\partial t} + Az + \beta z\right) dx\,dt - \int_\Sigma \frac{\partial \bar{p}}{\partial n_A} z\ \ ds\,dt + \int_\Omega \bar{p}(0)z(0)\ \ dx \tag{20}$$

$$= -\left\langle \mathcal{B}(F'_y(\bar{y},\bar{u})^*\bar{\zeta}), z \right\rangle_{b,Q\cup\Omega_T} \qquad \text{for all } z \in Y_q.$$

Let u be in U_{ad} and let y_u be the solution of (1) corresponding to u. Since the function $\bar{y} - y_u$ belongs to $C_0(Q \cup \Omega_T)$, by setting $z = \bar{y} - y_u$ in (20), we obtain

$$\int_Q p(\bar u-u)dx\,dt=\int_Q p\left(\frac{\partial(\bar y-y_u)}{\partial t}+A(\bar y-y_u)+\beta(\bar y-y_u)\right)dx\,dt$$
$$=-\Big\langle\mathcal{B}(F'_y(\bar y,\bar u)^*\bar\zeta),\bar y-y_u\Big\rangle_{b,Q\cup\Omega_T}=-\Big\langle\mathcal{B}(F'_y(\bar y,\bar u)^*\bar\zeta),\bar y-y_u\Big\rangle_{\mathcal{M}(\overline{Q}),C(\overline{Q})}$$
$$=-\Big\langle\bar\zeta,F'_y(\bar y,\bar u)(\bar y-y_u)\Big\rangle_{*,Q}.$$

Consequently, from (19), it follows that

$$\Big\langle\bar\zeta,F'_y(\bar y,\bar u)(y_u-\bar y)+F'_u(\bar y,\bar u)(u-\bar u)\Big\rangle_{*,Q}\geq 0. \tag{21}$$

On the other hand, for $\rho\in]0,1[$ and $u\in U_{ad}$, let us set $u_\rho=\bar u+\rho(u-\bar u)$, and let y_{u_ρ} be the solution of (1) corresponding to u_ρ. It is clear that (y_{u_ρ},u_ρ) is admissible for $(\widetilde{\mathcal{P}})$. By applying Lemma 7.1, and using the convexity of F, we obtain

$$\lim_{\rho\searrow 0}\frac{J(y_{u_\rho},u_\rho)-J(\bar y,\bar u)}{\rho}=\Big[F(\bar y,\bar u);\ F'_y(\bar y,\bar u)(y_u-\bar y)+F'_u(\bar y,\bar u)(u-\bar u)\Big]$$
$$\leq J(y_u,u)-J(\bar y,\bar u).$$

We claim that $[F(\bar y,\bar u);\ F'_y(\bar y,\bar u)(y_u-\bar y)+F'_u(\bar y,\bar u)(u-\bar u)]\geq 0$. Argue by contradiction and suppose the contrary. From (17), it follows that

$$\Big\langle\bar\zeta,F'_y(\bar y,\bar u)(y_u-\bar y)+F'_u(\bar y,\bar u)(u-\bar u)\Big\rangle_{*,Q}<0,$$

which contradicts (21). Therefore $J(y_u,u)\geq J(\bar y,\bar u)$. □

5.2. Problems with partially continuous controls. In this section, we are concerned with problems of the form $(\mathcal{P})$ when the admissible controls are continuous on an open set $\mathcal{A}$ of $\overline{Q}$. In particular, when $\mathcal{A}=\overline{Q}$, we recover the result stated in the introduction for the problem $(\mathcal{P}_c)$ (see Remark 5.5). In this case, the multiplier $\bar\zeta\in(L^\infty(Q))'$ may be replaced by $\mathcal{B}(\bar\zeta)$, that is, a Radon measure on $\overline{Q}$. If the set of admissible controls are continuous on an open subset $\mathcal{A}\subset\overline{Q}$, then we show that the operator $\Lambda_{\bar\zeta}$ intervenes only on $\overline{Q}\setminus\mathcal{A}$.

COROLLARY 5.3. *Assume that* A1–A4 *are fulfilled and that* $U_{ad}\subset C(\overline{Q})$. *If* $(\bar y,\bar u)$ *is a solution of* $(\mathcal{P})$, *then there exist* $\bar\mu\in\mathcal{M}(\overline{Q})$ *and* $\bar\eta\in\mathcal{M}(\overline{Q})$ *such that* $(\bar\mu,\bar\eta,\bar p)$ *satisfies* (9), (10), (11), (12), *and* (22), *where* $\bar p\in L^1(0,T;W_0^{1,1}(\Omega))$ *is the solution of*

$$\begin{cases}-\dfrac{\partial\bar p}{\partial t}+A\bar p+\Phi'_y(\cdot,\bar y,\bar u)\bar p+g'_y(\cdot,\bar y)^*\bar\mu|_Q+F'_y(\bar y,\bar u)^*\bar\eta|_Q=0 & \text{in } Q,\\[2ex] \bar p(T)+g'_y(\cdot,\bar y)^*\bar\mu|_{\Omega_T}+F'_y(\bar y,\bar u)^*\bar\eta|_{\Omega_T}=0 & \text{in } \Omega.\end{cases}$$

Proof. The proof is based on arguments similar to those of Corollary 5.4 and is omitted. The corollary corresponds to the case when $\mathcal{A}=\overline{Q}$. □

COROLLARY 5.4. *Let* $\mathcal{A}$ *be an open subset of* $\overline{Q}$. *Assume that* A1–A4 *are fulfilled and that* $U_{ad}\subset(C_b(\mathcal{A})\cap L^\infty(Q))$. *If* $(\bar y,\bar u)$ *is a solution of* $(\mathcal{P})$, *then there exist* $\bar\mu\in\mathcal{M}^+(\overline{Q})$ *and* $\bar\zeta\in(L^\infty(Q))'$ *such that* $(\bar\mu,\bar\zeta)$ *satisfies* (9), (10), (11), *and* (12),

$$\int_Q\bar p\ \Phi'_u(x,t,\bar y,\bar u)(u-\bar u)\,dx\,dt \tag{22}$$

$$+\Big\langle \mathcal{B}(\bar\zeta), F'_u(\bar y,\bar u)(u-\bar u)\Big\rangle_{b,\mathcal{A}} + \int_{\overline{Q}\setminus\mathcal{A}} \Lambda_{\bar\zeta}\Big(F'_u(\bar y,\bar u)(u-\bar u)\Big)\, d\mathcal{B}(|\bar\zeta|) \ge 0$$

for all $u \in U_{ad}$, where $\Lambda_{\bar\zeta}$ is the operator associated with $\bar\zeta$ and defined in Theorem 4.1 *and $\bar p \in L^1(0,T;W_0^{1,1}(\Omega))$ is the solution of*

$$\begin{cases} -\dfrac{\partial \bar p}{\partial t} + A\bar p + \Phi'_y(\cdot,\bar y,\bar u)\bar p + g'_y(\cdot,\bar y)^*\bar\mu|_Q + F'_y(\bar y,\bar u)^*\mathcal{B}(\bar\zeta)|_{\mathcal{A}} \\ \\ +\mathcal{B}(F'_y(\bar y,\bar u)^*\bar\zeta)|_{Q\setminus\mathcal{A}} = 0 \qquad \text{in } Q, \\ \\ \bar p(T) + g'_y(\cdot,\bar y)^*\bar\mu|_{\Omega_T} + \mathcal{B}(F'_y(\bar y,\bar u)^*\bar\zeta)|_{\Omega_T} = 0 \qquad \text{in } \Omega. \end{cases}$$

Proof. Due to Theorem 4.2, there exist $\bar\mu \in \mathcal{M}(\overline{Q})$, $\bar\zeta \in (L^\infty(Q))'$, and $\bar p \in L^1(0,T;W_0^{1,1}(\Omega))$ such that (9)–(14) hold. Since $U_{ad} \subset C_b(\mathcal{A}) \cap L^\infty(Q)$, for every $u \in U_{ad}$ and every $z \in C(\overline{Q})$, the functions $F'_y(\bar y,\bar u)z$ and $F'_u(\bar y,\bar u)u$ belong to $C_b(\mathcal{A}) \cap L^\infty(Q)$. From Theorem 4.1, it follows that

$$\Big\langle \bar\zeta, F'_u(\bar y,\bar u)u\Big\rangle_{*,Q} = \Big\langle \mathcal{B}(\bar\zeta), F'_u(\bar y,\bar u)u\Big\rangle_{b,\mathcal{A}} + \int_{\overline{Q}\setminus\mathcal{A}} \Lambda_{\bar\zeta}\Big(F'_u(\bar y,\bar u)u\Big)\, d\mathcal{B}(|\zeta|)$$

for all $u \in U_{ad}$. On the other hand, with the definition of $\bar p$ and with (8), we have

$$\begin{aligned} &\int_Q \left(\bar p\frac{\partial z}{\partial t} + \sum_{i,j} a_{ij}D_i z D_j \bar p + \Phi'_y(x,t,\bar y,\bar u)\bar p z\right) dx\,dt \\ &= -\Big\langle g'_y(\cdot,\bar y)^*\bar\mu + \mathcal{B}(F'_y(\bar y,\bar u)^*\bar\zeta), z\Big\rangle_{b,Q\cup\Omega_T} = -\Big\langle \bar\mu, g'_y(\cdot,\bar y)z\Big\rangle_{b,Q\cup\Omega_T} \\ &-\Big\langle F'_y(\bar y,\bar u)^*\mathcal{B}(\bar\zeta), z\Big\rangle_{b,\mathcal{A}} - \Big\langle \mathcal{B}(F'_y(\bar y,\bar u)^*\bar\zeta), z\Big\rangle_{b,(Q\setminus\mathcal{A})\cup\Omega_T} \end{aligned}$$

for every $z \in C^1(\overline{Q}) \cap C_0(Q\cup\Omega_T)$. The proof is complete. ∎

REMARK 5.5. If there are no state constraints, by using Corollaries 5.1 and 5.3, we recover the optimality conditions stated in the introduction for problem $(\mathcal{P}_c)$.

5.3. Comparison with other results. In this section, we apply our general result to problem $(\mathcal{P})$ when the function F is assumed to satisfy the following additional assumption.

A6. There exists a positive constant a such that $a \le F(y,u)$ for all $(y,u) \in \mathbb{R}^2$.

This kind of assumption is used in [9, p. 198], [13], and [5].

COROLLARY 5.6. *Assume that* A1–A4 *and* A6 *are fulfilled. If $(\bar y,\bar u)$ is a solution of $(\mathcal{P})$, then there exist $\bar\mu \in \mathcal{M}(\overline{Q})$, $\bar\zeta \in (L^\infty(Q))'$, and $\bar p \in L^1(0,T;W_0^{1,1}(\Omega))$ such that the conditions* (9)–(14) *hold. Moreover, we have*

$$\bar\zeta \ge 0,$$

$$\int_{Q_0} \bar p\ \Phi'_u(x,t,\bar y,\bar u)(u-\bar u)\,dx\,dt \ge 0 \qquad \text{for all } u \in U_0, \tag{23}$$

where

$$Q_0 = \{(x,t) \in Q \mid F(\bar y(x,t),\bar u(x,t)) < ||F(\bar y,\bar u)||_{\infty,Q}\},$$

$$U_0 = \{u \in U_{ad} \mid u(x,t) = \bar{u}(x,t) \quad \textit{for almost all } (x,t) \in Q \setminus Q_0\}.$$

Proof. Due to Theorem 4.2, there exist $\bar{\mu} \in \mathcal{M}(\overline{Q})$, $\bar{\zeta} \in (L^\infty(Q))'$, and $\bar{p} \in L^1(0,T;W_0^{1,1}(\Omega))$ such that (9)–(14) hold.

• Let us prove that $\bar{\zeta} \geq 0$. Let $\phi \in L^\infty(Q)$ be such that $\phi \geq 0$ and $||\phi||_{\infty,Q} \neq 0$. It is clear that

$$0 \leq \frac{\phi(x,t)}{||\phi||_{\infty,Q}}\, a \leq a \leq F(\bar{y}(x,t), \bar{u}(x,t)) \qquad \text{for all } (x,t) \in Q.$$

Taking (15) into account, and since $0 < a \leq F$, we have

$$\left\langle \bar{\zeta}, \frac{\phi\ a}{||\phi||_{\infty,Q}} \right\rangle_{*,Q} = \left\langle \bar{\zeta}, F(\bar{y},\bar{u}) - \left(F(\bar{y},\bar{u}) - \frac{\phi\ a}{||\phi||_{\infty,Q}} \right) \right\rangle_{*,Q} \geq 0.$$

Therefore, $\langle \bar{\zeta}, \phi \rangle_{*,Q} \geq 0$ for all nonnegative functions $\phi \in L^\infty(Q)$.

• Let us prove (23). Suppose that meas $(Q_0) > 0$, and let $\epsilon > 0$ be such that $Q_\epsilon = \{(x,t) \in Q \mid F(\bar{y}(x,t), \bar{u}(x,t)) \leq ||F(\bar{y},\bar{u})||_{\infty,Q} - \epsilon\} \neq \emptyset$.

Let us set

$$U_\epsilon = \{u \in U_{ad} \mid u(x,t) = \bar{u}(x,t) \ \text{ for almost all } (x,t) \in Q \setminus Q_\epsilon\}.$$

First, we prove that $\bar{\zeta}(Q_\epsilon) = 0$ for all $\epsilon > 0$. Suppose that $\bar{\zeta}(Q_\epsilon) > 0$ for some $\epsilon > 0$, and consider

$$F_\epsilon(s,t) = \begin{cases} F(\bar{y}(x,t), \bar{u}(x,t)) & \text{in } Q \setminus Q_\epsilon, \\ \\ F(\bar{y}(x,t), \bar{u}(x,t)) + \frac{\epsilon}{2} & \text{in } Q_\epsilon. \end{cases}$$

Due to the definition of Q_ϵ, we have

$$||F_\epsilon||_{\infty,Q} = \max\left(||F_\epsilon||_{\infty,Q\setminus Q_\epsilon}, ||F_\epsilon||_{\infty,Q_\epsilon} \right)$$

$$= \max\left(||F(\bar{y},\bar{u})||_{\infty,Q\setminus Q_\epsilon}, ||F(\bar{y},\bar{u}) + \frac{\epsilon}{2}||_{\infty,Q_\epsilon} \right)$$

$$\leq \max\left(||F(\bar{y},\bar{u})||_{\infty,Q\setminus Q_\epsilon}, ||F(\bar{y},\bar{u})||_{\infty,Q} - \frac{\epsilon}{2} \right) \leq ||F(\bar{y},\bar{u})||_{\infty,Q}.$$

By using (15), we obtain

$$\frac{\epsilon}{2}\ \bar{\zeta}(Q_\epsilon) = \left\langle \bar{\zeta}, F_\epsilon - F(\bar{y},\bar{u}) \right\rangle_{*,Q} \leq 0.$$

This contradicts $\bar{\zeta}(Q_\epsilon) > 0$. Consequently, $\bar{\zeta}(Q_\epsilon) = 0$ for all $\epsilon > 0$. Therefore, from the definition of U_ϵ, we deduce that

$$\left\langle \bar{\zeta}, F'_u(\bar{y},\bar{u})(u-\bar{u}) \right\rangle_{*,Q} = 0 \qquad \text{for all } u \in U_\epsilon \text{ and for all } \epsilon > 0.$$

The optimality condition (14) can be rewritten as

$$\int_{Q_\epsilon} \bar{p}\ \Phi'_u(x,t,\bar{y},\bar{u})(u-\bar{u})\, dx\, dt \geq 0 \qquad \text{for all } u \in U_\epsilon \text{ and for all } \ \epsilon > 0.$$

Since $Q_0 = \cup_{\epsilon>0}\ Q_\epsilon$ and $U_0 = \cap_{\epsilon>0}\ U_\epsilon$, it follows that

$$\int_{Q_0} \bar{p}\ \Phi'_u(x,t,\bar{y},\bar{u})(u-\bar{u})\, dx\, dt \geq 0 \qquad \text{for all } u \in U_0.$$

The proof is complete. ∎

In the case when F does not explicitly depend on u, we have the following result.

COROLLARY 5.7. *Assume that* A1–A4 *are fulfilled. Assume in addition that* $F \equiv F(y)$. *If* $(\bar{y}, \bar{u})$ *is a solution of* $(\mathcal{P})$, *then there exist* $\bar{\mu} \in \mathcal{M}(\overline{Q})$, $\bar{\nu} \in \mathcal{M}(\overline{Q})$, *and* $\bar{p} \in L^1(0,T; W_0^{1,1}(\Omega))$ *such that the following conditions hold:*

$(\bar{\nu}, \bar{\mu})$ *satisfies* (9)–(12),

$$\begin{cases} -\dfrac{\partial \bar{p}}{\partial t} + A\bar{p} + \Phi'_y(\cdot, \bar{y}, \bar{u})\bar{p} + g'_y(\cdot, \bar{y})^* \bar{\mu}|_Q + F'_y(\bar{y})^* \bar{\nu}|_Q = 0 & \text{in } Q, \\ \bar{p}(T) + g'_y(\cdot, \bar{y})^* \bar{\mu}|_{\Omega_T} + F'_y(\bar{y})^* \bar{\nu}|_{\Omega_T} = 0 & \text{in } \Omega, \end{cases}$$

$$\int_Q \bar{p}\ \Phi'_u(x,t,\bar{y},\bar{u})(u - \bar{u})\, dx\, dt \geq 0 \qquad \text{for all } u \in U_{ad}.$$

Moreover, if $||F(\Psi)||_{C(\overline{\Sigma} \cup \Omega_0)} \neq ||F(\bar{y})||_{C(\overline{Q})}$, *then* $\bar{\nu}|_{Q \cup \Omega_T} \neq 0$.

Proof. Due to Theorem 4.2, there exist $\bar{\mu} \in \mathcal{M}(\overline{Q})$, $\bar{\zeta} \in (L^\infty(Q))'$, and $\bar{p} \in L^1(0,T; W_0^{1,1}(\Omega))$ such that (9)–(14) hold. Due to Theorem 4.1, for all $\phi \in C(\overline{Q})$, we have

$$\Big\langle \zeta, \phi \Big\rangle_{*,Q} = \Big\langle \mathcal{B}(\zeta), \phi \Big\rangle_{\mathcal{M}(\overline{Q}), C(\overline{Q})},$$

$$\Big\langle F'_y(\bar{y})^* \zeta, \phi \Big\rangle_{*,Q} = \Big\langle \zeta, F'_y(\bar{y})\phi \Big\rangle_{*,Q} = \Big\langle \mathcal{B}(\zeta), F'_y(\bar{y})\phi \Big\rangle_{\mathcal{M}(\overline{Q}), C(\overline{Q})}$$
$$= \Big\langle F'_y(\bar{y})^* \mathcal{B}(\zeta), \phi \Big\rangle_{\mathcal{M}(\overline{Q}), C(\overline{Q})}.$$

The result follows from these properties by setting $\mathcal{B}(\bar{\zeta}) = \bar{\nu}$. It remains to prove that $(\bar{\mu}|_{Q \cup \Omega_T}, \bar{\nu}|_{Q \cup \Omega_T}) \neq 0$. Argue by contradiction and suppose that $\bar{\mu}|_{Q \cup \Omega_T} = \bar{\nu}|_{Q \cup \Omega_T} = 0$. From (9), it follows that $\bar{\nu}|_{\overline{\Sigma} \cup \Omega_0} \neq 0$. Consider the solution z_F of

$$\frac{\partial z}{\partial t} + Az = 0 \quad \text{in } Q, \quad z = F(\Psi)|_\Sigma \quad \text{in } \Sigma, \quad z(0) = F(\Psi)|_{\Omega \times \{0\}} \quad \text{in } \Omega.$$

The function $z_F - F(\bar{y})$ belongs to $C_0(Q \cup \Omega_T)$, and

$$\Big\langle \bar{\nu}, z_F - F(\bar{y}) \Big\rangle_{\mathcal{M}(\overline{Q}), C(\overline{Q})} = \Big\langle \bar{\nu}, z_F - F(\bar{y}) \Big\rangle_{b, Q \cup \Omega_T} = 0. \tag{24}$$

On the other hand, since $||z_F||_{C(\overline{Q})} = ||F(\Psi)||_{C_b(\overline{\Sigma} \cup \Omega_0)} < ||F(\bar{y})||_{C(\overline{Q})}$, we can easily see that $z_F - F(\bar{y})$ belongs to $\{\chi \in C(\overline{Q}) \mid [F(\bar{y}); \chi] < 0\}$. Due to (12), it follows that

$$\Big\langle \bar{\nu}, z_F - F(\bar{y}) \Big\rangle_{\mathcal{M}(\overline{Q}), C(\overline{Q})} = \Big\langle \bar{\nu}, z_F - F(\bar{y}) \Big\rangle_{b, Q \cup \Omega_T} < 0.$$

This contradicts (24). The proof is complete. ∎

6. The compactification of Q. Let $\mathcal{O}$ be a locally compact subset of $\overline{Q}$ and let $L^\infty(\mathcal{O})$ be the space of essentially bounded measurable functions on $\mathcal{O}$. Denote by $(L^\infty(\mathcal{O}))'$ the dual space of $L^\infty(\mathcal{O})$.

THEOREM 6.1 (see [7, Theorem 11, p. 445]). *There exist a compact Hausdorff space* $\mathcal{O}^\#$ *and an isometric homomorphism* τ *from* $L^\infty(\mathcal{O})$ *onto* $C(\mathcal{O}^\#)$. *The isomor-*

phism τ maps nonnegative functions into nonnegative functions. Moreover, τ is an algebraic isomorphism in the sense that if $h = h_1 h_2$ almost everywhere on $\mathcal{O}$, then $\tau(h) = \tau(h_1)\tau(h_2)$. If f is an arbitrary real continuous function and h is in $L^\infty(\mathcal{O})$, then $\tau(f(h)) = f(\tau(h))$.

REMARK 6.2 (see [7, Theorem 26, p. 278]). Since τ is an algebraic homomorphism from $C(\overline{\mathcal{O}})$ into $C(\mathcal{O}^\#)$, there exists a continuous mapping i, from $\mathcal{O}^\#$ into $\overline{\mathcal{O}}$, such that

$$\tau(\phi) = \phi \circ i \qquad \text{for all } \phi \in C(\overline{\mathcal{O}}).$$

Moreover, observe that τ^{-1} (the inverse of τ) satisfies

$$\tau^{-1}(h\ell) = \tau^{-1}(h)\ \tau^{-1}(\ell) \qquad \text{for all } h \in C(\mathcal{O}^\#) \text{ and for all } \ell \in C(\mathcal{O}^\#).$$

The set $C(\overline{\mathcal{O}}) \otimes C(\mathcal{O}^\#)$, of linear combinations of functions of the form $\phi\psi$, where $\phi \in C(\overline{\mathcal{O}})$ and $\psi \in C(\mathcal{O}^\#)$, is a subspace of $C(\overline{\mathcal{O}} \times \mathcal{O}^\#)$. The following result gives interesting properties for the elements of $\mathcal{M}(\overline{\mathcal{O}} \times \mathcal{O}^\#)$ (the dual space of $C(\overline{\mathcal{O}} \times \mathcal{O}^\#)$).

LEMMA 6.3. *Let η be a Radon measure on $\overline{\mathcal{O}} \times \mathcal{O}^\#$, let $\mathcal{B}(\eta) \in \mathcal{M}(\overline{\mathcal{O}})$ be the projection of η on $\overline{\mathcal{O}}$, and let $\mathcal{B}(|\eta|) \in \mathcal{M}^+(\overline{\mathcal{O}})$ be the projection of $|\eta|$ on $\overline{\mathcal{O}}$. There exists a bounded linear tranformation $\Lambda_\eta : C(\mathcal{O}^\#) \longrightarrow L^\infty_{\mathcal{B}(|\eta|)}(\overline{\mathcal{O}})$ such that*

$$\left\langle \eta, \phi\psi \right\rangle_\# = \int_{\overline{\mathcal{O}}} \Lambda_\eta(\psi)\phi \, d\mathcal{B}(|\eta|) \quad \textit{for all } (\phi, \psi) \in C(\overline{\mathcal{O}}) \times C(\mathcal{O}^\#). \tag{25}$$

($\langle \cdot, \cdot \rangle_\#$ is the duality pairing between $\mathcal{M}(\overline{\mathcal{O}} \times \mathcal{O}^\#)$ and $C(\overline{\mathcal{O}} \times \mathcal{O}^\#)$, τ is defined in Theorem 6.1.)

Proof. The proof is word for word the same as the proof of Lemma 4.4 in [1]. □

Due to Theorem 6.1, the measure $\zeta \in (L^\infty(\mathcal{O}))'$ can be identified with $\hat{\zeta} \in \mathcal{M}(\mathcal{O}^\#)$ via the formula

$$\left\langle \hat{\zeta}, \xi \right\rangle_{\mathcal{M}(\mathcal{O}^\#), C(\mathcal{O}^\#)} = \left\langle \zeta, \tau^{-1}(\xi) \right\rangle_{(L^\infty(\mathcal{O}))', L^\infty(\mathcal{O})} \qquad \text{for all } \xi \in C(\mathcal{O}^\#). \tag{26}$$

Let i be the continuous mapping defined in Remark 6.2 and let e be the continuous mapping, from $\mathcal{O}^\#$ into $\overline{\mathcal{O}} \times \mathcal{O}^\#$, defined by

$$e(q^\#) = (i(q^\#), q^\#) \qquad \text{for all } q^\# \in \mathcal{O}^\#.$$

To each $\hat{\zeta} \in \mathcal{M}(\mathcal{O}^\#)$ (defined by (26)), we associate $\hat{\hat{\zeta}} \in \mathcal{M}(\overline{\mathcal{O}} \times \mathcal{O}^\#)$ defined by

$$\left\langle \hat{\hat{\zeta}}, \psi \right\rangle_\# = \left\langle \hat{\zeta}, \psi \circ e \right\rangle_{\mathcal{M}(\mathcal{O}^\#), C(\mathcal{O}^\#)} = \left\langle \zeta, \tau^{-1}(\psi \circ e) \right\rangle_{(L^\infty(\mathcal{O}))', L^\infty(\mathcal{O})} \tag{27}$$

for all $\psi \in C(\overline{\mathcal{O}} \times \mathcal{O}^\#)$.

THEOREM 6.4. *Let $\zeta \in (L^\infty(\mathcal{O}))'$ and let $\hat{\hat{\zeta}} \in \mathcal{M}(\overline{\mathcal{O}} \times \mathcal{O}^\#)$ be the measure associated with ζ. There exists a bounded linear transformation $\Lambda_{\hat{\hat{\zeta}}} : C(\mathcal{O}^\#) \longrightarrow L^\infty_{\mathcal{B}(|\hat{\hat{\zeta}}|)}(\overline{\mathcal{O}})$ such that*

$$\left\langle \hat{\hat{\zeta}}, \phi\tau(h) \right\rangle_\# = \left\langle \mathcal{B}(\hat{\hat{\zeta}}), \phi h \right\rangle_{b,\mathcal{A}} + \int_{\overline{\mathcal{O}} \setminus \mathcal{A}} \phi \Lambda_{\hat{\hat{\zeta}}}(\tau(h)) \, d\mathcal{B}(|\hat{\hat{\zeta}}|) \tag{28}$$

for all $(\phi, h) \in C(\overline{\mathcal{O}}) \times (C_b(\mathcal{A}) \cap L^\infty(\mathcal{O}))$ and for all open subsets $\mathcal{A}$ of $\overline{\mathcal{O}}$.

Proof. See the proof of Theorem 4.7 in [1]. □

THEOREM 6.5. *Let ζ be in $(L^\infty(\mathcal{O}))'$. For $\ell \in L^\infty(\mathcal{O})$, define the measure $\ell^*\zeta \in (L^\infty(\mathcal{O}))'$ by*

$$\Big\langle \ell^*\zeta, h\Big\rangle_{(L^\infty(\mathcal{O}))',L^\infty(\mathcal{O})} = \Big\langle \zeta, \ell h\Big\rangle_{(L^\infty(\mathcal{O}))',L^\infty(\mathcal{O})} \qquad \textit{for all } h \in L^\infty(\mathcal{O}).$$

Then, the measures $\hat{\hat{\zeta}}$ and $\widehat{\widehat{\ell^*\zeta}}$, respectively, associated with ζ and $\ell^*\zeta$ via (27)*, satisfy*

$$\Big\langle \widehat{\widehat{\ell^*\zeta}}, \psi\Big\rangle_{\#} = \Big\langle \tau(\ell)^*\hat{\hat{\zeta}}, \psi\Big\rangle_{\#} \qquad \textit{for all } \psi \in C(\overline{\mathcal{O}} \times \mathcal{O}^{\#}). \tag{29}$$

Proof. With (27), we can identify the measure $\ell^*\mu$ with $\widehat{\widehat{\ell^*\mu}} \in \mathcal{M}(\overline{\mathcal{O}} \times \mathcal{O}^{\#})$. Moreover,

$$\Big\langle \widehat{\widehat{\ell^*\zeta}}, \psi\Big\rangle_{\#} = \Big\langle \ell^*\zeta, \tau^{-1}(\psi \circ e)\Big\rangle_{(L^\infty(\mathcal{O}))',L^\infty(\mathcal{O})} = \Big\langle \zeta, \ell\tau^{-1}(\psi \circ e)\Big\rangle_{(L^\infty(\mathcal{O}))',L^\infty(\mathcal{O})}$$

$$= \Big\langle \zeta, \tau^{-1}\tau(\ell)\ \tau^{-1}(\psi \circ e)\Big\rangle_{(L^\infty(\mathcal{O}))',L^\infty(\mathcal{O})} = \Big\langle \zeta, \tau^{-1}\Big(\tau(\ell)(\psi \circ e)\Big)\Big\rangle_{(L^\infty(\mathcal{O}))',L^\infty(\mathcal{O})}$$

$$= \Big\langle \zeta, \tau^{-1}\Big((\tau(\ell)\psi) \circ e\Big)\Big\rangle_{(L^\infty(\mathcal{O}))',L^\infty(\mathcal{O})} = \Big\langle \hat{\hat{\zeta}}, \tau(\ell)\psi\Big\rangle_{\#} = \Big\langle \tau(\ell)^*\hat{\hat{\zeta}}, \psi\Big\rangle_{\#}$$

for all $\psi \in C(\overline{\mathcal{O}} \times \mathcal{O}^{\#})$. □

COROLLARY 6.6. *Let $\ell \in L^\infty(\mathcal{O})$, $\zeta \in (L^\infty(\mathcal{O}))'$ and let $\hat{\hat{\zeta}} \in \mathcal{M}(\overline{\mathcal{O}} \times \mathcal{O}^{\#})$ be the measure associated with ζ. Then,*

$$\Big\langle \zeta, \phi h\Big\rangle_{(L^\infty(\mathcal{O}))',L^\infty(\mathcal{O})} = \Big\langle \hat{\hat{\zeta}}, \phi\tau(h)\Big\rangle_{\#} \qquad \textit{for all } (\phi, h) \in C(\overline{\mathcal{O}}) \times L^\infty(\mathcal{O}), \tag{30}$$

$$\Big\langle \ell^*\zeta, \phi h\Big\rangle_{(L^\infty(\mathcal{O}))',L^\infty(\mathcal{O})} = \Big\langle \tau(\ell)^*\hat{\hat{\zeta}}, \phi\ \tau(h)\Big\rangle_{\#} \quad \textit{for all } (\phi, h) \in C(\overline{\mathcal{O}}) \times L^\infty(\mathcal{O}).$$

In particular, we have

$$\Big\langle \ell^*\zeta, \phi\Big\rangle_{(L^\infty(\mathcal{O}))',L^\infty(\mathcal{O})} = \Big\langle \tau(\ell)^*\hat{\hat{\zeta}}, \phi\Big\rangle_{\#} = \Big\langle \mathcal{B}\Big(\tau(\ell)^*\hat{\hat{\zeta}}\Big), \phi\Big\rangle_{\mathcal{M}(\overline{\mathcal{O}}),C(\overline{\mathcal{O}})}$$

for all $\phi \in C(\overline{\mathcal{O}})$.

Proof. To prove (30), it is sufficient to observe that

$$\Big\langle \zeta, \phi h\Big\rangle_{(L^\infty(\mathcal{O}))',L^\infty(\mathcal{O})} = \Big\langle \zeta, \tau^{-1}\tau(\phi h)\Big\rangle_{(L^\infty(\mathcal{O}))',L^\infty(\mathcal{O})} = \Big\langle \hat{\zeta}, \tau(\phi h)\Big\rangle_{\mathcal{M}(\mathcal{O}^{\#}),C(\mathcal{O}^{\#})}$$

$$= \Big\langle \hat{\zeta}, \tau(\phi)\tau(h)\Big\rangle_{\mathcal{M}(\mathcal{O}^{\#}),C(\mathcal{O}^{\#})} = \Big\langle \hat{\zeta}, (\phi \circ i)\ \tau(h)\Big\rangle_{\mathcal{M}(\mathcal{O}^{\#}),C(\mathcal{O}^{\#})}$$

$$= \Big\langle \hat{\zeta}, (\phi\ \tau(h)) \circ e\Big\rangle_{\mathcal{M}(\mathcal{O}^{\#}),C(\mathcal{O}^{\#})} = \Big\langle \hat{\hat{\zeta}}, \phi\tau(h)\Big\rangle_{\#} \qquad \text{for all } (\phi, h) \in C(\overline{\mathcal{O}}) \times L^\infty(\mathcal{O}).$$

By taking (30) and (29) into account, it follows that

$$\Big\langle \ell^*\zeta, \phi h\Big\rangle_{(L^\infty(\mathcal{O}))',L^\infty(\mathcal{O})} = \Big\langle \widehat{\widehat{\ell^*\zeta}}, \phi\ \tau(h)\Big\rangle_{\#} = \Big\langle \tau(\ell)^*\hat{\hat{\zeta}}, \phi\ \tau(h)\Big\rangle_{\#}.$$

The proof is complete. □

PROPOSITION 6.7. *Let $\ell \in (L^\infty(\mathcal{O}) \cap C_b(\mathcal{A}))$, where $\mathcal{A}$ is an open subset of $\overline{\mathcal{O}}$. Let $\zeta \in (L^\infty(\mathcal{O}))'$ and let $\hat{\hat{\zeta}} \in \mathcal{M}(\overline{\mathcal{O}} \times \mathcal{O}^\#)$ be the measure associated with ζ. Then,*

$$\Big\langle \zeta, \phi h\Big\rangle_{(L^\infty(\mathcal{O}))',L^\infty(\mathcal{O})} = \Big\langle \tau(\ell)^*\hat{\hat{\zeta}}, \phi\Big\rangle_{\#} = \Big\langle \mathcal{B}(\hat{\hat{\zeta}}), \ell\phi\Big\rangle_{b,\mathcal{A}} + \Big\langle \mathcal{B}(\tau(\ell)^*\hat{\hat{\zeta}}), \phi\Big\rangle_{b,\overline{\mathcal{O}}\setminus\mathcal{A}}$$

for all $\phi \in C(\overline{\mathcal{O}})$.

Proof. From Corollary 6.6 and from (28), for all $\phi \in C(\overline{\mathcal{O}})$, we have

$$\Big\langle \zeta, \phi h\Big\rangle_{(L^\infty(\mathcal{O}))',L^\infty(\mathcal{O})} = \Big\langle \tau(\ell)^*\hat{\hat{\mu}}, \phi\Big\rangle_{\#} = \Big\langle \mathcal{B}\Big(\tau(\ell)^*\hat{\hat{\zeta}}\Big), \phi\Big\rangle_{\mathcal{M}(\overline{\mathcal{O}}),C(\overline{\mathcal{O}})} \tag{31}$$

$$= \Big\langle \mathcal{B}\Big(\tau(\ell)^*\hat{\hat{\zeta}}\Big), \phi\Big\rangle_{b,\mathcal{A}} + \Big\langle \mathcal{B}\Big(\tau(\ell)^*\hat{\hat{\zeta}}\Big), \phi\Big\rangle_{b,\overline{\mathcal{O}}\setminus\mathcal{A}}.$$

Let $(f_k)_k$ be a sequence of continuous functions with compact support in $\mathcal{A}$, with values in $[0,1]$, and converging to 1 on every compact subset included in $\mathcal{A}$. For $\phi \in C(\overline{\mathcal{O}})$, the integrals $\int_{\mathcal{A}} \phi\ d[\mathcal{B}(\tau(\ell)^*\hat{\hat{\zeta}})]$ and $\int_{\mathcal{A}} \ell\phi\ d\mathcal{B}(\hat{\hat{\zeta}})$ are obtained by passing to the limit in $\int_{\mathcal{A}} \phi f_k\ d[\mathcal{B}(\tau(\ell)^*\hat{\hat{\mu}})]$ and in $\int_{\mathcal{A}} \ell\phi f_k\ d\mathcal{B}(\hat{\hat{\zeta}})$. Let us still denote by f_k the extension of f_k by zero to $\overline{\mathcal{O}} \setminus \mathcal{A}$. Since the functions ϕf_k and $\ell\phi f_k$ belong to $C(\overline{\mathcal{O}})$, we have

$$\int_{\mathcal{A}} \phi f_k\, d\Big[\mathcal{B}\Big(\tau(\ell)^*\hat{\hat{\zeta}}\Big)\Big] = \Big\langle \mathcal{B}\Big(\tau(\ell)^*\hat{\hat{\zeta}}\Big), \phi f_k\Big\rangle_{\mathcal{M}(\overline{\mathcal{O}}),C(\overline{\mathcal{O}})} = \Big\langle \hat{\hat{\zeta}}, \tau(\ell)\ \phi f_k\Big\rangle_{\#}$$

$$= \Big\langle \zeta, \ell\phi f_k\Big\rangle_{(L^\infty(\mathcal{O}))',L^\infty(\mathcal{O})} = \Big\langle \mathcal{B}\Big(\hat{\hat{\zeta}}\Big), \ell\phi f_k\Big\rangle_{\mathcal{M}(\overline{\mathcal{O}}),C(\overline{\mathcal{O}})} = \int_{\mathcal{A}} \ell\phi f_k\, d\,\mathcal{B}\Big(\hat{\hat{\zeta}}\Big).$$

It follows that

$$\int_{\mathcal{A}} \phi\, d\Big[\mathcal{B}\Big(\tau(\ell)^*\hat{\hat{\zeta}}\Big)\Big] = \int_{\mathcal{A}} \ell\phi\, d\,\mathcal{B}\Big(\hat{\hat{\zeta}}\Big). \tag{32}$$

The conclusion follows from (31) and (32). □

Proof of Theorem 4.1. Let us identify the measure $\zeta \in (L^\infty(Q))'$ with $\hat{\hat{\zeta}} \in \mathcal{M}(\overline{Q} \times Q^\#)$. By setting $\Lambda_\zeta = \Lambda_{\hat{\hat{\zeta}}} \circ \tau$, the properties (6) and (7) follow from (25) and (28). Assertion (8) follows from Proposition 6.7. □

7. Proof of necessary optimality conditions. To obtain optimality conditions, we prove some differential calculus rules stated below.

LEMMA 7.1. *For every $0 < \rho < 1$ and every $u_1,\ u_2 \in L^\infty(Q)$, we set $u_\rho = u_1 + \rho u_2$. If y_ρ and y_1 are the solutions of* (1) *corresponding to u_ρ and u_1, then*

$$y_\rho = y_1 + \rho z + r_\rho \qquad \textit{with} \quad \lim_{\rho\to 0} \frac{1}{\rho} ||r_\rho||_{C(\overline{Q})} = 0, \tag{33}$$

$$F(y_\rho, u_\rho) = F(y_1, u_1) + \rho(F'_y(y_1, u_1)z + F'_u(y_1, u_1)u_2) + \tilde{r}_\rho, \quad \lim_{\rho\to 0} \frac{1}{\rho} ||\tilde{r}_\rho||_{\infty,Q} = 0, \tag{34}$$

where z is the weak solution of

$$\frac{\partial z}{\partial t} + Az + \Phi'_y(\cdot, y_1, u_1)z = -\Phi'_u(\cdot, y_1, u_1)u_2 \quad in\ Q,\ z = 0\ on\ \Sigma,\ z(0) = 0\ in\ \Omega.$$

Proof. The function $\zeta_\rho = \frac{y_\rho - y_1}{\rho} - z$ is the weak solution in $C(\overline{Q})$ of

$$\frac{\partial \zeta}{\partial t} + A\zeta + a_\rho \zeta = (a - a_\rho)z + (b - b_\rho)u_2 \ \text{ in } Q, \quad \zeta = 0 \quad \text{on } \Sigma, \quad \zeta(0) = 0 \quad \text{in } \Omega,$$

where $a_\rho = \int_0^1 \Phi'_y(\cdot, y_1 + \theta(y_\rho - y_1), u_\rho)d\theta$, $b_\rho = \int_0^1 \Phi'_u(\cdot, y_1, u_1 + \theta(u_\rho - u_1))d\theta$, $a = \Phi'_y(\cdot, y_1, u_1)$, and $b = \Phi'_u(\cdot, y_1, u_1)$. Moreover, by Proposition 3.1, we have

$$||\zeta_\rho||_{C(\bar{Q})} \le C(||a - a_\rho||_{q,Q} + ||b - b_\rho||_{q,Q}) \quad \text{for all } q > \frac{N}{2} + 1,$$

where $C \equiv C(T, \Omega, N, q)$ is a constant independent of ρ. Since $(u_\rho)_\rho$ converges to u_1 in $L^\infty(Q)$, the sequence $(y_\rho)_\rho$ converges to y_1 in $C(\overline{Q})$ and $(a_\rho, b_\rho)_\rho$ converges to (a, b) in $L^\theta(Q) \times L^\theta(Q)$ for all $\theta < \infty$. We have established (33). To prove (34), observe that

$$\frac{F(y_\rho, u_\rho) - F(y_1, u_1)}{\rho}$$

$$= \frac{F(y_\rho, u_\rho) - F(y_1 + \rho z, u_\rho)}{\rho} + \frac{F(y_1 + \rho z, u_\rho) - F(y_1, u_\rho)}{\rho} + \frac{F(y_1, u_\rho) - F(y_1, u_1)}{\rho}.$$

Due to (33) and A2, we have

$$\lim_{\rho \searrow 0} \Big|\Big| \frac{F(y_\rho, u_\rho) - F(y_1 + \rho z, u_\rho)}{\rho} \Big|\Big|_{\infty,Q} = \lim_{\rho \searrow 0} \Big(||\tilde{F}_\rho||_{\infty,Q} \ \frac{||r_\rho||_{\infty,Q}}{\rho} \Big) = 0,$$

where $\tilde{F}_\rho = \int_0^1 F'_y(\theta y_\rho + (1 - \theta)(y_1 + \rho z), u_\rho)d\theta$. Therefore,

$$\lim_{\rho \searrow 0} \Big|\Big| \frac{F(y_\rho, u_\rho) - F(y_1, u_1)}{\rho} - (F'_y(y_1, u_1)z + F'_u(y_1, u_1)u_2) \Big|\Big|_{\infty,Q}$$

$$\le \lim_{\rho \searrow 0} \Big|\Big| \frac{F(y_1 + \rho z, u_\rho) - F(y_1, u_\rho)}{\rho} - F'_y(y_1, u_1)z \Big|\Big|_{\infty,Q}$$

$$+ \lim_{\rho \searrow 0} \Big|\Big| \frac{F(y_1, u_\rho) - F(y_1, u_1)}{\rho} - F'_u(y_1, u_1)u_2 \Big|\Big|_{\infty,Q} = 0.$$

The proof is complete. □

For $u \in U_{ad}$, let us denote by z_u the solution of

$$\frac{\partial z}{\partial t} + Az + \Phi'_y(\cdot, \bar{y}, \bar{u})z = -\Phi'_u(\cdot, \bar{y}, \bar{u})\ u \ \text{ in } Q, \quad z = 0 \ \text{ on } \Sigma, \ \ z(0) = 0 \ \text{ in } \Omega.$$

LEMMA 7.2. *Assume that* A1–A4 *are fulfilled. Let $(\bar{y}, \bar{u})$ be a solution of $(\mathcal{P})$. Let us set*

$$\mathcal{S} = \Big\{ (\xi, \chi) \in C(\overline{Q}) \times L^\infty(Q) \mid \textit{there exists } u \in U_{ad} \textit{ such that}$$

$$\xi = g(\cdot, \bar{y}) + g'_y(\cdot, \bar{y})(z_u - z_{\bar{u}}) \quad \chi = F'_y(\bar{y}, \bar{u})\ (z_u - z_{\bar{u}}) + F'_u(\bar{y}, \bar{u})(u - \bar{u}) \Big\},$$

$$\mathcal{D} = \text{ int } \mathcal{C} \times \{\chi \in L^\infty(Q) \mid [F(\bar y, \bar u); \chi] < 0\}.$$

Then there exist $\bar\mu \in \mathcal{M}(\overline{Q})$ *and* $\bar\zeta \in (L^\infty(Q))'$ *such that*

$$\langle \bar\mu, \xi_1\rangle_{\mathcal{M}(\overline{Q}),C(\overline{Q})} + \langle \bar\zeta, \chi_1\rangle_{*,Q} > \langle \bar\mu, \xi_2\rangle_{\mathcal{M}(\overline{Q}),C(\overline{Q})} + \langle \bar\zeta, \chi_2\rangle_{*,Q} \tag{35}$$

for all $(\xi_1, \chi_1) \in \mathcal{S}$ *and for all* $(\xi_2, \chi_2) \in \mathcal{D}$ *and*

$$\langle \bar\mu, \xi_1\rangle_{\mathcal{M}(\overline{Q}),C(\overline{Q})} + \langle \bar\zeta, \chi_1\rangle_{*,Q} \geq \langle \bar\mu, \xi_2\rangle_{\mathcal{M}(\overline{Q}),C(\overline{Q})} + \langle \bar\zeta, \chi_2\rangle_{*,Q} \tag{36}$$

for all $(\xi_1, \chi_1) \in \mathcal{S}$ *and for all* $(\xi_2, \chi_2) \in \overline{\mathcal{D}} = \mathcal{C} \times \mathbf{cl}_\infty\{\chi \in L^\infty(Q) \mid [F(\bar y, \bar u); \chi] < 0\}$.

Proof. The sets $\mathcal{S}$ and $\mathcal{D}$ are convex, and $\mathcal{D}$ is open. Let us prove that $\mathcal{S} \cap \mathcal{D} = \emptyset$. Argue by contradiction and suppose that there exists $u_o \in U_{ad}$ such that

$$g(\cdot, \bar y) + g'_y(\cdot, \bar y)(z_{u_o} - z_{\bar u}) \in \text{ int } \mathcal{C}, \tag{37}$$

$$\Big[F(\bar y, \bar u); F'_y(\bar y, \bar u)\ (z_{u_o} - z_{\bar u}) + F'_u(\bar y, \bar u)(u_o - \bar u)\Big] < 0. \tag{38}$$

Let $u_\rho = \bar u + \rho(u_o - \bar u)$, $g_\rho = g(\cdot, \bar y) + \frac{1}{\rho}(g(y_\rho) - g(\bar y))$, where y_ρ is the solution of (1) corresponding to u_ρ. From (37), (38), and Lemma 7.1, it follows that

$$\lim_{\rho \searrow 0} g_\rho \in \text{ int } \mathcal{C} \quad \text{and} \quad \lim_{\rho \searrow 0} \frac{||F(y_\rho, u_\rho)||_{\infty,Q} - ||F(\bar y, \bar u)||_{\infty,Q}}{\rho} < 0.$$

Therefore, there exists $\rho_o > 0$ such that, for every $0 < \rho \leq \rho_o < 1$, we have

$$g(\cdot, y_\rho) = \rho\ g_\rho + (1 - \rho)\ g(\bar y) \in \text{ int } \mathcal{C} \quad \text{and} \quad J(y_\rho, v_\rho) < J(\bar y, \bar v).$$

This contradicts the optimality of $(\bar y, \bar u)$ and proves that $\mathcal{S} \cap \mathcal{D} = \emptyset$. From a geometric version of the Hahn–Banach theorem (the Eidelheit theorem [12]), there exists $(\bar\mu, \bar\zeta) \in \mathcal{M}(\overline{Q}) \times (L^\infty(Q))'$ such that

$$\inf_{(\xi_1,\chi_1)\in\mathcal{S}} \Big(\langle \bar\mu, \xi_1\rangle_{\mathcal{M}(\overline{Q}),C(\overline{Q})} + \langle \bar\zeta, \chi_1\rangle_{*,Q}\Big)$$

$$> \ \langle \bar\mu, \xi_2\rangle_{\mathcal{M}(\overline{Q}),C(\overline{Q})} + \langle \bar\zeta, \chi_2\rangle_{*,Q} \quad \text{for all } (\xi_2, \chi_2) \in \mathcal{D}$$

and

$$\inf_{(\xi_1,\chi_1)\in\mathcal{S}} \Big(\langle \bar\mu, \xi_1\rangle_{\mathcal{M}(\overline{Q}),C(\overline{Q})} + \langle \bar\zeta, \chi_1\rangle_{*,Q}\Big)$$

$$\geq \sup_{(\xi_2,\chi_2)\in\overline{\mathcal{D}}} \Big(\langle \bar\mu, \xi_2\rangle_{\mathcal{M}(\overline{Q}),C(\overline{Q})} + \langle \bar\zeta, \chi_2\rangle_{*,Q}\Big).$$

The proof is complete. □

7.1. Proof of Theorem 4.2. Let z_{θ_o} be a function belonging to int $\mathcal{C}$ and satisfying $z_{\theta_o} \equiv g(\cdot, \bar\Psi)$ in $\overline{\Sigma} \cup \overline{\Omega}_0$. (The existence of such a function follows from A4.)

• If we set $\chi_1 = 0$, $\xi_1 = g(\cdot, \bar y)$, and $\xi_2 = z_{\theta_o}$ in (35), we obtain

$$\Big\langle \bar\mu, g(\cdot, \bar y) - z_{\theta_o}\Big\rangle_{\mathcal{M}(\overline{Q}),C(\overline{Q})} = \Big\langle \bar\mu, g(\cdot, \bar y) - z_{\theta_o}\Big\rangle_{b,Q\cup\Omega_T} > \Big\langle \bar\zeta, \chi\Big\rangle_{*,Q}$$

for all $\chi \in \{y \in L^\infty(Q) \mid [F(\bar y, \bar u); y] < 0\}$. Conditions (9) and (12) easily follow from this inequality.

• The following is a direct consequence of Remark 4.3:

(i) If we set $\xi_1 = g(\cdot, \bar{y})$, $\xi_2 = z \in \mathcal{C}$ fixed, and $\chi_1 = \chi_2 = 0$ in (36), we obtain (10).

(ii) If we set $\xi_1 = \xi_2 = g(\cdot, \bar{y})$ and $\chi_1 = 0$ in (36), we obtain (11).

(iii) Let $u \in U_{ad}$. By setting $\xi_1 = g(\cdot, \bar{y}) + g'_y(\cdot, \bar{y})(z_u - z_{\bar{u}})$, $\chi_1 = F'_y(\bar{y}, \bar{u})\ (z_u - z_{\bar{u}}) + F'_u(\bar{y}, \bar{u})(u - \bar{u})$, $\chi_2 = 0$, and $\xi_2 = g(\cdot, \bar{y})$ in (36), we obtain

$$\left\langle \bar{\mu}, g'_y(\cdot, \bar{y})(z_u - z_{\bar{u}}) \right\rangle_{\mathcal{M}(\overline{Q}), C(\overline{Q})} + \left\langle \bar{\zeta}, F'_y(\bar{y}, \bar{u})(z_u - z_{\bar{u}}) + F'_u(\bar{y}, \bar{u})(u - \bar{u}) \right\rangle_{*,Q} \geq 0. \tag{39}$$

Let $\bar{p}$ be the weak solution of (13). With the Green formula of Theorem 3.4 and with Theorem 4.1, we obtain

$$\left\langle \bar{\mu}, g'_y(\cdot, \bar{y})(z_u - z_{\bar{u}}) \right\rangle_{\mathcal{M}(\overline{Q}), C(\overline{Q})} + \left\langle \bar{\zeta}, F'_y(\bar{y}, \bar{u})(z_u - z_{\bar{u}}) \right\rangle_{*,Q}$$

$$= \left\langle g'_y(\cdot, \bar{y})^* \bar{\mu} + \mathcal{B}(F'_y(\bar{y}, \bar{u})^* \bar{\zeta}), z_u - z_{\bar{u}} \right\rangle_{b, Q \cup \Omega_T} = \int_Q \bar{p}\ \Phi'_u(x, t, \bar{y}, \bar{u})(u - \bar{u})\, dx\, dt.$$

This equality, with (39), gives (14). The proof is complete. □

REFERENCES

[1] N. Arada and J.-P. Raymond, *Necessary optimality conditions for control problems and the Stone–Čech compactification*, SIAM J. Control Optim., 37 (1999), pp. 1011–1032.

[2] N. Arada and J.-P. Raymond, *Dirichlet Boundary Control of Semilinear Parabolic Equations. Part 1: Problems with No State Constraints*, Report 98-05, UMR-CNRS 5640, Toulouse, France.

[3] N. Arada and J.-P. Raymond, *Optimality conditions for state-constrained Dirichlet boundary control problems*, J. Optim. Theory Appl., 102 (1999), pp. 51–68.

[4] V. Barbu, *Analysis and Control of Nonlinear Infinite-Dimentional Systems*, Academic Press, Boston, MA, 1993.

[5] E. N. Barron, *The Pontryagin maximum principle for minimax problems of optimal control*, Nonlinear Anal., 15 (1990), pp. 1155–1165.

[6] E. Casas, *Boundary control of semilinear elliptic equations with pointwise state constraints*, SIAM J. Control Optim., 31 (1993), pp. 993–1006.

[7] N. Dunford and J. T. Schwartz, *Linear Operators, Part* 1, Interscience Publishers, New York, London, 1958.

[8] H. O. Fattorini, *Infinite-Dimensional Optimization and Control Theory*, Encyclopedia Math. Appl., Cambridge University Press, Cambridge, UK, 1999.

[9] X. J. Li and J. M. Yong, *Optimal Control Theory for Infinite-Dimensional Systems*, Birkhäuser, Boston, MA, 1995.

[10] B. S. Mordukhovich and K. Zhang, *Dirichlet boundary control of parabolic systems with pointwise state constraints*, in International Conference on Control and Estimations of Distributed Parameter Systems, Vorau, W. Desch, F. Kappel, and K. Kunish, eds., Birkhäuser-Verlag, Basel, 1998, pp. 231–246.

[11] B. S. Mordukhovich and K. Zhang, *Minimax control of parabolic systems with Dirichlet boundary condition and state constraints*, Appl. Math. Optim., 36 (1997), pp. 323–360.

[12] T. Roubiček, *Relaxation in Optimization Theory and Variational Calculus*, de Gruyter Ser. Nonlinear Anal. Appl., Walter de Gruyter, Berlin, 1997.

[13] J. Yong, *A minimax control problem for second order elliptic partial differential equations*, Kodai Math. J., 16 (1993), pp. 469–486.

[14] W. Yu, *Identification for parabolic distributed parameter systems with constraints on the parameters and the state*, SIAM J. Control Optim., 33 (1995), pp. 1801–1815.

SIAM J. CONTROL OPTIM.
Vol. 38, No. 1, pp. 272–293

INEXACT SQP INTERIOR POINT METHODS AND LARGE SCALE OPTIMAL CONTROL PROBLEMS*

F. LEIBFRITZ† AND E. W. SACHS‡

Abstract. Optimal control problems with partial differential equations lead to large scale nonlinear optimization problems with constraints. An efficient solver which takes into account the structure and also the size of the problem is an inexact sequential quadratic programming method where the quadratic problems are solved iteratively. Based on a reformulation as a mixed nonlinear complementarity problem we give a measure of when to terminate the iterative quadratic program solver. For the latter we use an interior point algorithm. Under standard assumptions, local linear, superlinear, and quadratic convergence can be proved. The numerical application is an optimal control problem from nonlinear heat conduction.

Key words. inexact SQP method, interior point method, optimal control, partial differential equations

AMS subject classifications. 35K55, 65M99, 49N99, 90C06, 90C30, 93A15, 65K05

PII. S0363012996298795

1. Introduction.

In this paper we consider large scale optimization problems with equality and inequality constraints which occur, for instance, from the discretization of optimal control problems with partial differential equations and state constraints.

As an example we use a problem which arises in the control of heating processes of industrial kilns. The goal is to heat the furnace such that the interior of the metal ingots or ceramic charges inside the kiln follows a prescribed temperature profile. Furthermore, constraints on the temperature, so-called state constraints, are imposed during the heating process. This problem can be formulated as an optimal control problem with a nonlinear diffusion equation which is controlled through the boundary. The objective function is given in a least squares framework. The state constraints are upper bounds on the temperature and are so-called hard constraints, in contrast to constraints on the controls. It is well known that control problems with state constraints require advanced numerical methods for their solution.

We present a method from *sequential quadratic programming* (SQP) for the solution of the resulting finite-dimensional optimization problem. In general, SQP methods reduce the nonlinear problem to a sequence of quadratic subproblems. For large problems, the solution of a single quadratic program might be very expensive. Even for small problems, the computational effort in SQP methods is dominated by solving the quadratic subproblems. In the view of global and local convergence properties of the SQP methods, it is advisable to solve each of the quadratic subproblems by

*Received by the editors February 12, 1996; accepted for publication (in revised form) June 17, 1999; published electronically December 21, 1999. This research was partially supported by the Bundesministerium für Bildung, Wissenschaft, Forschung und Technologie (BMBF) through the research initiative "Anwendungsorientierte Verbundvorhaben auf dem Gebiet der Mathematik."

http://www.siam.org/journals/sicon/38-1/29879.html

†Universität Trier, FB IV, "Graduiertenkolleg Mathematische Optimierung," D–54286 Trier, Germany (leibfr@uni-trier.de). The research of this author was supported in part by the "Deutsche Forschungsgemeinschaft" through the Graduiertenkolleg "Mathematische Optimierung" at the University of Trier.

‡Universität Trier, FB IV, Mathematik and "Graduiertenkolleg Mathematische Optimierung," D–54286 Trier, Germany (sachs@uni-trier.de).

an efficient algorithm only up to a certain accuracy. Furthermore, the structure of the matrices occurring in each quadratic subproblem should be taken into account. Moreover, the presence of inequality constraints poses a further difficulty for these algorithms. In our work we use an interior point algorithm for the solution of the quadratic subproblems (cf. [6], [12], [23], [24], and the references therein). Since this is an iterative method it has two major advantages which can be used in the context of large scale optimization problems. First, the sparsity of the matrices can be used in an efficient way. Second, the quadratic subproblems can be solved only approximately. In other words, the inner iterative process can be terminated during the first iterations of the outer SQP method when one is still far away from a solution at an early stage, saving computing time. Then one can adaptively tighten the accuracy of the solutions of the quadratic problems as the iteration progresses. This leads to an inexact SQP approach which offers a trade-off between the accuracy of solving each subproblem and the amount of work for solving them (cf. [3], [4], and [19]). An immediate question is just how accurately should the subproblems be solved? And is there a measure when to terminate the iterative quadratic program solver? Furthermore, does an inexact SQP approach possess the same asymptotic convergence behavior as an exact SQP method? We will show that theoretical and practical answers to the above questions are available (cf. [5], [8], [10], [14, section 7], [16], [17], [22], and the references therein).

Converting the Karush–Kuhn–Tucker conditions of the nonlinear optimization problem into a mixed nonlinear complementarity problem enables us to define a quantity which gives a practical measure of how close a given vector is at a solution of the considered problem. This quantity determines how accurately the quadratic subproblems will be solved. Moreover, we prove under standard assumptions that a linear, superlinear, or quadratic rate of the outer SQP method is retained, although the quadratic subproblems are not solved exactly. Furthermore, the interior point algorithm itself can be shown to be superlinearly or quadratically convergent. The linear systems of the interior point approach are solved by an iterative equation solver like GMRES, in order to make use of the inherent sparsity structure, also with an adaptive termination criterion. All the convergence results will be verified by numerical results for a discretized control problem, as mentioned above.

Note that the convergence results for the inexact SQP method are of a local nature and no globalization strategies are incorporated. These require second order sufficiency conditions for optimality at the solution. This implies that the corresponding quadratic subproblems to be solved by interior point methods exhibit the proper convexity requirements.

In section 2 we state the finite-dimensional optimization problem with nonlinear equality and inequality constraints. The corresponding Karush–Kuhn–Tucker conditions will be restated into a mixed nonlinear complementarity problem. After that we present the inexact SQP algorithm for determining a solution of the mixed nonlinear complementarity problem. Furthermore, we discuss the notion of a regular solution and state some sufficient conditions for the regularity of this problem class.

Essential to an inexact SQP method is a practical measure of how close a given vector is to being a solution of the mixed nonlinear complementarity problem. Therefore, in section 3 we give a quantity which can be used as a measure of inexactness. Some results will justify the use of this quantity.

The local convergence analysis for the inexact SQP algorithm can be found in section 4. Under standard assumptions we prove that this method achieves the same

asymptotic convergence rates as an exact SQP approach. A similar result can be found in Pang [19] for nonmixed complementarity problems.

In section 5 an interior point method is described which is an iterative scheme to solve the quadratic subproblems, or, equivalently, the mixed linear complementarity problems occurring in the inexact SQP iteration. Furthermore, we state some local convergence results. In particular, the first result yields a superlinear convergence rate of the Karush–Kuhn–Tucker system of the quadratic subproblem to zero. Second, under stronger assumptions we obtain locally the quadratic convergence of the generated sequence to a solution of the subproblem.

Finally, in section 6, we use a discretized parabolic state-constrained control problem to support our theoretical results. It shows that the combination of an inexact SQP method with an interior point algorithm works very well for the considered problem class.

Throughout this paper we use the following notation. If $\varphi : \mathbb{R}^n \to \mathbb{R}^m$ is a continuously differentiable function, we use the notation

$$\varphi_z(z) = [(\varphi_1)_z(z), \dots, (\varphi_m)_z(z)] \in \mathbb{R}^{n\times m}.$$

Thus, the matrix $\varphi_z(z)$ has as columns the gradients $(\varphi_1)_z(z), \dots, (\varphi_m)_z(z)$ and is the transpose of the Jacobian matrix of the function φ. For vectors $\zeta \in \mathbb{R}^n$ and $\vartheta \in \mathbb{R}^n$ we also use

$$\min(\zeta) = \min_{1\le i\le n} \zeta_i, \quad \max(\zeta) = \max_{1\le i\le n} \zeta_i, \quad \chi = \min(\zeta, \vartheta), \quad \chi = \max(\zeta, \vartheta),$$

where $\chi \in \mathbb{R}^n$ is vector valued with the "min" ("max") interpreted as componentwise minimum (maximum). The index l will be used as the inexact SQP iteration counter and the index k denotes the interior point iteration counter of the considered algorithms.

2. Inexact SQP method for nonlinear optimization problems. In this section we consider a local inexact SQP method for the solution of the following finite-dimensional minimization problem with nonlinear equality and inequality constraints:

$$\text{(NP)} \qquad \min_z \ F(z) \quad \text{s.t.} \quad f(z) = 0, \quad g(z) \le 0,$$

where $z \in \mathbb{R}^n, F : \mathbb{R}^n \to \mathbb{R}, f : \mathbb{R}^n \to \mathbb{R}^{m_e}, g : \mathbb{R}^n \to \mathbb{R}^m$. We make the following assumptions on (NP):

Assumption 2.1.

(i) There exists an optimal solution z^* of (NP).

(ii) The functions F, f, g are twice continuously differentiable in an open ball $U(z^*)$.

(iii) The linear independent constraint qualification is fulfilled, i.e., the gradients of the active constraints are linearly independent at z^*.

Using the Karush–Kuhn–Tucker theorem, Assumption 2.1 implies that there exist $w^* \in \mathbb{R}^{m_e}$ and $y^* \in \mathbb{R}^m$, $y^* \ge 0$, such that (z^*, w^*, y^*) satisfies the first order necessary optimality conditions for (NP), i.e.,

$$F_z(z^*) + f_z(z^*)w^* + g_z(z^*)y^* = 0, \tag{2.1}$$

$$f(z^*) = 0, \tag{2.2}$$

$$g(z^*) \le 0, \tag{2.3}$$

$$(y^*)^T g(z^*) = 0. \tag{2.4}$$

These conditions can be converted into a *mixed nonlinear complementarity problem* due to the fact that some variables (z and w) are not restricted to be nonnegative (cf., for example, [9], [19], [20], or [21]). The mixed nonlinear complementarity problem, defined in terms of two mappings $h_1 : \mathbb{R}^{n+m_e+m} \to \mathbb{R}^{n+m_e}$ and $h_2 : \mathbb{R}^{n+m_e+m} \to \mathbb{R}^m$, is to determine a triple of vectors $(z, w, y) \in \mathbb{R}^{n+m_e+m}$ such that

$$h_1(z,w,y) = 0, \quad h_2(z,w,y) \geq 0, \quad y \geq 0, \quad \text{and} \quad y^T h_2(z,w,y) = 0. \tag{2.5}$$

If we define the functions by

$$h_1(z,w,y) = \begin{bmatrix} F_z(z) + f_z(z)\, w + g_z(z)\, y \\ f(z) \end{bmatrix} \quad \text{and} \quad h_2(z,w,y) = -g(z), \tag{2.6}$$

then the Karush–Kuhn–Tucker conditions (2.1)–(2.4) correspond precisely to the mixed nonlinear complementarity problem (2.5). In the following, we look for a solution $(z^*, w^*, y^*) \in \mathbb{R}^{n+m_e+m}$ such that

$$\begin{aligned} y^* \geq 0, \quad & {y^*}^T g(z^*) = 0, \\ h_1(z^*, w^*, y^*) = 0, \quad & \text{and} \quad h_2(z^*, w^*, y^*) \geq 0, \end{aligned} \tag{MNCP}$$

where the mappings h_1 and h_2 are given by (2.6).

SQP methods reduce the solution of the nonlinear problem (NP) to the numerical solution of a sequence of quadratic subproblems. In particular, a SQP method for a problem with equality constraints determines the step as the solution of a subproblem with quadratic objective function and linear constraints. For $w \in \mathbb{R}^{m_e}$, $y \in \mathbb{R}^m$, define the Lagrangian function by

$$\mathcal{L}(z,w,y) = F(z) + w^T f(z) + y^T g(z). \tag{2.7}$$

A common variant of the SQP algorithm for (NP) including inequality constraints obtains a new iterate $(z^{l+1}, w^{l+1}, y^{l+1})$ from the current iterate (z^l, w^l, y^l) by solving at each stage the quadratic program

$$\begin{aligned} \min_{\Delta z} \ & \tfrac{1}{2}\, \Delta z^T\, \mathcal{L}_{zz}(z^l, w^l, y^l)\, \Delta z + F_z^T(z^l)\, \Delta z \\ \text{s.t.} \ & f_z^T(z^l)\, \Delta z + f(z^l) = 0, \\ & g_z^T(z^l)\, \Delta z + g(z^l) \leq 0. \end{aligned} \tag{QP}_l$$

Then the new iterate z^{l+1} is set to $z^l + \Delta z$ and the new Lagrange multipliers w^{l+1}, y^{l+1} are obtained from the multipliers at the solution of $(\mathrm{QP})_l$. The necessary optimality conditions for this subproblem can be restated as a *mixed linear complementarity problem*, i.e.,

$$\begin{aligned} & \text{find } (z,w,y) \text{ such that} \\ & y \geq 0, \quad -g_z^T(z^l)(z - z^l) - g(z^l) \geq 0, \\ & y^T \left(-g_z^T(z^l)(z - z^l) - g(z^l)\right) = 0, \\ & \begin{bmatrix} F_z(z^l) \\ f(z^l) \end{bmatrix} + \begin{bmatrix} \mathcal{L}_{zz}(z^l, w^l, y^l) & f_z(z^l) & g_z(z^l) \\ f_z^T(z^l) & O & O \end{bmatrix} \begin{bmatrix} z - z^l \\ w \\ y \end{bmatrix} = 0. \end{aligned} \tag{MLCP}_l$$

For $\Delta z = z^{l+1} - z^l$ the point $(\Delta z, w^{l+1}, y^{l+1})$ satisfies the Karush–Kuhn–Tucker conditions of $(\mathrm{QP})_l$. Hence, it is also a solution vector of the mixed linear complementarity problem.

Under suitable regularity conditions the sequence $\{(z^l, w^l, y^l)\}$ generated by a SQP approach is well defined and locally convergent. A drawback of this method is the cost of solving the quadratic subproblem, or, equivalently, the mixed linear complementarity problem at each iteration. Computing the exact solution can be expensive if the number of unknowns is large and may not be justified when far from a solution. Therefore, one is led to use an iterative method, e.g., an interior point algorithm, and to solve the subproblem only approximately. This observation leads to an inexact SQP method which offers a trade-off between the accuracy of solving the subproblems and the amount of work involved in solving them (cf. [3], [4], and [19]).

Pang [19] suggests an inexact Newton method for solving a nonmixed nonlinear complementarity problem of the following form. Find a vector $\xi \in \mathbb{R}^n$ such that

$$\xi \geq 0, \quad \varphi(\xi) \geq 0, \quad \xi^T \varphi(\xi) = 0.$$

In this paper we formulate an inexact SQP algorithm for determining a solution of the mixed nonlinear complementarity problem (MNCP). If we define the mixed nonlinear complementarity function $\hat{h} : \mathbb{R}^{n+m_e+m} \to \mathbb{R}^{n+m_e+m}$ (cf. [11]) by

$$\hat{h}(z, w, y) = \begin{bmatrix} \mathcal{L}_z(z, w, y) \\ f(z) \\ \min(y, -g(z)) \end{bmatrix} \tag{2.8}$$

and its linearization at a given point (z^l, w^l, y^l) near the solution (z^*, w^*, y^*) by

$$\hat{h}^l(z, w, y) = \begin{bmatrix} F_z(z^l) + f_z(z^l)w + g_z(z^l)y + \mathcal{L}_{zz}(z^l, w^l, y^l)(z - z^l) \\ f(z^l) + f_z^T(z^l)(z - z^l) \\ \min(y, -g(z^l) - g_z^T(z^l)(z - z^l)) \end{bmatrix}, \tag{2.9}$$

where the *min*-operator is interpreted as the componentwise minimum of the vector-valued entries, then we obtain the following algorithm.

Algorithm 2.2. Inexact SQP Method.

Choose a local starting point (z^0, w^0, y^0) and a termination criterion $\varepsilon > 0$.

For $l = 0, 1, 2, \ldots$ until $||\hat{h}(z^l, w^l, y^l)|| \leq \varepsilon$ do

(I) Choose $\lambda_l > 0$ and find a vector (z, w, y) of $(\text{MLCP})_l$ which satisfies

$$||\hat{h}^l(z, w, y)|| \leq \lambda_l \, ||\hat{h}(z^l, w^l, y^l)||. \tag{2.10}$$

(II) Set $(z^{l+1}, w^{l+1}, y^{l+1}) = (z, w, y)$.

The inexact SQP method of Algorithm 2.2 generates in every iteration a new point $(z^{l+1}, w^{l+1}, y^{l+1})$ according to the *approximation rule* (2.10). The nonnegative sequence $\{\lambda_l\}$ controls the level of accuracy and will be specified later.

The approximation rule (2.10) as a measure of inexactness will be motivated in the next section. Pang [19] introduced this error bound in the context of nonlinear complementarity problems. Furthermore, the rule (2.10) is similar to the residual rule used in [3] for the solution of nonlinear equations with an inexact Newton method.

In order to state the convergence of the inexact SQP method and to motivate the usefulness of (2.10) as a measure of inexactness, we discuss the notion of a *regular solution*, which was introduced by Robinson [21] in the context of generalized equations.

For our purpose, we define regularity in terms of the mixed nonlinear complementarity problem (MNCP).

DEFINITION 2.3 (cf. [21, p. 45]). *Let* (z^*, w^*, y^*) *be a solution of* (MNCP). *Then* (z^*, w^*, y^*) *is called regular if there exists a neighborhood* U_δ^* *of* (z^*, w^*, y^*) *and a scalar* $\gamma^* > 0$ *such that for every perturbation vector* $\rho = (\rho_1, \rho_2, \rho_3)^T \in \mathbb{R}^{n+m_e+m}$ *with* $||\rho|| < \gamma^*$, *there is a unique solution* $(z(\rho), w(\rho), y(\rho)) \in U_\delta^*$ *that solves the perturbed mixed linear complementarity problem defined by*

$$(\text{MLCP})_\rho \qquad \begin{array}{c} y(\rho) \geq 0, \quad y(\rho)^T \left(-g_z^T(z^*)(z(\rho) - z^*) - g(z^*) - \rho_3\right) = 0 \\ -g_z^T(z^*)(z(\rho) - z^*) - g(z^*) - \rho_3 \geq 0, \end{array}$$

$$\begin{bmatrix} F_z(z^*) \\ f(z^*) \end{bmatrix} + \begin{bmatrix} \mathcal{L}_{zz}(z^*, w^*, y^*) & f_z(z^*) & g_z(z^*) \\ f_z^T(z^*) & O & O \end{bmatrix} \begin{bmatrix} z(\rho) - z^* \\ w(\rho) \\ y(\rho) \end{bmatrix} - \begin{bmatrix} \rho_1 \\ \rho_2 \end{bmatrix} = 0.$$

Moreover, the solution $(z(\rho), w(\rho), y(\rho))$ *is Lipschitz continuous in* ρ; *i.e., there is a constant* $L > 0$ *such that for every* $\rho, \hat{\rho}$ *with* $||\rho|| < \gamma^*, ||\hat{\rho}|| < \gamma^*$ *we have*

$$\|(z(\rho), w(\rho), y(\rho)) - (z(\hat{\rho}), w(\hat{\rho}), y(\hat{\rho}))\| \leq L \, \|\rho - \hat{\rho}\|.$$

This regularity condition can be considered a generalization of the requirement that the Jacobian matrix be nonsingular in the case of solving systems of nonlinear equations. Robinson [21] states some sufficient conditions for the regularity of (MNCP). In particular, he proved under Assumption 2.4 that (z^*, w^*, y^*) is a regular solution of (MNCP). Hence, the mixed nonlinear complementarity problem is locally solvable in a vicinity of (z^*, w^*, y^*). Without loss of generality, $g(z^*)$ and y^* can be partitioned as

$$(g_+(z^*), g_0(z^*), g_-(z^*))^T \in \mathbb{R}^{r+s+t}, \qquad \left(y_+^*, y_0^*, y_-^*\right)^T \in \mathbb{R}^{r+s+t},$$

where $m = r + s + t$ and

$$\begin{array}{rcl@{\qquad}rcl} g_+(z^*) & = & 0, & y_+^* & > & 0, \\ g_0(z^*) & = & 0, & y_0^* & = & 0, \\ g_-(z^*) & < & 0, & y_-^* & = & 0. \end{array} \tag{2.11}$$

With this partition we can formulate a special strong second order sufficiency condition and a linear independence condition. We make the following assumption.

Assumption 2.4.

(i) The matrix $[f_z^T(z^*)\,(g_+)_z^T(z^*)\,(g_0)_z^T(z^*)]^T$ has full row rank.

(ii) For all $\Delta z \neq 0$, $\Delta z \in \mathbb{R}^n$ satisfying $\Delta z^T f_z(z^*) = 0$ and $\Delta z^T (g_+)_z(z^*) = 0$:

$$\Delta z^T \, \mathcal{L}_{zz}(z^*, w^*, y^*) \, \Delta z > 0.$$

With these conditions one can show the following result.

THEOREM 2.5 (see Robinson [21, Theorem 4.1]). *Let Assumption* 2.1(ii) *and Assumption* 2.4 *be fulfilled. Furthermore, let* (z^*, w^*, y^*) *be a solution of* (MNCP). *Then,* (z^*, w^*, y^*) *is a regular solution of* (MNCP).

If a solution of (MNCP) satisfies the conditions of Assumption 2.4, then Theorem 2.5 guarantees the regularity of the problem. In this case we know from Definition 2.3 that there is a unique solution of $(\text{MLCP})_\rho$ in a vicinity of (z^*, w^*, y^*). Hence, we can locally determine a solution of the mixed nonlinear complementarity problem

with the inexact SQP method of Algorithm 2.2. Furthermore, the regular solution (z^*, w^*, y^*) of (MNCP) is also an optimal solution of (NP). In particular, we can state the following result.

THEOREM 2.6 (cf. Bertsekas [1, Theorem 1.30]). *Let Assumption* 2.1(ii) *and Assumption* 2.4 *be fulfilled. Furthermore, let* (z^*, w^*, y^*) *be a solution of* (MNCP). *Then,* z^* *is a strict local minimum of the nonlinear problem* (NP).

3. Measure of inexactness. The main advantage of the approximation rule (2.10) is the savings in the computational effort during the early iterations; i.e., at each stage of the inexact SQP method of Algorithm 2.2 a new iterate is defined as an inaccurate solution of the subproblem $(\text{MLCP})_l$. The use of an inexact method raises the question of how to measure the accuracy of the approximate solution to a mixed nonlinear complementarity problem, or, in other words, how close is a given vector to a solution of (MNCP)?

For the motivation of the approximation rule (2.10) as a *measure of inexactness*, first, we consider the mixed linear complementarity problem $(\text{MLCP})_l$. Second, similar results will be given for the mixed nonlinear complementarity problem (MNCP).

We assume that $(z^l, w^l, y^l), l \geq 0$, is fixed and given. For such a point, we define the matrix $M \in \mathbb{R}^{(n+m_e+m)\times(n+m_e+m)}$ and the vectors $r_1, s_1 \in \mathbb{R}^{n+m_e}$ by

$$M = \begin{bmatrix} \nabla h_1(z^l, w^l, y^l) \\ \nabla h_2(z^l, w^l, y^l) \end{bmatrix}, \; r_1 = \begin{bmatrix} F_z(z^l) \\ f(z^l) \end{bmatrix}, \; s_1 = \begin{bmatrix} z - z^l \\ w \end{bmatrix},$$

where

$$\begin{aligned} \nabla h_1(z^l, w^l, y^l) &= \begin{bmatrix} \mathcal{L}_{zz}(z^l, w^l, y^l) & f_z(z^l) & g_z(z^l) \\ f_z^T(z^l) & O & O \end{bmatrix}, \\ \nabla h_2(z^l, w^l, y^l) &= \begin{bmatrix} -g_z^T(z^l) & O & O \end{bmatrix}, \end{aligned}$$

and rewrite $(\text{MLCP})_l$ as

$$\begin{aligned} & y \geq 0, \quad y^T \left(-g_z^T(z^l)(z - z^l) - g(z^l)\right) = 0, \\ & -g_z^T(z^l)(z - z^l) - g(z^l) \geq 0, \quad r_1 + \nabla h_1(z^l, w^l, y^l) \begin{bmatrix} s_1 \\ y \end{bmatrix} = 0. \end{aligned} \tag{3.1}$$

Obviously, a vector $(\tilde{z}, \tilde{w}, \tilde{y})$ is an exact solution of $(\text{MLCP})_l$, or, equivalently, of (3.1) if and only if the linearized complementarity function (2.9) is equal to zero.

LEMMA 3.1. *Let* (z^l, w^l, y^l) *be fixed. Furthermore, let* $\hat{h}^l$ *be defined by* (2.9). *Then*

$$\hat{h}^l(\tilde{z}, \tilde{w}, \tilde{y}) \equiv 0 \quad \Longleftrightarrow \quad (\tilde{z}, \tilde{w}, \tilde{y}) \quad \textit{is a solution of } (\text{MLCP})_l.$$

Thus, the quantity $||\hat{h}^l(z, w, y)||$ is a practical measure of how close a given vector is to being a solution of $(\text{MLCP})_l$. The following result justifies the use of the above quantity as a measure of inexactness, which can be proved by a simple modification of the proof of [19, Lemma 1].

PROPOSITION 3.2. *Let* (z^l, w^l, y^l) *be fixed. Furthermore, let* $(\tilde{z}, \tilde{w}, \tilde{y})$ *be a regular solution of* (3.1). *Then there exists a neighborhood* $U_{\tilde{\delta}}$ *of* $(\tilde{z}, \tilde{w}, \tilde{y})$ *and* $\hat{\lambda} > 0$ *such that for all* $(z, w, y) \in U_{\tilde{\delta}}$, *we obtain*

$$||(z, w, y) - (\tilde{z}, \tilde{w}, \tilde{y})|| \leq \hat{\lambda} \, ||\hat{h}^l(z, w, y)||.$$

Under the assumption of regularity Proposition 3.2 guarantees that the quantity $||\hat{h}^l||$ is a good measure of the inexactness in a neighborhood of a solution to the mixed linear complementarity problem. One can expect that the smaller $||\hat{h}^l||$ is, the closer a point (z, w, y) is to the solution $(\tilde{z}, \tilde{w}, \tilde{y})$ of (3.1).

Similar to the above discussion, we can show that an iterate (z^l, w^l, y^l) is an exact solution of the mixed nonlinear complementarity problem (MNCP) if and only if the complementarity function (2.8) is zero. In particular, we state the following result.

LEMMA 3.3. *Let $\hat{h}$ be defined by* (2.8). *Then*

$$\hat{h}(z^*, w^*, y^*) \equiv 0 \quad \Longleftrightarrow \quad (z^*, w^*, y^*) \quad \textit{is a solution of} \text{ (MNCP)}.$$

Thus, the quantity $||\hat{h}(z^l, w^l, y^l)||$ measures how close the current iterate is to being a solution of (MNCP). We can expect that the larger $||\hat{h}(z^l, w^l, y^l)||$ is, the further away the current point is from an exact solution of the mixed nonlinear complementarity problem. On the other hand, the smaller it is, the more accurately the next subproblem is solved. In particular, we have the following proposition.

PROPOSITION 3.4. *Let (z^*, w^*, y^*) be a regular solution of the mixed nonlinear complementarity problem* (MNCP). *Then there exists a neighborhood U^* of (z^*, w^*, y^*) and a positive scalar $\hat{\lambda}_l$ such that whenever a point (z^l, w^l, y^l) is in U^*, we obtain*

$$||(z^l, w^l, y^l) - (z^*, w^*, y^*)|| \leq \hat{\lambda}_l \, ||\hat{h}(z^l, w^l, y^l)||.$$

Proof. Let $L > 0$, $\gamma^* > 0$, and U^*_δ be specified as in Definition 2.3. From the continuity of $\hat{h}$ we obtain the existence of a neighborhood U^* of (z^*, w^*, y^*) such that for all points (z^l, w^l, y^l) in U^* we get

$$v^l = \Big((z^l, w^l, y^l) - (\hat{h}_1(z^l, w^l, y^l), \hat{h}_2(z^l, w^l, y^l), \hat{h}_3(z^l, w^l, y^l))\Big)^T \in U^*_\delta, \quad ||\rho^l|| < \gamma^*,$$

where $\rho^l = (I - M)\hat{h}(z^l, w^l, y^l)$. Then, by the definition of $\hat{h}$ and $\hat{h}^l(z^l, w^l, y^l) = \hat{h}(z^l, w^l, y^l)$, it follows that $\hat{h}^l(v^l; \rho^l) = 0$, where $\hat{h}^l(v; \rho) := \hat{h}^l(v) - \rho$ denotes the perturbed linearized complementarity function according to $(\text{MLCP})_\rho$. Hence, the vector $v^l \in U^*_\delta$ is a solution of a perturbed mixed linear complementarity problem of the form $(\text{MLCP})_\rho$. Furthermore, the exact solution (z^*, w^*, y^*) of (MNCP) is regular. Then, by Definition 2.3 we know that $||(v^l_1, v^l_2, v^l_3) - (z^*, w^*, y^*)|| \leq L \, ||(I - M)\hat{h}(z^l, w^l, y^l)||$. Therefore, with $\hat{\lambda}_l = 1 + L \, ||I - M||$, we can deduce the desired result. ∎

4. Convergence rate for the inexact SQP method. In this section we state a local convergence result for the inexact SQP Algorithm 2.2. This method produces at every stage a new iterate according to the approximation rule (2.10). From the discussion of the previous section, we know that an iterate (z^l, w^l, y^l) is an exact solution of the mixed nonlinear complementarity problem (MNCP) if and only if the mixed nonlinear complementarity function (2.8) is equal to zero. Thus, the quantity $||\hat{h}(z^l, w^l, y^l)||$ measures how close the current iterate is to being a solution of (MNCP). As in Proposition 3.4 we can expect that the larger $||\hat{h}(z^l, w^l, y^l)||$ is, the further away the current point is from an exact solution of (MNCP). On the other hand, the smaller it is, the more accurately the next subproblem is solved.

We need two lemmas to establish the local convergence result. The first lemma is easy to prove and guarantees that under a sufficient differentiability assumption on

F, f, and g, the mixed nonlinear complementarity function $\hat{h}$ is Lipschitz continuous. The second lemma is a consequence of regularity. It states under the sole assumption of regularity that for sufficiently small perturbations the perturbed mixed linear complementarity problems are uniquely solvable in a vicinity of a regular solution of (MNCP). This result is a special case of Robinson [21, Theorem 2.4].

LEMMA 4.1. *Let F, f, and g be twice continuously differentiable. Then the mixed complementarity function $\hat{h}$ defined in* (2.8) *is Lipschitz continuous; i.e., there exist positive constants c_f, c_g, and $c_{\mathcal{L}}$ such that for all $(z^1, w^1, y^1), (z^2, w^2, y^2) \in \mathbb{R}^{n+m_e+m}$ we have*

$$||\hat{h}(z^1, w^1, y^1) - \hat{h}(z^2, w^2, y^2)|| \leq \hat{L}\, ||(z^1, w^1, y^1) - (z^2, w^2, y^2)||, \tag{4.1}$$

where $\hat{L} := c_{\mathcal{L}} + 2\, \max(1, c_f) + 2\, \sqrt{m}\, \max(1, c_g)$.

LEMMA 4.2. *Let (z^*, w^*, y^*) be a solution of* (MNCP). *Furthermore, suppose that Assumption* 2.1(ii) *and Assumption* 2.4 *are fulfilled. Then there exist a scalar $\gamma^* > 0$, two neighborhoods $U^*_{\delta_1}$ and $U^*_{\delta_2}$ of (z^*, w^*, y^*), and a Lipschitz constant $L > 0$ such that for all $(z, w, y) \in U^*_{\delta_1}$ and every perturbation $\rho = (\rho_1, \rho_2, \rho_3)^T \in \mathbb{R}^{n+m_e+m}$ with $||\rho|| < \gamma^*$, there is a unique solution vector*

$$\xi(z, w, y; \rho) := (\xi_1, \xi_2, \xi_3)^T \in U^*_{\delta_2}$$

depending on z, w, y, and ρ that solves the perturbed mixed linear complementarity problem defined by

$$\begin{gathered}
\xi_3 \geq 0, \quad \xi_3^T \left(-g_z^T(z)(\xi_1 - z) - g(z) - \rho_3\right) = 0, \\
-g_z^T(z)(\xi_1 - z) - g(z) - \rho_3 \geq 0, \\
\begin{bmatrix} F_z(z) \\ f(z) \end{bmatrix} + \begin{bmatrix} \mathcal{L}_{zz}(z, w, y) & f_z(z) & g_z(z) \\ f_z^T(z) & O & O \end{bmatrix} \begin{bmatrix} \xi_1 - z \\ \xi_2 \\ \xi_3 \end{bmatrix} - \begin{bmatrix} \rho_1 \\ \rho_2 \end{bmatrix} = 0.
\end{gathered}$$

Moreover, if $||\rho|| < \gamma^*$ and $||\hat{\rho}|| < \gamma^*$, then

$$||(\xi_1, \xi_2, \xi_3) - (\hat{\xi}_1, \hat{\xi}_2, \hat{\xi}_3)|| \leq L\, ||\rho - \hat{\rho}||,$$

where $\hat{\xi}(z, w, y; \hat{\rho}) := (\hat{\xi}_1, \hat{\xi}_2, \hat{\xi}_3)$ depends on z, w, y, and $\hat{\rho}$.

Now we are ready to show the local convergence result. A similar result can be found in Pang [19] for nonmixed nonlinear complementarity problems. For mixed nonlinear complementarity problems, i.e., those with equality as well as inequality relations, considered in this paper, the problem formulation, the definition of $\hat{h}$, and other relations have to be modified as described above. In particular, we are able to prove the following theorem.

THEOREM 4.3. *Let (z^*, w^*, y^*) be a solution triple of* (MNCP). *Furthermore, suppose that Assumption* 2.1(ii) *and Assumption* 2.4 *are fulfilled. Let $\hat{L} > 0$ be the Lipschitz constant specified in Lemma* 4.1 *and let $L > 0$ be the scalar as given in Lemma* 4.2. *Moreover, assume that for all l we have for some $\lambda \in (0, 1)$ and $\tilde{\lambda} > 1$*

$$\lambda_l \leq \lambda / \tilde{\lambda}, \tag{4.2}$$

where $\tilde{\lambda} = \max(1, \hat{L})\, (\, 1 + \max(1, L)\, c\,)$ and $c > 0$ is a given constant satisfying (4.8). *Then*

(i) *There exists an open ball* U^* *including* (z^*, w^*, y^*) *such that for all starting data* $(z^0, w^0, y^0) \in U^*$, *the sequence* $\{(z^l, w^l, y^l)\}$ *generated by the inexact SQP Algorithm* 2.2 *is well defined and converges to* (z^*, w^*, y^*) *with a linear rate; i.e., there is a constant* $r \in (0, 1)$ *such that*

$$||(z^{l+1}, w^{l+1}, y^{l+1}) - (z^*, w^*, y^*)|| \leq r\ ||(z^l, w^l, y^l) - (z^*, w^*, y^*)||; \tag{4.3}$$

(ii) *If, in addition to the assumptions in* (i), $\lim_{l\to\infty} \lambda_l = 0$, *then the rate of convergence is superlinear;*

(iii) *Suppose, in addition to the assumptions in* (i), *that the second derivatives of* F, f, *and* g *satisfy a Lipschitz condition in a neighborhood of* z^*. *Then, if*

$$\lambda_l \leq \hat{\beta}\ \|\hat{h}(z^l, w^l, y^l)\| \tag{4.4}$$

for some $\hat{\beta} > 0$, *the convergence of* $\{(z^l, w^l, y^l)\}$ *to* (z^*, w^*, y^*) *is quadratic.*

Proof. Let $U^*_{\delta_1}$, $U^*_{\delta_2}$, $\gamma^* > 0$, and $L > 0$ be specified as in Lemma 4.2. Since $\hat{h}(z^*, w^*, y^*) = 0$, Lemma 4.1 implies that with the Lipschitz constant $\hat{L} > 0$,

$$\left\|\hat{h}(z, w, y)\right\| = \left\|\hat{h}(z, w, y) - \hat{h}(z^*, w^*, y^*)\right\| \leq \hat{L}\ \|(z, w, y) - (z^*, w^*, y^*)\| \tag{4.5}$$

holds for all $(z, w, y) \in U^*_{\delta_1}$. By restricting $U^*_{\delta_1}$, if necessary, we can assume that

$$\lambda\ \|(z, w, y) - (z^*, w^*, y^*)\| < \gamma^* \quad \forall\, (z, w, y) \in U^*_{\delta_1} \text{ and } \lambda \in (0, 1). \tag{4.6}$$

Since h_1 and h_2 are continuously differentiable functions, we can assume, by restricting $U^*_{\delta_1}$ further, the existence of constant $\varepsilon > 0$ such that

$$\left\| \begin{bmatrix} h_1(z^*, w^*, y^*) \\ h_2(z^*, w^*, y^*) \end{bmatrix} - \begin{bmatrix} h_1(z, w, y) \\ h_2(z, w, y) \end{bmatrix} - \begin{bmatrix} \nabla h_1(z, w, y) \\ \nabla h_2(z, w, y) \end{bmatrix} ((z^*, w^*, y^*) - (z, w, y))^T \right\|$$

$$\leq \varepsilon\ \|(z, w, y) - (z^*, w^*, y^*)\| < \gamma^* \quad \forall\, (z, w, y) \in U^*_{\delta_1} \tag{4.7}$$

and $r := L\,\varepsilon + \lambda < 1$. Moreover, the norm of the gradient of h_1 and h_2 is bounded for all points in a vicinity of the regular solution (z^*, w^*, y^*). Hence, there is a constant $c > 0$ such that

$$\left\| \begin{bmatrix} I_1 & O \\ O & I_2 \end{bmatrix} - \begin{bmatrix} \nabla h_1(z, w, y) \\ \nabla h_2(z, w, y) \end{bmatrix} \right\| \leq c \quad \forall\, (z, w, y) \in U^*_{\delta_1}. \tag{4.8}$$

(i) We interpret the residual vector $\hat{h}^l$ as perturbation $\tilde{\rho}$ for solving $(\text{MLCP})_l$ at the lth stage of the inexact SQP algorithm and show that $(z^{l+1}, w^{l+1}, y^{l+1})$ solves a perturbed mixed linear complementarity problem. Then, choosing two perturbations ρ, $\hat{\rho}$ and perturbing $(z^{l+1}, w^{l+1}, y^{l+1})$ by $\tilde{\rho}$ (denoted by ξ^{l+1}), we conclude that ξ^{l+1} and (z^*, w^*, y^*) each solve a perturbed mixed linear complementarity problem. Finally, using the Lipschitz continuity of these solutions with respect to the perturbations and adding $0 = \tilde{\rho} - \tilde{\rho}$ to the left-hand side of (4.3), we show that (4.3) holds.

Let (z^0, w^0, y^0) be chosen in $U^*_{\delta_1}$. In general, assume that the point (z^l, w^l, y^l) lies in $U^*_{\delta_1}$. Let $(z^{l+1}, w^{l+1}, y^{l+1})$ be generated by the inexact SQP Algorithm 2.2; i.e., $(z^{l+1}, w^{l+1}, y^{l+1})$ is an approximate solution of $(\text{MLCP})_l$ satisfying (2.10). Setting $\tilde{\rho} := (\tilde{\rho}_1, \tilde{\rho}_2, \tilde{\rho}_3)^T = \hat{h}^l(z^{l+1}, w^{l+1}, y^{l+1})$, then it follows by (2.10), (4.2), and (4.5) that

$$||\tilde{\rho}|| \leq \lambda_l \hat{L} ||(z^l, w^l, y^l) - (z^*, w^*, y^*)|| \leq \frac{\lambda ||(z^l, w^l, y^l) - (z^*, w^*, y^*)||}{1 + \max(1, L)c}. \tag{4.9}$$

Thus, (4.9), together with (4.6), implies $||\tilde{\rho}|| < \gamma^*$. Using Lemma 4.2, we deduce that $(z^{l+1}, w^{l+1}, y^{l+1}) \in U^*_{\delta_2}$ satisfies uniquely the following perturbed mixed linear complementarity problem:

$$
\begin{aligned}
y^{l+1} \geq 0, \quad & \left(y^{l+1}\right)^T \left(-g_z^T(z^l)(z^{l+1} - z^l) - g(z^l) - \tilde{\rho}_3\right) = 0, \\
& -g_z^T(z^l)(z^{l+1} - z^l) - g(z^l) - \tilde{\rho}_3 \geq 0, \\
& \begin{bmatrix} F_z(z^l) \\ f(z^l) \end{bmatrix} + \nabla h_1(z^l, w^l, y^l) \begin{bmatrix} z^{l+1} - z^l \\ w^{l+1} \\ y^{l+1} \end{bmatrix} - \begin{bmatrix} \tilde{\rho}_1 \\ \tilde{\rho}_2 \end{bmatrix} = 0.
\end{aligned} \tag{4.10}
$$

Defining $\xi^{l+1} := \left(\xi_1^{l+1}, \xi_2^{l+1}, \xi_3^{l+1}\right)^T = \left((z^{l+1}, w^{l+1}, y^{l+1}) - (\tilde{\rho}_1, \tilde{\rho}_2, \tilde{\rho}_3)\right)^T$ and

$$
\rho := (\rho_1, \rho_2, \rho_3)^T = \left(\begin{bmatrix} I_1 & O \\ O & I_2 \end{bmatrix} - \begin{bmatrix} \nabla h_1(z^l, w^l, y^l) \\ \nabla h_2(z^l, w^l, y^l) \end{bmatrix} \right) \tilde{\rho},
$$

it follows by (4.6), (4.8), and (4.9), that $||\rho|| < \gamma^*$. Then, with the definition of $\tilde{\rho}$, ξ^{l+1}, ρ, and $y^{l+1} \geq \tilde{\rho}_3$, the vector ξ^{l+1} solves

$$
\begin{aligned}
\xi_3^{l+1} \geq 0, \quad & \left(\xi_3^{l+1}\right)^T \left(-g_z^T(z^l)(\xi_1^{l+1} - z^l) - g(z^l) - \rho_3\right) = 0, \\
& -g_z^T(z^l)(\xi_1^{l+1} - z^l) - g(z^l) - \rho_3 \geq 0, \\
& \begin{bmatrix} F_z(z^l) \\ f(z^l) \end{bmatrix} + \nabla h_1(z^l, w^l, y^l) \begin{bmatrix} \xi_1^{l+1} - z^l \\ \xi_2^{l+1} \\ \xi_3^{l+1} \end{bmatrix} - \begin{bmatrix} \rho_1 \\ \rho_2 \end{bmatrix} = 0.
\end{aligned} \tag{4.11}
$$

Hence, the continuity of $\hat{h}^l$, $\tilde{\rho} = \hat{h}^l(z^{l+1}, w^{l+1}, y^{l+1})$ and Lemma 4.2 imply that the vector $\xi^{l+1} \in U^*_{\delta_2}$ is the unique solution of (4.11). Setting $\hat{\rho} = [\hat{\rho}_{1,2}\ \hat{\rho}_3]^T$, where

$$
\begin{bmatrix} \hat{\rho}_{1,2} \\ \hat{\rho}_3 \end{bmatrix} = \begin{bmatrix} h_1(z^*, w^*, y^*) \\ h_2(z^*, w^*, y^*) \end{bmatrix} - \begin{bmatrix} h_1(z^l, w^l, y^l) \\ h_2(z^l, w^l, y^l) \end{bmatrix} - \begin{bmatrix} \nabla h_1(z^l, w^l, y^l) \\ \nabla h_2(z^l, w^l, y^l) \end{bmatrix} \begin{bmatrix} z^* - z^l \\ w^* - w^l \\ y^* - y^l \end{bmatrix},
$$

then by (4.7) we obtain $||\hat{\rho}|| \leq \varepsilon ||(z^*, w^*, y^*) - (z^l, w^l, y^l)|| < \gamma^*$. Since

$$
\begin{aligned}
h_1(z^*, w^*, y^*) &= h_1(z^l, w^l, y^l) + \hat{\rho}_{1,2} + \nabla h_1(z^l, w^l, y^l)((z^*, w^*, y^*) - (z^l, w^l, y^l))^T = 0, \\
h_2(z^*, w^*, y^*) &= h_2(z^l, w^l, y^l) + \hat{\rho}_3 + \nabla h_2(z^l, w^l, y^l)((z^*, w^*, y^*) - (z^l, w^l, y^l))^T \geq 0,
\end{aligned}
$$

it follows that the regular solution (z^*, w^*, y^*) satisfies the following perturbed mixed linear complementarity problem:

$$
\begin{aligned}
y^* \geq 0, \quad & \left(y^*\right)^T \left(-g_z^T(z^l)(z^* - z^l) - g(z^l) + \hat{\rho}_3\right) = 0, \\
& -g_z^T(z^l)(z^* - z^l) - g(z^l) + \hat{\rho}_3 \geq 0, \\
& h_1(z^l, w^l, y^l) + \hat{\rho}_{1,2} + \nabla h_1(z^l, w^l, y^l)((z^*, w^*, y^*) - (z^l, w^l, y^l))^T = 0.
\end{aligned} \tag{4.12}
$$

Then, Lemma 4.2 guarantees the Lipschitz continuity of the above solutions with respect to the perturbations; i.e., we obtain

$$
||(\xi_1^{l+1}, \xi_2^{l+1}, \xi_3^{l+1}) - (z^*, w^*, y^*)|| \leq L\, ||\rho - \hat{\rho}||. \tag{4.13}
$$

Hence, with $||\hat{\rho}|| \le \varepsilon||(z^*, w^*, y^*) - (z^l, w^l, y^l)||$, (2.10), (4.2), (4.8), (4.9), and (4.13), we have

$$\begin{aligned}||(z^{l+1}, w^{l+1}, y^{l+1}) - (z^*, w^*, y^*)|| &\le (\lambda_l \hat{L}(1 + Lc) + L\varepsilon)||(z^l, w^l, y^l) - (z^*, w^*, y^*)|| \\ &\le (\lambda + L\varepsilon)||(z^l, w^l, y^l) - (z^*, w^*, y^*)||.\end{aligned}$$

Since $r := \lambda + L\,\varepsilon < 1$ and $(z^l, w^l, y^l) \in U^*_{\delta_1}$ we obtain $(z^{l+1}, w^{l+1}, y^{l+1}) \in U^*_{\delta_1}$. Then, it follows that the point $(z^{l+1}, w^{l+1}, y^{l+1})$ generated by the inexact SQP Algorithm 2.2 is well defined, and by induction, we can conclude that the sequence $\{(z^l, w^l, y^l)\}$ is well defined. Furthermore, $\{(z^l, w^l, y^l)\}$ converges to (z^*, w^*, y^*) with a linear rate.

(ii) Suppose, in addition, that $\lim_{l\to\infty} \lambda_l = 0$. If we replace in (4.7) $r := L\,\varepsilon + \lambda < 1$, the constant ε by $\varepsilon_l > 0$ with $\lim_{l\to\infty} \varepsilon_l = 0$, then the same analysis as in (i) yields

$$||(z^{l+1}, w^{l+1}, y^{l+1}) - (z^*, w^*, y^*)|| \le r_l\, ||(z^l, w^l, y^l) - (z^*, w^*, y^*)||,$$

where $r_l = \lambda_l\, \hat{L}\, (1 + L\, c) + L\, \varepsilon_l \to 0$. But this implies the superlinear convergence.

(iii) Let, in addition, the assumptions in (iii) be fulfilled. The differentiability of h_1, h_2, and the Lipschitz condition imply in a neighborhood of (z^*, w^*, y^*) the Lipschitz continuity of ∇h_1 and ∇h_2. Then, by [18, Theorem 3.2.12], there exists $\beta > 0$ such that $||\hat{\rho}|| \le (\beta/2)\, ||(z^l, w^l, y^l) - (z^*, w^*, y^*)||^2 \quad \forall\, (z^l, w^l, y^l) \in U^*_{\delta_1}$. Thus, the same analysis as in (i) yields

$$||(z^{l+1}, w^{l+1}, y^{l+1}) - (z^*, w^*, y^*)|| \le \alpha\, ||(z^l, w^l, y^l) - (z^*, w^*, y^*)||^2,$$

where $\alpha = \hat{\beta}\, \hat{L}^2\, (1 + L\, c) + (L\, \beta/2)$. This and $(z^l, w^l, y^l) \to (z^*, w^*, y^*)$ establish the quadratic convergence rate of the considered sequence. □

This theorem shows that it is sufficient for the local convergence behavior to solve the quadratic subproblem, or, equivalently, the mixed linear complementarity problem only up to a certain accuracy. Dependent on the choice of the control sequence $\{\lambda_l\}$ in (2.10), this results in a linear, superlinear, or quadratic rate of convergence.

5. Interior point method for solving the quadratic subproblem. We have seen in the previous section that the quadratic subproblem, or, equivalently, the mixed linear complementarity problem in the inexact SQP method of Algorithm 2.2 should be solved by an iterative scheme. Since there are inequality constraints involved, we use an *interior point method*.

In order to describe the primal-dual interior point approach for the quadratic subproblem in general, we define

$$\begin{aligned} &A = f_z^T(z^l), \quad C = g_z^T(z^l), \quad Q = \mathcal{L}_{zz}(z^l, w^l, y^l), \\ &b = -f(z^l), \quad c = F_z(z^l), \quad d = -g(z^l), \quad \text{and} \quad \xi = \Delta z = z - z^l, \end{aligned} \tag{5.1}$$

where (z^l, w^l, y^l), $l = 0, 1, 2, \ldots$, is a given SQP iterate and $A \in \mathbb{R}^{m_e \times n}, C \in \mathbb{R}^{m \times n}, Q \in \mathbb{R}^{n \times n}, b \in \mathbb{R}^{m_e}, c \in \mathbb{R}^n, d \in \mathbb{R}^m$, and $\xi \in \mathbb{R}^n$. Then, the subproblem $(\mathrm{QP})_l$ can be rewritten as a general quadratic program, i.e.,

$$\min_{\xi}\ \frac{1}{2}\, \xi^T\, Q\, \xi + c^T \xi \quad \text{s.t.} \quad A\xi - b = 0, \quad C\xi - d \le 0. \tag{QP}$$

Let ξ^* be an optimal solution of (QP) and let the inequality constraints and μ^* be partitioned as

$$\nu_{\mathcal{A}^*} = d_{\mathcal{A}^*} - C_{\mathcal{A}^*}\, \xi^* = 0, \quad \mu^*_{\mathcal{A}^*} > 0 \quad \text{and} \quad \nu_{\mathcal{I}^*} = d_{\mathcal{I}^*} - C_{\mathcal{I}^*}\, \xi^* > 0, \quad \mu^*_{\mathcal{I}^*} = 0, \tag{5.2}$$

where $\mathcal{A}^*$ denotes the index set of the active and $\mathcal{I}^*$ denotes the index set of the nonbinding constraints at an optimal solution of (QP), and $\mu^* \in \mathbb{R}^m, \mu^* \geq 0$ denotes the Lagrange multiplier of the inequality constraints of (QP). With this partition we can formulate Assumption 5.1, which should hold throughout this section at an optimal solution ξ^* of (QP).

Assumption 5.1.

(i) There exists an optimal solution ξ^* of (QP).

(ii) Q is symmetric and positive definite on the null space $\mathcal{N}^*$ at ξ^*, where

$$\mathcal{N}^* := \{\xi \in \mathbb{R}^n \mid A\xi = 0, C_{\mathcal{A}^*}\xi = 0\}. \tag{5.3}$$

(iii) The matrix $[A\ C_{\mathcal{A}^*}]^T$ has full rank at the solution ξ^* of (QP).

For the special case (5.1) the following statement can be made. If Assumption 2.4 is fulfilled, we obtain that Assumption 5.1 holds in a neighborhood of ξ^*. In particular, Assumption 5.1(ii) is a convexity condition for the quadratic program and, together with Assumption 5.1(iii), a regularity condition, which guarantees the local solvability of (QP). Using the Karush–Kuhn–Tucker theorem, this assumption implies the existence of $\pi^* \in \mathbb{R}^{m_e}$ and $\mu^* \in \mathbb{R}^m, \mu^* \geq 0$, such that the point (ξ^*, π^*, μ^*) satisfies the Karush–Kuhn–Tucker conditions for (QP). Introducing a slack variable $\nu \in \mathbb{R}^m$, defined by $\nu = d - C\xi \geq 0$, the necessary optimality conditions can be restated as a mixed linear complementarity problem.

We arrive at the following nonlinear system with nonnegativity constraints on the variables μ and ν:

$$J(\xi, \pi, \mu, \nu) = \begin{pmatrix} Q\xi + A^T\pi + C^T\mu + c \\ A\xi - b \\ C\xi + \nu - d \\ NMe \end{pmatrix} = 0, \quad (\mu, \nu) \geq 0, \tag{5.4}$$

where $J : \mathbb{R}^{n+m_e+2m} \to \mathbb{R}^{n+m_e+2m}$ and $e = (1, \ldots, 1)^T \in \mathbb{R}^m$ and diagonal matrices $N, M \in \mathbb{R}^{m \times m}$

$$N = \operatorname{diag}(\nu_1, \ldots, \nu_m) \quad \text{and} \quad M = \operatorname{diag}(\mu_1, \ldots, \mu_m).$$

We denote the primal-dual feasibility set defined by

$$\mathcal{F} := \{(\xi, \pi, \mu, \nu) \in \mathbb{R}^{n+m_e+2m} \mid (\xi, \pi, \mu, \nu) \,\text{satisfies}\, (5.4)\} \tag{5.5}$$

and impose the following assumption.

Assumption 5.2. The interior of the primal-dual feasibility set is nonempty, i.e.,

$$\mathcal{F}^+ = \{(\xi, \pi, \mu, \nu) \in \mathcal{F} \mid \mu > 0, \nu > 0\} \neq \emptyset.$$

We say that points in the set $\mathcal{F}$ are feasible for (5.4) and strictly feasible, if they are elements of $\mathcal{F}^+$.

The interior point algorithm we consider in this paper is motivated by the application of the logarithmic barrier approach (cf. [7]) to the quadratic program. The unique minimum of the corresponding logarithmic barrier problem, if it exists, is characterized by the Karush–Kuhn–Tucker condition

$$J(\xi, \pi, \mu, \nu) - \begin{pmatrix} 0 \\ 0 \\ 0 \\ \eta e \end{pmatrix} = 0, \quad (\mu, \nu) > 0, \tag{5.6}$$

where $\eta > 0$ is the logarithmic barrier parameter and the function J is defined in (5.4).

The primal-dual interior point framework can now be formulated (cf., for example, [12] or [24] and the references therein). Recall that J is given by (5.4).

ALGORITHM 5.3. INTERIOR POINT METHOD.

Choose a starting point $(\xi^0, \pi^0, \mu^0, \nu^0) \in \mathcal{F}^+$.

For $k = 0, 1, 2, \ldots$ *until convergence do*

(I) *Choose* $\sigma_k \in [0, 1)$ *and set* $\eta_k = \sigma_k ({\mu^k}^T \nu^k)/m$.

(II) *Generate the search direction* $(\delta\xi^k, \delta\pi^k, \delta\mu^k, \delta\nu^k)$ *by solving the following system of equations:*

$$J'(\xi^k, \pi^k, \mu^k, \nu^k) \begin{pmatrix} \delta\xi^k \\ \delta\pi^k \\ \delta\mu^k \\ \delta\nu^k \end{pmatrix} = -J(\xi^k, \pi^k, \mu^k, \nu^k) + \begin{pmatrix} 0 \\ 0 \\ 0 \\ \eta_k e \end{pmatrix}. \tag{5.7}$$

(III) *Choose* $\tau_k \in (0, 1)$ *and compute the step length by*

$$\theta_k = \frac{-\tau_k}{\min_{1 \le i \le m}([(M^k)^{-1}\delta\mu^k]_i, [(N^k)^{-1}\delta\nu^k]_i, -\tau_k)}. \tag{5.8}$$

(IV) *Set* $(\xi^{k+1}, \pi^{k+1}, \mu^{k+1}, \nu^{k+1}) = (\xi^k, \pi^k, \mu^k, \nu^k) + \theta_k(\delta\xi^k, \delta\pi^k, \delta\mu^k, \delta\nu^k)$.

The following proposition guarantees the nonsingularity of the Jacobian matrix J' of the perturbed Newton system (5.7) in a neighborhood of the solution of (5.4).

PROPOSITION 5.4. *Let Assumption* 5.1 *be fulfilled. If* $\mu^k > 0$, $\nu^k > 0$, *and* $(\xi^k, \pi^k, \mu^k, \nu^k) \to (\xi^*, \pi^*, \mu^*, \nu^*)$, *then the Jacobian*

$$J'(\xi^k, \pi^k, \mu^k, \nu^k) = \begin{bmatrix} Q & A^T & C^T & O \\ A & O & O & O \\ C & O & O & I \\ O & O & N^k & M^k \end{bmatrix} \tag{5.9}$$

of the perturbed system (5.7) *is nonsingular for* k *sufficiently large.*

Proposition 5.4 ensures that the iterates produced by the interior point method of Algorithm 5.3 are locally well defined. Its proof is a direct consequence of the following result. Hence, we omit the proof.

A variant of the interior point method of Algorithm 5.3 generates a search direction by solving the following reduced symmetric, indefinite system of equations:

$$\begin{bmatrix} Q & A^T & C^T \\ A & O & O \\ C & O & -(M^k)^{-1}N^k \end{bmatrix} \begin{bmatrix} \delta\xi^k \\ \delta\pi^k \\ \delta\mu^k \end{bmatrix} = \begin{bmatrix} -Q\xi^k - A^T\pi^k - C^T\mu^k - c \\ -A\xi^k + b \\ -\eta_k(M^k)^{-1}e - C\xi^k + d \end{bmatrix}. \tag{5.10}$$

The eliminated variable $\delta\nu^k$ is set to

$$\delta\nu^k = \eta_k(M^k)^{-1}e - N^k e - (M^k)^{-1}N^k\delta\mu^k. \tag{5.11}$$

If we replace the perturbed Newton system (5.7) by (5.10) and (5.11), we can prove that the iterates produced by this variant of Algorithm 5.3 are locally well defined.

PROPOSITION 5.5. *Let Assumption* 5.1 *be fulfilled,* $\mu^k > 0$, $\nu^k > 0$, *and let* $D^k = (M^k)^{-1}N^k$ *be positive definite. If* $(\mu^k, \nu^k) \to (\mu^*, \nu^*)$ *then the matrix*

$$\hat{J}^k = \begin{bmatrix} Q & A^T & C^T \\ A & O & O \\ C & O & -D^k \end{bmatrix} \tag{5.12}$$

is nonsingular for k *sufficiently large.*

Proof. Let the assumptions be fulfilled and let $D^k = (M^k)^{-1}N^k$. Without loss of generality, the matrices C, D^k can be partitioned as

$$C = [C_{\mathcal{A}^*}\ C_{\mathcal{I}^*}]^T \quad \text{and} \quad D^k = \operatorname{diag}(D^k_{\mathcal{A}^*}, D^k_{\mathcal{I}^*}),$$

where $\mathcal{A}^*$ corresponds with the index set of the active and $\mathcal{I}^*$ with the index set of the nonbinding constraints at the solution of (QP). Note that the submatrices $D^k_{\mathcal{A}^*}$ and $D^k_{\mathcal{I}^*}$ are positive definite and that $\lim_{k\to\infty} D^k_{\mathcal{A}^*} = D^*_{\mathcal{A}^*} = 0$ and $\lim_{k\to\infty}(D^k_{\mathcal{I}^*})^{-1} = (D^*_{\mathcal{I}^*})^{-1} = 0$.

Assume that a subsequence $\{\hat{J}^{k_i}\}$ of $\{\hat{J}^k\}$ is singular. Then there exists a subsequence $\{(\mu^{k_i}, \nu^{k_i})\}$ of $\{(\mu^k, \nu^k)\}$ such that for $(z^{k_i}, w^{k_i}, y^{k_i}_{\mathcal{A}^*}, y^{k_i}_{\mathcal{I}^*}) \neq 0$

$$\begin{aligned} Qz^{k_i} + A^T w^{k_i} + C^T_{\mathcal{A}^*} y^{k_i}_{\mathcal{A}^*} + C^T_{\mathcal{I}^*} y^{k_i}_{\mathcal{I}^*} &= 0, \\ Az^{k_i} &= 0, \\ C_{\mathcal{A}^*} z^{k_i} - D^{k_i}_{\mathcal{A}^*} y^{k_i}_{\mathcal{A}^*} &= 0 \iff y^{k_i}_{\mathcal{A}^*} = (D^{k_i}_{\mathcal{A}^*})^{-1} C_{\mathcal{A}^*} z^{k_i}, \\ C_{\mathcal{I}^*} z^{k_i} - D^{k_i}_{\mathcal{I}^*} y^{k_i}_{\mathcal{I}^*} &= 0 \iff y^{k_i}_{\mathcal{I}^*} = (D^{k_i}_{\mathcal{I}^*})^{-1} C_{\mathcal{I}^*} z^{k_i}. \end{aligned} \tag{5.13}$$

If $z^{k_i} = 0$, we can conclude from the last equation of (5.13) that $y^{k_i}_{\mathcal{I}^*} = 0$. This, together with Assumption 5.1(iii), implies $w^{k_i} = 0$ and $y^{k_i}_{\mathcal{A}^*} = 0$. Hence, $z^{k_i} \neq 0$. From this sequence we can extract a subsequence, i.e., $\{(1/||z^{k_i}||)(z^{k_i}, w^{k_i}, y^{k_i}_{\mathcal{A}^*}, y^{k_i}_{\mathcal{I}^*})\}$, which we denote by $\{(z^{k_i}, w^{k_i}, y^{k_i}_{\mathcal{A}^*}, y^{k_i}_{\mathcal{I}^*})\}$, where $\lim_{i\to\infty} z^{k_i} = z^*$ and $||z^{k_i}|| = 1$ for some z^*. Then we have

$$\begin{aligned} Qz^{k_i} + C^T_{\mathcal{A}^*}(D^{k_i}_{\mathcal{A}^*})^{-1} C_{\mathcal{A}^*} z^{k_i} + C^T_{\mathcal{I}^*}(D^{k_i}_{\mathcal{I}^*})^{-1} C_{\mathcal{I}^*} z^{k_i} + A^T w^{k_i} &= 0, \\ Az^{k_i} &= 0. \end{aligned} \tag{5.14}$$

Premultiplying the first equation in (5.14) by $(z^{k_i})^T$ we deduce from (5.14) that

$$(z^{k_i})^T Q z^{k_i} + (z^{k_i})^T C^T_{\mathcal{A}^*}(D^{k_i}_{\mathcal{A}^*})^{-1} C_{\mathcal{A}^*} z^{k_i} + (z^{k_i})^T C^T_{\mathcal{I}^*}(D^{k_i}_{\mathcal{I}^*})^{-1} C_{\mathcal{I}^*} z^{k_i} = 0. \tag{5.15}$$

Then we obtain

$$(z^{k_i})^T C^T_{\mathcal{A}^*}(D^{k_i}_{\mathcal{A}^*})^{-1} C_{\mathcal{A}^*}\ z^{k_i} = -\ (z^{k_i})^T Q\ z^{k_i} - (z^{k_i})^T C^T_{\mathcal{I}^*}(D^{k_i}_{\mathcal{I}^*})^{-1} C_{\mathcal{I}^*}\ z^{k_i}.$$

Note that for all k_i we have

$$(z^{k_i})^T \left(C^T_{\mathcal{A}^*}(D^{k_i}_{\mathcal{A}^*})^{-1} C_{\mathcal{A}^*} \right) z^{k_i} \geq 0 \quad \text{and} \quad (z^{k_i})^T \left(C^T_{\mathcal{I}^*}(D^{k_i}_{\mathcal{I}^*})^{-1} C_{\mathcal{I}^*} \right) z^{k_i} \geq 0.$$

Moreover, for the subsequence $\{(\mu^{k_i}, \nu^{k_i})\}$ of $\{(\mu^k, \nu^k)\}$ we know that $D^{k_i}_{\mathcal{A}^*} \to 0$ and $(D^{k_i}_{\mathcal{I}^*})^{-1} \to 0$ if $i \to \infty$. Then we can conclude that

$$\limsup_{i\to\infty}\ (z^{k_i})^T C^T_{\mathcal{A}^*}(D^{k_i}_{\mathcal{A}^*})^{-1} C_{\mathcal{A}^*}\ z^{k_i} \leq -\ (z^*)^T Q\ z^*.$$

Since $D^{k_i}_{\mathcal{A}^*} \to 0$, we obtain $C_{\mathcal{A}^*}\, z^{k_i} \to C_{\mathcal{A}^*}\, z^* = 0$. Now, we deduce from (5.15) that $(z^*)^T Q z^* \leq 0$, which is a contradiction to Assumption 5.1 (ii). ∎

In the last part of this section we state some local convergence results for the interior point method of Algorithm 5.3. The first theorem yields a superlinear convergence rate of the Karush–Kuhn–Tucker system J to zero without the assumption of nonsingularity of the Jacobian J' at the solution of (5.4). Then, Theorem 5.6 can be obtained by making straightforward modifications to Theorems 3.1 and 3.3 of Zhang, Potra, and Tapia [24]. Its proof is omitted.

THEOREM 5.6. *Let $\{(\xi^k, \pi^k, \mu^k, \nu^k)\}$ be a sequence generated by the interior point method of Algorithm* 5.3 *and assume that $(\xi^k, \pi^k, \mu^k, \nu^k)$ converge to $(\xi^*, \pi^*, \mu^*, \nu^*)$. Furthermore, suppose that*

(i) *strict complementarity holds at the solution of* (5.4)*, i.e., that for all $i = 1, \ldots, m$ we have either $\mu_i^* > 0$ and $\nu_i^* = 0$ or $\mu_i^* = 0$ and $\nu_i^* > 0$;*
(ii) *the sequence $\{\kappa_k\}$ is bounded, where $\kappa_k = {\mu^k}^T \nu^k / (m \min(M^k N^k e))$;*
(iii) *Q is symmetric and positive semidefinite on $\mathcal{N}^*$;*
(iv) *$\sigma_k \to 0$ and $\tau_k \to 1$.*

Then for k sufficiently large the convergence of the sequence $\{J(\xi^k, \pi^k, \mu^k, \nu^k)\}$ to zero is superlinear and the point $(\xi^*, \pi^*, \mu^*, \nu^*)$ solves problem (5.4).

With some additional work, we can actually demonstrate that the complementarity sequence $\{M^k N^k e\}$ componentwise converges to zero with a superlinear rate.

Theorem 5.8 states the quadratic convergence of the sequence $\{(\xi^k, \pi^k, \mu^k, \nu^k)\}$ to $(\xi^*, \pi^*, \mu^*, \nu^*)$ in a vicinity of the solution. To obtain this result we have to assume the nonsingularity of the Jacobian matrix J' at the solution of (5.4). In particular, we state the following lemma, which is a direct consequence of Proposition 5.5.

LEMMA 5.7. *Let $(\xi^*, \pi^*, \mu^*, \nu^*)$ be a solution of* (5.4) *and assume that*

(i) *strict complementarity holds at the solution of* (5.4)*, i.e., that for all $i = 1, \ldots, m$ we have either $\mu_i^* > 0$ and $\nu_i^* = 0$ or $\mu_i^* = 0$ and $\nu_i^* > 0$;*
(ii) *the matrix $[A\ C_{\mathcal{A}^*}]^T$ has full rank at ξ^*;*
(iii) *Q is symmetric and positive definite on the null space $\mathcal{N}^*$ at ξ^*.*

Then the Jacobian matrix $J'(\xi^*, \pi^*, \mu^*, \nu^*)$ is nonsingular.

In particular, we have the following quadratic convergence result, which is attributed to Zhang, Tapia, and Dennis [23] in the context of linear programming, and it can be proved by making straightforward modifications to [23, Theorem 4.3]. Its proof is omitted.

THEOREM 5.8. *Let $\{(\xi^k, \pi^k, \mu^k, \nu^k)\}$ be a sequence generated by the interior point method of Algorithm* 5.3 *and assume that $(\xi^k, \pi^k, \mu^k, \nu^k)$ converges to $(\xi^*, \pi^*, \mu^*, \nu^*)$. Furthermore, suppose that*

(i) *strict complementarity holds at the solution of* (5.4)*, i.e., that for all $i = 1, \ldots, m$ we have either $\mu_i^* > 0$ and $\nu_i^* = 0$ or $\mu_i^* = 0$ and $\nu_i^* > 0$;*
(ii) *the matrix $[A\ C_{\mathcal{A}^*}]^T$ has full rank at ξ^*;*
(iii) *Q is symmetric and positive definite on the null space $\mathcal{N}^*$;*
(iv) *the parameters σ_k and τ_k satisfy in every iteration*

$$0 \leq \sigma_k \leq \min(\sigma, c_1 {\mu^k}^T \nu^k) \quad \text{and} \quad \max(\tau, 1 - c_2 {\mu^k}^T \nu^k) \leq \tau_k < 1,$$

where $\sigma \in [0, 1)$, $\tau \in (0, 1)$, and c_1, c_2 are positive constants.

Then the convergence of the sequence $\{(\xi^k, \pi^k, \mu^k, \nu^k)\}$ to $(\xi^*, \pi^*, \mu^*, \nu^*)$ is quadratic.

6. Numerical solution of a parabolic state-constrained control problem. As an example for an application of the inexact SQP interior point method we use a

parabolic control problem that has already been presented in the literature. Burger and Pogu [2] use Newton's method and a conjugate gradient method. In order to avoid the repeated solution of a nonlinear parabolic boundary value problem, Kupfer and Sachs [13] apply a reduced SQP method. Both consider an unconstrained parabolic control problem. In Leibfritz and Sachs [15] a state-constrained problem is solved by an exact SQP approach combined with an interior point solver for the quadratic subproblems and a band LR decomposition for the linear systems. In [16] the authors present implementation details of an inexact SQP interior point method and some numerical results.

For motivation we formulate the control problem in infinite dimension. The goal is to solve the discretized counterpart with our inexact SQP interior point algorithm. For further details we refer to Leibfritz and Sachs [15], [16]. Let $\phi(x,t)$ denote the temperature at time $t \in [0,T]$ and at $x \in [0,1]$, where $x = 0$ is on the boundary and $x = 1$ is inside the probe. Furthermore, we use the following notation:

C:	heat capacity,	λ:	heat conduction,
q:	source term,	p:	reference profile,
u:	control,	ϕ:	state,
$\tilde{\phi}_0$:	initial temperature distribution,	$\phi_{\max}$:	maximal temperature.

Then the parabolic state-constrained optimal control problem for a diffusion equation with boundary inputs and state constraints is given as follows:

$$\text{Minimize} \int_0^T (\phi(1,t) - p(t))^2 \, dt + \alpha \int_0^T u^2(t) \, dt, \tag{6.1}$$

subject to all $(\phi(x,t), u(t))$ satisfying the diffusion equation

$$C(\phi(x,t))\phi_t(x,t) - [\lambda(\phi(x,t))\phi_x(x,t)]_x = q(x,t), \quad (x,t) \in (0,1) \times (0,T), \tag{6.2}$$

with initial and boundary conditions

$$\begin{aligned} \lambda(\phi(0,t))\phi_x(0,t) &= g[\phi(0,t) - u(t)], & t \in (0,T), \\ \lambda(\phi(1,t))\phi_x(1,t) &= 0, & t \in (0,T), \\ \phi(x,0) &= \tilde{\phi}_0(x), & x \in (0,1), \end{aligned} \tag{6.3}$$

and the state constraint,

$$\phi(x,t) \leq \phi_{\max}, \quad (x,t) \in (0,1) \times (0,T). \tag{6.4}$$

Note that the optimal control problem has inequality constraints on the state which avoids overheating during the process. The positive constant α imposes a penalty on large values of the control. In Figure 6.1 we see that for the unconstrained problem, after the start of the process at $t = 0$, the heating leads to rather large temperatures at the boundary, where the probe is heated. This effect can be removed by an upper limit on these temperatures (see, e.g., Figure 6.2), which leads to the state constraint (6.4).

In the numerical tests we use the following parameters for the state-constrained optimal control problem:

$$T = 12.0, \quad \tilde{\phi}_0(x) \equiv 0, \quad x \in (0,1), \quad \phi_{\max} = 2.2, \quad \alpha = 1.0 \cdot 10^{-4}, \quad g = 1.0.$$

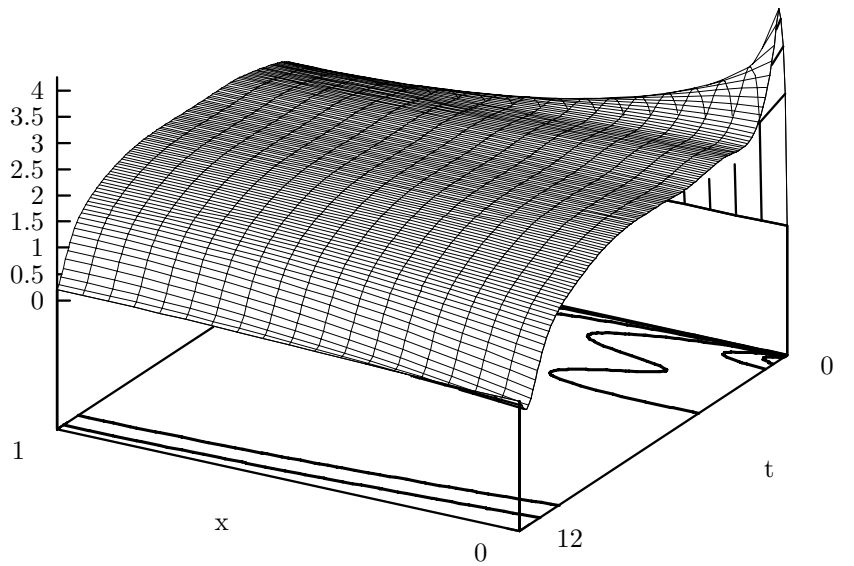


FIG. 6.1. *Unconstrained temperature distribution.*

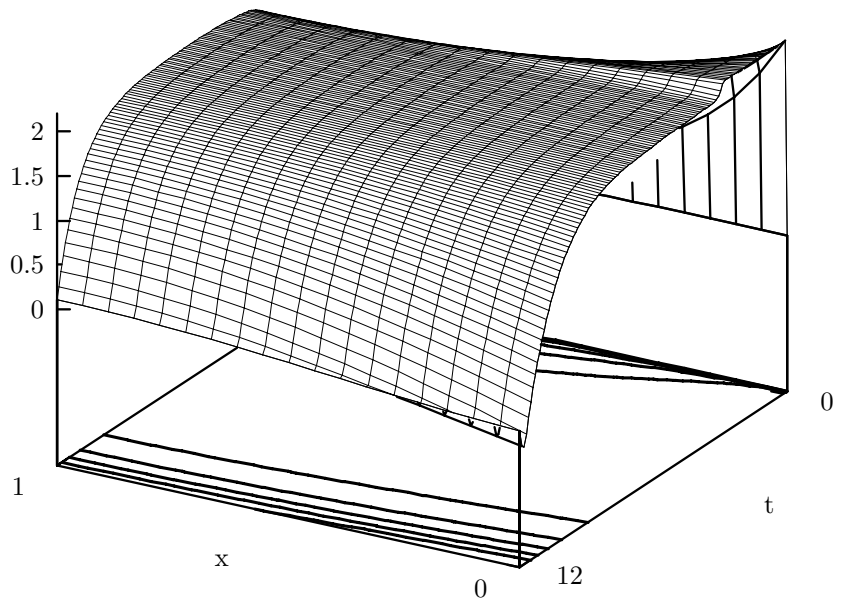


FIG. 6.2. *Computed temperature distribution* $\phi(x,t)$ *with* $\phi_{\max} = 2.2$.

For the discretization parameters N of the space grid and M of the time grid we use different values. The nonlinearities in the diffusion equation are described by

$$C(t) = q_1 + q_2 t, \quad \lambda(t) = r_1 + r_2 t,$$

and the source term at $(x,t) \in (0,1) \times (0,T)$ is given by

$$\begin{aligned} q(x,t) = &\left[\rho(q_1 + 2q_2) + \pi^2(r_1 + 2r_2)\right] \exp(\rho t) \cos(\pi x) \\ &- r_2 \pi^2 \exp(2\rho t) + \left(2r_2\pi^2 + \rho q_2\right) \exp(2\rho t) \cos(\pi x)^2, \end{aligned}$$

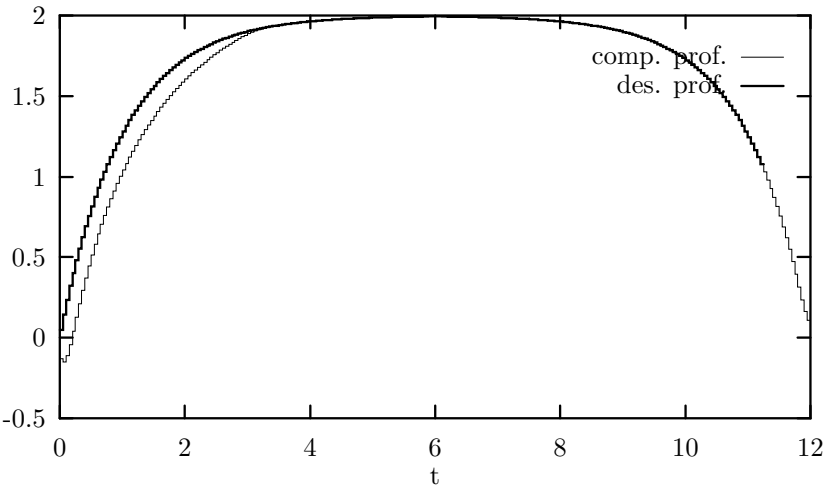


FIG. 6.3. *Temperature profile $p(t)$ and computed temperature profile $\phi(1,t)$.*

where $q_1 = r_1 = 4.0$, $q_2 = 1.0$, $r_2 = -1.0$, and $\rho = -1.0$. The temperature profile, which should be reached at $x = 1$ for $t \in (0,T)$, is defined by

$$p(t) = \begin{cases} 2 - 2\exp(\rho t), & 0 \le t \le 6, \\ 2 - 2\exp(\rho(T-t)), & 6 < t \le T, \\ 0, & t > T. \end{cases}$$

The graph in Figure 6.2 shows that the temperature is reduced at the right boundary and that the constraint is active for a certain time period. Furthermore, this figure demonstrates that the introduction of the inequality constraint avoids overheating of the whole process.

The desired and the achieved temperature distributions inside the probe at $x = 1$ are illustrated in Figure 6.3 and indicate that the approximation is satisfactory. In the time interval between $t = 0.0$ and $t \approx 3.0$, where the constraints are active on the right boundary, the computed temperature profile lies below the desired reference profile. After that, the profiles match each other. This profile exhibits a typical structure. During the time interval $[0,T]$ there is a heating phase followed by a phase where the temperature is kept nearly constant. Then the process is concluded by a cooling down phase.

The optimal control, which is due to the upper bound imposed on the temperature at the other boundary, is displayed in Figure 6.4.

TABLE 6.1
Quadratic convergence rate of the inexact SQP method.

	$N = 9$ $M = 40$		$N = 18$ $M = 120$		$N = 18$ $M = 240$	
l	IP	ε_l	IP	ε_l	IP	ε_l
1	7	$7.884 \cdot 10^{-1}$	18	$6.581 \cdot 10^{-0}$	7	$6.688 \cdot 10^{-0}$
2	5	$2.415 \cdot 10^{-3}$	5	$9.199 \cdot 10^{-1}$	5	$8.874 \cdot 10^{-1}$
3	6	$2.532 \cdot 10^{-8}$	6	$1.028 \cdot 10^{-3}$	7	$2.727 \cdot 10^{-3}$
4	–	–	6	$3.565 \cdot 10^{-8}$	6	$5.210 \cdot 10^{-8}$

In the last part of this section, we present several numerical tables for the state-

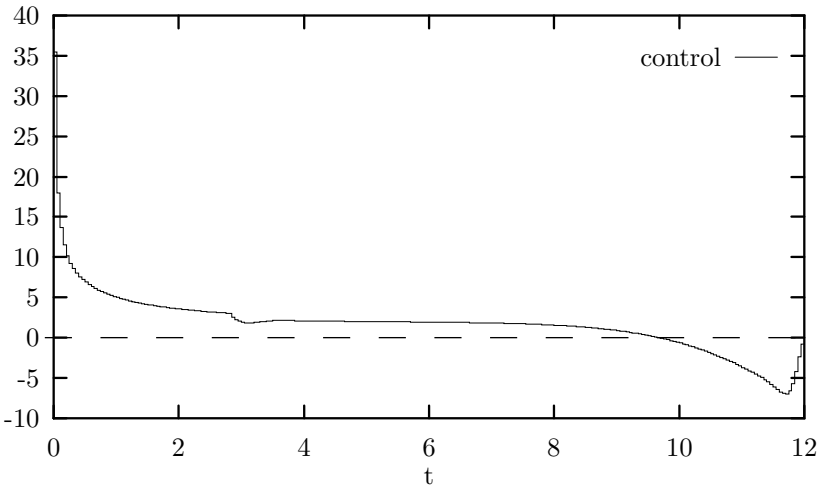


FIG. 6.4. *Computed optimal control $u(t)$ at $x = 0$.*

constrained parabolic control problem. All computations were performed in double-precision Fortran-77 on a Sun Sparcstation 10. In all numerical tests we have initialized our local inexact SQP Algorithm 2.2 with an approximate solution of the unconstrained control problem (cf. Figure 6.1). The forcing sequence $\{\lambda_l\}$ of the inexact SQP Algorithm 2.2 is chosen by (4.4). Then one can expect that the sequence $\{(z^l, w^l, y^l)\}$ generated by the inexact SQP Algorithm 2.2 converges locally to (z^*, w^*, y^*) with a quadratic rate (cf. Theorem 4.3). In every inexact SQP iteration, we solve the subproblems with the interior point method of Algorithm 5.3. There, we have chosen the parameters σ_k and τ_k as in Theorem 5.8 (iv) and we have used the constants

$$c_1 = c_2 = 1.0, \quad \sigma = 0.001, \quad \text{and} \quad \tau = 0.99995.$$

TABLE 6.2
Quadratic convergence rate of the interior point algorithm.

	$N = 18\ M = 120$			$N = 18\ M = 240$		
l	k	ε_k	ε_{k-1}^2	k	ε_k	ε_{k-1}^2
2	4	$2.912 \cdot 10^{-02}$	–	4	$6.940 \cdot 10^{-02}$	–
	5	$1.031 \cdot 10^{-04}$	$8.482 \cdot 10^{-04}$	5	$3.756 \cdot 10^{-03}$	$4.817 \cdot 10^{-03}$
3	4	$4.363 \cdot 10^{-02}$	–	6	$8.474 \cdot 10^{-04}$	–
	5	$6.481 \cdot 10^{-04}$	$1.911 \cdot 10^{-03}$	7	$1.580 \cdot 10^{-07}$	$7.181 \cdot 10^{-07}$
	6	$1.023 \cdot 10^{-07}$	$4.201 \cdot 10^{-07}$			
4	5	$3.286 \cdot 10^{-10}$	–	5	$1.357 \cdot 10^{-09}$	–
	6	$4.317 \cdot 10^{-19}$	$1.080 \cdot 10^{-19}$	6	$6.196 \cdot 10^{-18}$	$1.841 \cdot 10^{-18}$

The interior point algorithm will be terminated if the approximation rule (2.10) is satisfied and the linear system (5.10) is solved iteratively by GMRES. Finally, we terminate the inexact SQP approach if $\varepsilon_l = ||\hat{h}(z^l, w^l, y^l)|| \leq 10^{-08} \cdot \sqrt{M(3N+4)}$.

Table 6.1 shows the convergence rate of the inexact SQP method. There, IP indicates the number of interior point iterations for each inexact SQP iteration. We

observe that the inexact SQP approach locally achieves a quadratic rate of convergence according to Theorem 4.3. The results clearly demonstrate numerically the theoretical properties of the inexact SQP algorithm. Furthermore, we see that this approach achieves asymptotically the same convergence rates as an exact SQP method. In Table 6.2 we illustrate the quadratic convergence of the sequence generated by the interior point method during the last few inner iterations. During the first inexact SQP iteration the interior point algorithm terminates before achieving the vicinity of the solution of (QP), where we can observe the asymptotic rates. After that, we obtain a quadratic rate during the last few iterations according to Theorem 5.8. Furthermore, this table indicates that each quadratic subproblem is solved more accurately the more the iteration progresses. Finally, Table 6.3 shows the superlinear convergence of the duality gap during the last few iterations of the interior point algorithm according to Theorem 5.6.

TABLE 6.3
Superlinear convergence of the duality gap.

	$N = 18\ M = 120$			$N = 18\ M = 240$		
l	k	$d_k = {\mu^k}^T \nu^k$	d_k/d_{k-1}	k	$d_k = {\mu^k}^T \nu^k$	d_k/d_{k-1}
2	4	$8.083 \cdot 10^{-03}$	0.99813	4	$3.576 \cdot 10^{-03}$	0.99466
	5	$6.491 \cdot 10^{-05}$	0.00803	5	$9.937 \cdot 10^{-05}$	0.02779
3	4	$6.484 \cdot 10^{-05}$	0.99793	6	$3.835 \cdot 10^{-06}$	0.11196
	5	$3.001 \cdot 10^{-06}$	0.04628	7	$4.595 \cdot 10^{-08}$	0.01198
	6	$4.930 \cdot 10^{-09}$	0.00164			
4				2	$9.445 \cdot 10^{-05}$	0.99952
	3	$5.452 \cdot 10^{-06}$	0.99767	3	$9.436 \cdot 10^{-06}$	0.09990
	4	$6.927 \cdot 10^{-09}$	0.00127	4	$1.465 \cdot 10^{-08}$	0.00155
	5	$4.823 \cdot 10^{-13}$	0.00007	5	$6.735 \cdot 10^{-13}$	0.00005
	6	$2.904 \cdot 10^{-21}$	0.000000006	6	$2.205 \cdot 10^{-20}$	0.0000003

Acknowledgments. We thank the two referees and the associate editor for their careful reading of the paper and the suggestions they made toward improving it.

REFERENCES

[1] D. P. BERTSEKAS, *Constrained Optimization and Lagrange Multiplier Methods*, Academic Press, Orlando, San Diego, New York, London, 1982.

[2] J. BURGER AND M. POGU, *Functional and numerical solution of a control problem originating from heat transfer*, J. Optim. Theory Appl., 68 (1991), pp. 49–73.

[3] R. S. DEMBO, S. C. EISENSTAT, AND T. STEIHAUG, *Inexact Newton methods*, SIAM J. Numer. Anal., 19 (1982), pp. 400–408.

[4] R. S. DEMBO AND T. STEIHAUG, *Truncated Newton algorithms for large-scale unconstrained optimization*, Math. Programming, 26 (1983), pp. 190–212.

[5] R. S. DEMBO AND U. TULOWITZKI, *Sequential truncated quadratic programming methods*, in Numerical Optimization 1984: Proceedings of the SIAM Conference on Numerical Optimization, Boulder, CO, P. T. Boggs, R. H. Byrd, and R. B. Schnabel, eds., SIAM, Philadelphia, 1985, pp. 83–101.

[6] A. S. EL-BAKRY, R. A. TAPIA, T. TSUCHIYA, AND Y. ZHANG, *On the formulation and theory of Newton interior-point method for nonlinear programming*, J. Optim. Theory Appl., 89 (1996), pp. 507–541.

[7] A. V. FIACCO AND G. P. MCCORMICK, *Nonlinear Programming: Sequential Unconstrained Minimization Techniques*, Classics Appl. Math. 4, SIAM, Philadelphia, 1990.

[8] R. FONTECILLA, *On inexact quasi-Newton methods for constrained optimization*, in Numerical Optimization 1984: Proceedings of the SIAM Conference on Numerical Optimization,

Boulder, CO, P. T. Boggs, R. H. Byrd, and R. B. Schnabel, eds., SIAM, Philadelphia, 1985, pp. 102–118.

[9] P. T. Harker and J.-S. Pang, *Finite-dimensional variational inequality and nonlinear complementarity problems: A survey of theory, algorithms and applications*, Math. Programming, 48 (1990), pp. 161–220.

[10] M. Heinkenschloss and L. N. Vicente, *Analysis of Inexact Trust-Region Interior Point SQP Algorithms*, Tech. Rep. ICAM 95–06–01, Department of Mathematics, Virginia Polytechnic Institute and State University, Blacksburg, 1995.

[11] C. Kanzow and H. Kleinmichel, *A class of Newton-type methods for equality and inequality constrained optimization*, Optim. Methods Softw., 5 (1995), pp. 173–198.

[12] M. Kojima, N. Meggido, T. Noma, and A. Yoshise, *A Unified Approach to Interior Point Algorithms for Linear Complementarity Problems*, Lecture Notes in Comput. Sci. 538, Springer-Verlag, Berlin, Heidelberg, New York, London, Paris, Tokyo, 1991.

[13] F.-S. Kupfer and E. W. Sachs, *Numerical solution of a nonlinear parabolic control problem by a reduced SQP method*, Comput. Optim. Appl., 1 (1992), pp. 113–135.

[14] F. Leibfritz, *Logarithmic Barrier Methods for Solving the Optimal Output Feedback Problem*, Tech. Rep. Trierer Forschungsberichte Mathematik/Informatik 95–16, Universität Trier, Germany, 1995.

[15] F. Leibfritz and E. W. Sachs, *Numerical solution of parabolic state constrained control problems using SQP- and interior-point-methods*, in Large Scale Optimization: State of the Art, W. W. Hager, D. W. Hearn, and P. M. Pardalos, eds., Kluwer, Norwell, MA, 1994, pp. 251–264.

[16] F. Leibfritz and E. W. Sachs, *SQP interior point methods for parabolic control problems*, in Control of Partial Differential Equations and Applications: Proceedings of the IFIP WG 7.2 International Conference, Laredo, Lecture Notes in Pure and Appl. Math. 174, E. Casas and J. Yvon, eds., Marcel Dekker, New York, 1995, pp. 181–192.

[17] W. Murray and F. J. Prieto, *A sequential quadratic programming algorithm using an incomplete solution of the subproblem*, SIAM J. Optim., 5 (1995), pp. 590–640.

[18] J. Ortega and W. C. Rheinboldt, *Iterative Solution of Nonlinear Equations in Several Variables*, Academic Press, Orlando, San Diego, New York, London, 1970.

[19] J.-S. Pang, *Inexact Newton methods for the nonlinear complementarity problem*, Math. Programming, 36 (1986), pp. 54–71.

[20] J.-S. Pang, *Complementarity problems*, in Handbook of Global Optimization, R. Horst and P. M. Pardalos, eds., Nonconvex Optimization and Its Applications 2, Kluwer Academic Publishers, Dordrecht, Boston, London, 1995, pp. 271–338.

[21] S. M. Robinson, *Strongly regular generalized equations*, Math. Oper. Res., 5 (1980), pp. 43–62.

[22] S. J. Wright, *Interior point methods for optimal control of discrete time systems*, J. Optim. Theory Appl., 77 (1993), pp. 161–187.

[23] Y. Zhang, R. Tapia, and J. E. Dennis, Jr., *On the superlinear and quadratic convergence of primal-dual interior point linear programming algorithms*, SIAM J. Optim., 2 (1992), pp. 304–324.

[24] Y. Zhang, R. Tapia, and F. Potra, *On the superlinear convergence of interior-point algorithms for a general class of problems*, SIAM J. Optim., 3 (1993), pp. 413–422.

SIAM J. CONTROL OPTIM.
Vol. 38, No. 1, pp. 294–312

ON THE LAGRANGE–NEWTON–SQP METHOD FOR THE OPTIMAL CONTROL OF SEMILINEAR PARABOLIC EQUATIONS*

FREDI TRÖLTZSCH†

Abstract. A class of Lagrange–Newton–SQP methods is investigated for optimal control problems governed by semilinear parabolic initial-boundary value problems. Distributed and boundary controls are given, restricted by pointwise upper and lower bounds. The convergence of the method is discussed in appropriate Banach spaces. Based on a weak second order sufficient optimality condition for the reference solution, local quadratic convergence is proved. The proof is based on the theory of Newton methods for generalized equations in Banach spaces.

Key words. optimal control, parabolic equation, semilinear equation, sequential quadratic programming, Lagrange–Newton method, convergence analysis

AMS subject classifications. 49J20, 49M15, 65K10, 49K20

PII. S0363012998341423

1. Introduction.

This paper is concerned with the numerical analysis of a sequential quadratic programming (SQP) method for optimal control problems governed by semilinear parabolic equations. We extend convergence results obtained in the author's papers [31] and [32] for simplified cases. Here, we allow for distributed and boundary control. Moreover, terminal, distributed, and boundary observation are included in the objective functional. In contrast to the former papers, where a semigroup approach was chosen to deal with the parabolic equations, the theory is now presented in the framework of weak solutions relying on papers by Casas [7], Raymond and Zidani [28], and Schmidt [30]. We refer also to Heinkenschloss and Tröltzsch [15], where the convergence of an SQP method is proved for the optimal control of a phase field model. Including first order sufficient optimality conditions in the considerations, we are able to essentially weaken the second order sufficient optimality conditions needed to prove the convergence of the method. These sufficient conditions tighten the gap to the associated necessary ones. However, the approach requires a quite extensive analysis.

SQP methods for the optimal control of ODEs have already been the subject of many papers. We refer, for instance, to the discussion of quadratic convergence and the associated numerical examples by Alt [1], [2], Alt and Malanowski [5], [6], to the mesh independence principle in Alt [3], and to the numerical application by Machielsen [27]. Moreover, we refer to the more extensive references therein. For a paper standing in some sense between the control of ODEs and PDEs we refer to Alt, Sontag and Tröltzsch [4], who investigated the control of weakly singular Hammerstein integral equations. The case of semilinear elliptic PDEs was considered by Unger [34].

Following recent developments for ordinary differential equations, we adopt here the relation between the SQP method and a generalized Newton method. This approach makes the whole theory more transparent. We are able to apply known results on the convergence of generalized Newton methods in Banach spaces assuming the so

*Received by the editors July 2, 1998; accepted for publication May 17, 1999; published electronically December 21, 1999. This research was supported by Deutsche Forschungsgemeinschaft, under Project number Tr 302/1-2.
http://www.siam.org/journals/sicon/38-1/34142.html
†Fakultät für Mathematik, Technische Universität Chemnitz, D-09107 Chemnitz, Germany (f.troeltzsch@mathematik.tu-chemnitz.de).

called strong regularity at the optimal reference point. In this way, the convergence analysis is shorter, and we are able to concentrate on specific questions arising from the presence of partial differential equations.

Once the convergence of the Newton method is shown, we still need an extensive analysis to make the theory complete. We have to ensure the strong regularity by sufficient conditions and to show that the Newton steps can be performed by solving linear-quadratic control problems (SQP-method). This interplay between the Newton method and the SQP method is a specific feature, which cannot be derived from general results in Banach spaces, since we have to discuss pointwise relations.

We should underline that this paper does not aim to discuss the numerical application of the method. Any computation has to be connected with a discretization of the problem. This gives rise to consider approximation errors, stability estimates, the interplay between mesh adaption and precision (particularly delicate for PDEs), and the numerical implementation. Besides the fact that some of these questions are still unsolved, the presentation of the associated theory would go far beyond the scope of one paper. We understand the analysis of our paper as a general line applicable to any proof of convergence for these numerical methods. Some test examples close to this paper were presented by Goldberg and Tröltzsch [11], [12]. The fast convergence of the SQP method is demonstrated there by examples in spatial domains of dimension one and two relying on a fine discretization of the problems. Lagrange–Newton-type methods were also discussed for PDEs by Heinkenschloss and Sachs [14], Ito and Kunisch [16], [17], Kelley and Sachs [19], [20], [21], Kupfer and Sachs [23], Heinkenschloss [13], and Kunisch and Volkwein [22], who report in much more detail on the numerical details needed for an effective implementation.

The paper is organized as follows. Section 2 is concerned with existence and uniqueness of weak solutions for the equation of state. After stating the problem and associated necessary and sufficient optimality conditions in section 3, the generalized Newton method is established in section 4. The strong stability of the generalized equation is discussed in section 5, while section 6 is concerned with performing the Newton steps by SQP steps.

2. The equation of state. The dynamics of our control system are described by the semilinear parabolic initial-boundary value problem

$$\begin{aligned} y_t(x,t) + \ \text{div}\ (\mathcal{A}(x)\,\text{grad}_x y(x,t)) + d(x,t,y(x,t),v(x,t)) &= 0 \quad \text{in } Q, \\ \partial_\nu y(x,t) + b(x,t,y(x,t),u(x,t)) &= 0 \quad \text{on } \Sigma, \\ y(x,0) - y_0(x) &= 0 \quad \text{on } \Omega\,. \end{aligned} \tag{2.1}$$

This system is considered in $Q = \Omega \times (0,T)$, where $\Omega \subset \mathbb{R}^N (N \geq 2)$ is a bounded domain and $T > 0$ a fixed time. By ∂_ν the co-normal derivative $\partial y/\partial \nu_A = -\nu^\top \mathcal{A} \nabla y$ is denoted, where ν is the outward normal on Γ. The functions u and v denote *boundary* and *distributed control*, $\Sigma = \Gamma \times (0,T)$, $\Gamma = \partial\Omega$, and y_0 is a fixed initial state function. Following [7] and [28], we impose the following assumptions on the data.

(A1) Γ is of class $C^{2,\alpha}$ for some $\alpha \in (0,1]$. The coefficients a_{ij} of the matrix $\mathcal{A} = (a_{ij})_{i,j=1,\ldots,N}$ belong to $C^{1,\alpha}(\overline{\Omega})$, and there is $m_0 > 0$ such that

$$-\xi^\top \mathcal{A}(x)\,\xi \geq m_0\,|\xi|^2 \quad \forall \xi \in \mathbb{R}^N \qquad \forall x \in \overline{\Omega}. \tag{2.2}$$

$\mathcal{A}(x)$ is (w.l.o.g.) symmetric .

(A2) The "distributed" nonlinearity $d = d(x,t,y,v)$ is defined on $\overline{Q} \times \mathbb{R}^2$ and satisfies the following Carathéodory-type condition:

(i) For all $(y,v) \in \mathbb{R}^2$, $d(\cdot,\cdot,y,v)$ is Lebesgue measurable on Q.
(ii) For almost all $(x,t) \in Q$, $d(x,t,\cdot,\cdot)$ is of class $C^{2,1}(\mathbb{R}^2)$.
The "boundary" nonlinearity $b = b(x,t,y,u)$ is defined on $\Sigma \times \mathbb{R}^2$ and is supposed to fulfill (i), (ii) with Σ substituted for Q.

In our setting, the controls u, v will be uniformly bounded by a certain constant K.

(A3) The functions d, b fulfill the *assumptions of boundedness and monotonicity*
(i)

$$|d(x,t,0,v)| \le d_K(x,t) \quad \forall (x,t) \in Q, |v| \le K, \tag{2.3}$$

where $d_K \in L^q(Q)$ and $q > \frac{N}{2} + 1$. There is a number $c_0 \in \mathbb{R}$ and a nondecreasing function $\eta : \mathbb{R}_+ \to \mathbb{R}_+$ such that

$$c_0 \le d_y(x,t,y,v) \le \eta(|y|) \tag{2.4}$$

for almost everywhere (a.e.) $(x,t) \in Q$, all $y \in \mathbb{R}$, all $|v| \le K$.
(ii)

$$|b(x,t,0,u)| \le b_K(x,t) \quad \forall (x,t) \in \Sigma, |u| \le K \tag{2.5}$$

and

$$c_0 \le b_y(x,t,y,u) \le \eta(|y|) \tag{2.6}$$

for a.e. $(x,t) \in \Sigma$, all $y \in \mathbb{R}$, all $|u| \le K$, where $b_K \in L^r(\Sigma)$, $r > N+1$.

The assumptions imply those supposed in [7], [28], since our controls are uniformly bounded. The $C^{2,1}$-assumption on d, b is not necessary for the discussion of the equation of state. We shall need it for the Lagrange–Newton method. Although the discussion of existence and uniqueness for the nonlinear system (2.1) is not necessary for our analysis we quote the following result from [7], [28].

THEOREM 2.1. *Suppose that* (A1)–(A3) *are satisfied,* $y_0 \in C(\overline{\Omega})$, $v \in L^\infty(Q)$, $u \in L^\infty(\Sigma)$. *Then the system* (2.1) *admits a unique weak solution* $y \in L^2(0,T;H^1(\Omega)) \cap C(\overline{Q})$.

A *weak solution* of (2.1) is a function y of $L^2(0,T;H^1(\Omega)) \cap C(\overline{Q})$ such that

$$\begin{aligned} &-\int_Q (y \cdot p_t + (\nabla_x y)^\top \mathcal{A}(x) \nabla_x p)\, dxdt + \int_Q d(x,t,y,v)\ p\ dxdt \\ &\quad + \int_\Sigma b(x,t,y,u)\ p\ dSdt - \int_\Omega y_0(x) p(x,0) dx = 0 \end{aligned} \tag{2.7}$$

holds $\forall p \in W_2^{1,1}(Q)$ satisfying $p(x,T) = 0$ [24]. In (2.7) we have assumed that $y \in C(\overline{Q})$ to make the nonlinearities d, b well defined. Theorem 2.1 was shown by a detailed discussion of regularity for an associated linear equation. This linear version of Theorem 2.1 is more important for our analysis. In what follows, we shall use the symbol $A = \operatorname{div}(\mathcal{A} \operatorname{grad} y)$. Moreover, we need the space $W(0,T) = \{y \in L^2(0,T;H^1(\Omega)) | y_t \in L^2(0,T;H^1(\Omega)')\}$[25], [26]. Regard the linear initial-boundary value problem

$$\begin{aligned} y_t + A\,y + a\,y &= v && \text{on } Q, \\ \partial_\nu y + b\,y &= u && \text{on } \Sigma, \\ y(0) &= y_0 && \text{on } \Omega. \end{aligned} \tag{2.8}$$

THEOREM 2.2. *Suppose that $a \in L^\infty(Q)$, $b \in L^\infty(\Sigma)$, $q > N/2+1$, $r > N+1$, $a(x,t) \geq c_0$, $b(x,t) \geq c_0$ a.e. on Q and Σ, respectively, and $y_0 \in C(\overline{\Omega})$. Then there is a constant $c_l = c(c_0, q, r, m_0, \Omega, T)$ not depending on a, b, v, u, y_0 such that*

$$\|y\|_{L^2(0,T;H^1(\Omega))} + \|y\|_{C(\overline{Q})} \leq c_l \left(\|v\|_{L^q(Q)} + \|u\|_{L^r(\Sigma)} + \|y_0\|_{C(\overline{\Omega})}\right) \tag{2.9}$$

holds for the weak solution of the linear system (2.8).

For the proof we refer to [7] or [28]. Equation (2.9) yields a similar estimate for $b \cdot y$. Regarding the linear system (2.8) with right-hand sides $v - ay$, $u - by$, and y_0, respectively, the L^2-theory of linear parabolic equations applies to derive

$$\|y\|_{W(0,T)} \leq c'(\|v\|_{L^q(Q)} + \|u\|_{L^r(\Sigma)} + \|y_0\|_{C(\overline{\Omega})}), \tag{2.10}$$

where c_l' depends also on $\|a\|_{L^\infty(Q)}, \|b\|_{L^\infty(\Sigma)}$. We shall work in the state space $Y = \{y \in W(0,T) \,|\, y_t + Ay \in L^q(Q),\ \partial_\nu y \in L^r(\Sigma)\, y(0) \in C(\overline{\Omega})\}$ endowed with the norm $\|y\|_Y := \|y\|_{W(0,T)} + \|y_t + Ay\|_{L^q(Q)} + \|\partial_\nu y\|_{L^r(\Sigma)} + \|y(0)\|_{C(\overline{\Omega})}$. Y is known to be continuously embedded into $C(\overline{Q})$. From (2.9) and (2.10) we get

$$\|y\|_Y \leq \tilde{c}_l \left(\|v\|_{L^q(Q)} + \|u\|_{L^r(\Sigma)} + \|y_0\|_{C(\overline{\Omega})}\right), \tag{2.11}$$

where $\tilde{c}_l$ depends on $c_0, q, r, m_0, \Omega, T, \|a\|_{L^\infty(Q)}, \|b\|_{L^\infty(\Sigma)}$. Further on, we shall need the Hilbert space $H = W(0,T) \times L^2(\Omega) \times L^2(\Sigma)$ equipped with the norm $\|(y,v,u)\|_H := (\|y\|^2_{W(0,T)} + \|v\|^2_{L^2(Q)} + \|u\|^2_{L^2(\Sigma)})^{1/2}$.

3. Optimal control problem and SQP method. Let $\varphi : \Omega \times \mathbb{R} \to R$, $f : Q \times \mathbb{R}^2 \to \mathbb{R}$, and $g : \Sigma \times \mathbb{R}^2 \to \mathbb{R}$ be given functions specified below. Consider the problem (P) to minimize

$$J(y,v,u) = \int_\Omega \varphi(x, y(x,T))dx + \int_Q f(x,t,y,v)dxdt + \int_\Sigma g(x,t,y,u)dSdt \tag{3.1}$$

subject to the state equation (2.1) and to the pointwise constraints on the control

$$v_a \leq v(x,t) \leq v_b \quad \text{a.e. on } Q, \tag{3.2}$$

$$u_a \leq u(x,t) \leq u_b \quad \text{a.e. on } \Sigma, \tag{3.3}$$

where v_a, v_b, u_a, u_b are given functions of $L^\infty(Q)$ and $L^\infty(\Sigma)$, respectively, such that $v_a \leq v_b$ a.e. on Q and $u_a \leq u_b$ a.e. on Σ. The *controls* v and u belong to the *sets of admissible controls*

$$V_{ad} = \{v \in L^\infty(Q) \,|\, v \text{ satisfies } (3.2)\}, \qquad U_{ad} = \{u \in L^\infty(\Sigma) \,|\, u \text{ satisfies } (3.3)\}.$$

(P) is a nonconvex programming problem, hence different local minima will possibly occur. Numerical methods will deliver a local minimum close to their starting point. Therefore, we do not restrict our investigations to global solutions of (P). We will assume later that a fixed *reference solution* is given satisfying certain first and second order optimality conditions (ensuring local optimality of the solution). For the same reason, we shall not discuss the problem of existence of global (optimal) solutions for (P).

In the next assumptions, D^2 will denote Hessian matrices of functions. The functions φ, f, d, g, and b are assumed to satisfy the following assumptions on smoothness and growth. Here, $\|\cdot\|$ denotes any useful norm for 2×2 matrices.

(A4) For all $x \in \Omega$, $\varphi(x,\cdot)$ belongs to $C^{2,1}(\mathbb{R})$ with respect to $y \in \mathbb{R}$, while $\varphi(\cdot,y)$, $\varphi_y(\cdot,y)$, and $\varphi_{yy}(\cdot,y)$ are bounded and measurable on Ω. There is a constant $c_K > 0$ such that

$$|\varphi_{yy}(x,y_1) - \varphi_{yy}(x,y_2)| \le c_K\,|y_1 - y_2| \tag{3.4}$$

holds $\forall\, y_i \in \mathbb{R}$ such that $|y_j| \le K$, $i = 1,2$.
For all $(x,t) \in Q$, $f(x,t,\cdot,\cdot)$ is of class $C^{2,1}(\mathbb{R}^2)$ with respect to $(y,v) \in \mathbb{R}^2$, while f, f_y, f_v, f_{yy}, f_{yv}, and f_{vv}, all depending on $(\cdot,\cdot,y,v)$, are bounded and measurable with respect to $(x,t) \in Q$. There is a constant $f_K > 0$ such that

$$\|D^2 f(x,t,y_1,v_1) - D^2 f(x,t,y_2,v_2)\| \le f_K\,(|y_1 - y_2| + |v_1 - v_2|) \tag{3.5}$$

holds $\forall\, y_i, v_i$ satisfying $|y_i| \le K$, $|v_i| \le K$, $i = 1,2$, and almost all $(x,t) \in Q$.
For d we require the same except the boundedness of $d(\cdot,\cdot,y,v)$.
The functions g and d satisfy analogous assumptions on $\Sigma \times \mathbb{R}^2$. In particular,

$$\|D^2 g(x,t,y_1,u_1) - D^2 g(x,t,y_2,u_2)\| \le g_K\,(|y_1 - y_2| + |u_1 - u_2|) \tag{3.6}$$

holds $\forall\, y_i, u_i$ satisfying $|y_i| \le K$, $|u_i| \le K$, $i = 1,2$, and almost all $(x,t) \in \Sigma$.

Let us recall the known standard *first order necessary optimality system* for a local minimizer (y,v,u) of (P). The triplet (y,v,u) has to satisfy, together with an *adjoint state* $p \in W(0,T)$, the state system (2.1), the constraints $v \in V_{ad}$, $u \in U_{ad}$, the *adjoint equation*

$$\begin{aligned} -p_t + A\,p + d_y(x,t,y,v)\,p &= f_y(x,t,y,v) && \text{in } Q,\\ \partial_\nu p + b_y(x,t,y,u)\,p &= g_y(x,t,y,u) && \text{on } \Sigma,\\ p\,(x,T) &= \varphi_y(x,y(x,T)) && \text{in } \Omega, \end{aligned} \tag{3.7}$$

and the *variational inequalities*

$$\int_Q (f_v(x,t,y,v) - d_v(x,t,y,v)\cdot p)(z - v)\,dxdt \ge 0 \quad \forall z \in V_{ad}, \tag{3.8}$$

$$\int_\Sigma (g_u(x,t,y,u) - b_u(x,t,y,u)\cdot p)(z - u)\,dSdt \ge 0 \quad \forall z \in U_{ad}. \tag{3.9}$$

We introduce for convenience the *Lagrange function* $\mathcal{L}$,

$$\begin{aligned} \mathcal{L}(y,v,u;p) &= J(y,v,u) - \int_Q \{(y_t + Ay + d(x,t,y,v)\}\,p\,dxdt\\ &\quad - \int_\Sigma \{\partial_\nu y + b(x,t,y,v)\}\,p\,dSdt, \end{aligned} \tag{3.10}$$

defined on $Y \times L^\infty(Q) \times L^\infty(\Sigma) \times W(0,T)$. $\mathcal{L}$ is of class $C^{2,1}$ with respect to (y,v,u) in $Y \times L^\infty(Q) \times L^\infty(\Sigma)$. Moreover, we define the *Hamilton functions*

$$H^Q = H^Q(x,t,y,p,v) = f(x,t,y,v) - p\,d(x,t,y,v), \tag{3.11}$$

$$H^\Sigma = H^\Sigma(x,t,y,p,u) = g(x,t,y,u) - p\,b(x,t,y,u), \tag{3.12}$$

containing the "nondifferential" parts of $\mathcal{L}$. Then the relations (3.7)–(3.9) imply

$$\mathcal{L}_y(y,v,u;p)h = 0 \quad \forall h \in W(0,T) \text{ satisfying } h(0) = 0, \tag{3.13}$$

$$\mathcal{L}_v(y,v,u;p)(z - v) = \int_Q H^Q_v(x,t,y,p,v)(z - v)\,dxdt \ge 0 \quad \forall z \in V_{ad}, \tag{3.14}$$

$$\mathcal{L}_u(y,v,u;p)(z - u) = \int_\Sigma H^\Sigma_u(x,t,y,p,u)(z - u)dSdt \ge 0 \quad \forall z \in U_{ad}. \tag{3.15}$$

Let us suppose once and for all that a fixed *reference triplet* $(\overline{y}, \overline{v}, \overline{u}) \in Y \times L^\infty(Q) \times L^\infty(\Sigma)$ is given satisfying together with $\overline{p} \in W(0,T)$ the optimality system. This system is not sufficient for local optimality. Therefore, we shall assume some kind of *second order sufficient conditions*. We have to consider them along with a *first order sufficient condition*. Following Dontchev, et al. [10], the sets

$$Q(\sigma) = \{(x,t) \in Q \,|\, |\, H_v^Q(x,t,\overline{y}(x,t),\overline{v}(x,t),\overline{p}(x,t))\,| \geq \sigma\}, \tag{3.16}$$

$$\Sigma(\sigma) = \{(x,t) \in \Sigma \,|\, |\, H_u^\Sigma(x,t,\overline{y}(x,t),\overline{u}(x,t),\overline{p}(x,t))\,| \geq \sigma\} \tag{3.17}$$

are defined for arbitrarily small but fixed $\sigma > 0$. On these sets, the controls $\overline{u}$ and $\overline{v}$ are uniquely defined by the first order optimality conditions; they attain the lower or upper bound of their control set. Here we do not need second order information, since first order sufficiency applies. Although the same holds for all points, where $|H_v^Q|$ and $|H_u^\Sigma|$ are positive, we need the level $\sigma > 0$ to make this property stable with respect to small perturbations. The sets $Q(\sigma)$ and $\Sigma(\sigma)$ define strongly active control constraints in the sense of optimization theory, since $|H_v^Q|$ and $|H_u^\Sigma|$ are Lagrange multipliers associated to the control constraints. However, we shall not need this interpretation. D^2H^Q and D^2H^Σ denote the Hessian matrices of H^Q, H^Σ with respect to (y,v) and (y,u) respectively, taken at the reference point. For instance,

$$D^2H^Q(x,t) = \begin{pmatrix} H_{yy}^Q(x,t,\overline{y}(x,t),\overline{v}(x,t),\overline{p}(x,t)) & H_{yv}^Q(x,t,\overline{y}(x,t),\overline{v}(x,t),\overline{p}(x,t)) \\ H_{vy}^Q(x,t,\overline{y}(x,t),\overline{v}(x,t),\overline{p}(x,t)) & H_{vv}^Q(x,t,\overline{y}(x,t),\overline{v}(x,t),\overline{p}(x,t)) \end{pmatrix}.$$

D^2H^Σ is defined analogously. Moreover, we introduce a quadratic form B depending on $h_i = (y_i, v_i, u_i) \in Y \times L^\infty(Q) \times L^\infty(\Sigma),\ i = 1,2$, by

$$\begin{aligned} B[h_1,h_2] \;=\; & \int_\Omega \varphi_{yy}(x,\overline{y}(x,T))y_1(x,T)y_2(x,T)\,dx + \int_Q (y_1,v_1)D^2H^Q(y_2,v_2)^\top\,dxdt \\ & + \int_\Sigma (y_1,u_1)D^2H^\Sigma(y_2,u_2)^\top\,dSdt. \end{aligned} \tag{3.18}$$

The *second order sufficient optimality condition* is defined as follows:

(SSC) There are $\delta > 0, \sigma > 0$ such that

$$B[h,h] \geq \delta \cdot \|h\|_H^2 \tag{3.19}$$

holds $\forall\, h = (y,v,u) \in W(0,T) \times L^2(Q) \times L^2(\Sigma)$, where $v \in V_{ad}, v(x,t) = 0$ on $Q(\sigma), u \in U_{ad}, u = 0$ on $\Sigma(\sigma)$, and y is the associated weak solution of the linearized equation

$$\begin{aligned} y_t + A\,y + d_y(\overline{y},\overline{v})\,y + d_v(\overline{y},\overline{v})\,v &= 0, \\ \partial_\nu y + b_y(\overline{y},\overline{u})\,y + b_u(\overline{y},\overline{u})\,u &= 0, \\ y(0) &= 0. \end{aligned} \tag{3.20}$$

Next we introduce the SQP method to solve the problem (P) iteratively. Let us first assume that the controls are unrestricted, that is $V_{ad} = L^\infty(Q)$, $U_{ad} = L^\infty(\Sigma)$. Then the optimality system (2.1), (3.7), (3.8), (3.9) is a nonlinear system of equations for the unknown functions v, p, y, u, which can be treated by the Newton method. (For unrestricted controls, the variational inequalities are equivalent to the equations

$f_v - d_v p = 0$ and $g_u - b_u p = 0$.) In each step of the method, a linear system of equations is to be solved. This linear system is the optimality system of a linear-quadratic optimal control problem without constraints on the controls, which can be solved instead of the linear system of equations.

Now consider again the constraints on the control. Then the optimality system is no longer a system of equations, since it contains the two variational inequalities (3.8) and (3.9). However, it is not difficult to generalize the linear-quadratic control problems by adding the control-constraints. This idea leads to the following iterative method: Suppose that (y_i, p_i, v_i, u_i), $i = 1, \dots, n$, have already been determined. Then $(y_{n+1}, v_{n+1}, u_{n+1})$ is computed by solving the following linear-quadratic optimal control problem (QP_n) :

(QP_n) Minimize

$$
\begin{aligned}
J_n(y,v,u) = &\int_\Omega \varphi_y^n \cdot y(T)dx + \int_Q (f_y^n \cdot y + f_v^n \cdot v)\, dxdt + \int_\Sigma (g_y^n\ y + g_u^n\ u) dSdt \\
&+\tfrac{1}{2}\int_\Omega \varphi_{yy}^n\ (y(T) - y_n(T))^2 dx + \frac{1}{2}\int_Q (y - y_n, v - v_n) D^2 H^{Q,n} \begin{pmatrix} y - y_n \\ v - v_n \end{pmatrix} dxdt \\
&+\tfrac{1}{2}\int_\Sigma (y - y_n, u - u_n) D^2 H^{\Sigma,n} \begin{pmatrix} y - y_n \\ u - u_n \end{pmatrix} dSdt
\end{aligned}
$$

(3.21)

subject to

$$
\begin{aligned}
y_t + A\, y + d^n + d_y^n\, (y - y_n) + d_v^n\, (v - v_n) &= 0, \\
\partial_\nu y + b^n + b_y^n\, (y - y_n) + b_u^n\, (u - u_n) &= 0, \\
y(0) &= y_0
\end{aligned}
\tag{3.22}
$$

and to

$$
v \in V_{ad}, u \in U_{ad}. \tag{3.23}
$$

In this setting, the notation $\varphi_y^n = \varphi_y(x, y_n(x,T))$, $\varphi_{yy}^n = \varphi_{yy}^n(x, y_n(x,T))$, $f_y^n = f_y^n(x,t,y_n(x,t),v_n(x,t))$, $D^2 H^{Q,n} = D^2 H_{(y,v,u)}(x,t,y_n(x,t),v_n(x,t),p_n(x,t))$ etc., was used. The associated adjoint state p_{n+1} is determined from

$$
\begin{aligned}
-p_t + A\, p + d_y^n\, (p - p_n) &= H_y^{Q,n} + H_{yy}^{Q,n}(y_{n+1} - y_n) + H_{yv}^{Q,n}(v_{n+1} - v_n), \\
p(T) &= \varphi_y^n + \varphi_{yy}^n\ (y_{n+1} - y_n)(T), \\
\partial_\nu p + b_y^n\, (p - p_n) &= H_y^{\Sigma,n} + H_{yy}^{\Sigma,n}(y_{n+1} - y_n) + H_{yu}^{\Sigma,n}(u_{n+1} - u_n).
\end{aligned}
\tag{3.24}
$$

In this way, a sequence of quadratic optimization problems is to be solved, giving the method the name sequential quadratic programming (SQP) method. The main aim of this paper is to show that this process exhibits a local quadratic convergence. We shall transform the optimality system into a *generalized equation*. Then we are able to interpret the SQP method as a Newton method for a generalized equation. This approach gives direct access to known results on the convergence of Newton methods. In the analysis, a specific difficulty arises from the fact that (QP_n) might be non-convex. It therefore may have multiple local minima. We shall have to restrict the control set to a sufficiently small neighborhood around the reference solution.

4. Generalized equation and Newton method. To transform the optimality system into a generalized equation, we reformulate the variational inequalities (3.8)–(3.9) as generalized equations, too.

Therefore, we define the *cones*

$$N^Q(v) = \begin{cases} \{z \in L^\infty(Q) \mid \int_Q z(\tilde{v} - v)dxdt \le 0 \quad \forall \tilde{v} \in V_{ad}\} \text{ if } v \in V_{ad}, \\ \emptyset \text{ if } v \notin V_{ad}, \end{cases} \tag{4.1}$$

$$N^\Sigma(u) = \begin{cases} \{z \in L^\infty(\Sigma) \mid \int_\Sigma z(\tilde{u} - u)dSdt \le 0 \quad \forall \tilde{u} \in U_{ad}\} \text{ if } u \in U_{ad}, \\ \emptyset \text{ if } u \notin U_{ad}. \end{cases} \tag{4.2}$$

Then (3.8) and (3.9) read $-H_v^Q(y,p,v) \in N^Q(v)$, $-H_u^\Sigma(y,p,u) \in N^\Sigma(u)$, or

$$0 \in H_v^Q(y,p,v) + N^Q(v), \tag{4.3}$$

$$0 \in H_u^\Sigma(y,p,u) + N^\Sigma(u). \tag{4.4}$$

(H_v^Q and H_u^Σ are Nemytskii operators defined analogously to H_y^Q, H_y^Σ.) The set-valued mappings $T_1 : v \mapsto N^Q(v)$ from $L^\infty(Q)$ to $2^{L^\infty(Q)}$ and $T_2 : u \mapsto N^\Sigma(u)$ from $L^\infty(\Sigma)$ to $2^{L^\infty(\Sigma)}$ have closed graph. We introduce now the subspace E of $\tilde{E} = L^\infty(Q) \times L^\infty(\Sigma) \times C(\overline{\Omega}))^2 \times L^\infty(Q) \times L^\infty(\Sigma)$, which contains all $\eta \in \tilde{E}$ of the form $\eta = (e_Q, e_\Sigma, 0, \gamma_Q, \gamma_\Sigma, \gamma_\Omega, \gamma_v, \gamma_u)$ and is endowed with the norm $\|\eta\|_E = \|e_Q\|_{L^\infty(Q)} + \|e_\Sigma\|_{L^\infty(\Sigma)} + \|\gamma_Q\|_{L^\infty(Q)} + \|\gamma_\Sigma\|_{L^\infty(\Sigma)} + \|\gamma_\Omega\|_{C(\overline{\Omega})} + \|\gamma_v\|_{L^\infty(Q)} + \|\gamma_u\|_{L^\infty(\Sigma)}$. We also need the space $W = Y \times Y \times L^\infty(Q) \times L^\infty(\Sigma)$ equipped with the norm $\|(y,p,v,u)\|_W = \|y\|_Y + \|p\|_Y + \|v\|_{L^\infty(\Omega)} + \|u\|_{L^\infty(\Sigma)}$. Moreover, define the set-valued mapping $T : W \to 2^E$ by

$$T(w) = (\{0\}, \{0\}, \{0\}, \{0\}, \{0\}, \{0\}, N^Q(v)N^\Sigma(u))$$

and $F : W \to E$ by $F(w) = (F_1(w), ..., F_8(w))$, where

$$\begin{aligned} F_1(w) &= y_t + A\,y + d(y,v), \\ F_2(w) &= \partial_\nu y + b(y,u), \\ F_3(w) &= y(0) - y_0, \\ F_4(w) &= -p_t + A\,p - H_y^Q(y,p,v), \\ F_5(w) &= \partial_\nu p - H_y^\Sigma(y,p,u), \\ F_6(w) &= p(T) - \varphi_y(y(T)), \\ F_7(w) &= H_v^Q(y,p,v), \\ F_8(w) &= H_u^\Sigma(y,p,u). \end{aligned}$$

In the definition of E, the third component is vanishing, since it will correspond to the initial condition $y(0) - y_0 = 0$, which is kept fixed in the generalized Newton method. The optimality system is clearly equivalent to the generalized equation

$$0 \in F(w) + T(w), \tag{4.5}$$

where F is of class $C^{1,1}$, and the set-valued mapping T has closed graph. Obviously, the reference solution $\overline{w} = (\overline{y}, \overline{p}, \overline{v}, \overline{u})$ satisfies (4.5). The *generalized Newton method* for solving (4.5) is similar to the Newton method for equations in Banach spaces.

Suppose that we have already computed $w_1, \dots, w_n$. Then w_{n+1} is to be determined by the generalized equation

$$0 \in F(w_n) + F'(w_n)(w - w_n) + T(w). \tag{4.6}$$

The convergence analysis of this method is closely related to the notion of *strong regularity* of (4.5). See Robinson [29]. The generalized equation (4.5) is said to be *strongly regular at* $\overline{w}$ if there are constants $r_1 > 0, r_2 > 0$, and $c_L > 0$ such that for all perturbations $e \in \mathcal{B}_{r_1}(0_E)$ the *linearized equation*

$$e \in F(\overline{w}) + F'(\overline{w})(w - \overline{w}) + T(w) \tag{4.7}$$

has in $\mathcal{B}_{r_2}(\overline{w})$ a unique solution $w = w(e)$, and the Lipschitz property

$$\|w(e_1) - w(e_2)\|_W \le c_L \|e_1 - e_2\|_E \tag{4.8}$$

holds $\forall\, e_1, e_2 \in \mathcal{B}_{r_1}(0_E)$. In the case of an equation $F(w) = 0$, we have $F(\overline{w}) = 0, T(w) = \{0\}$, and strong regularity means the existence and boundedness of $(F'(\overline{w}))^{-1}$. For other aspects of L-stability we refer to [9]. The following result gives a first answer to the convergence analysis of the generalized Newton method.

THEOREM 4.1. *Suppose that* (4.5) *is strongly regular at* $\overline{w}$. *Then there are* $r_{\mathcal{N}} > 0$ *and* $c_{\mathcal{N}} > 0$ *such that for each starting element* $w_1 \in \mathcal{B}_{r_{\mathcal{N}}}(\overline{w})$ *the generalized Newton method generates a unique sequence* $\{w_n\}_{n=1}^{\infty}$. *This sequence remains in* $\mathcal{B}_{r_{\mathcal{N}}}(\overline{w})$, *and it holds that*

$$\|w_{n+1} - \overline{w}\|_W \le c_{\mathcal{N}} \|w_n - \overline{w}\|_W^2 \qquad \forall n \in \mathbb{N}. \tag{4.9}$$

This result was apparently shown first by Josephy [18]. Generalizations can be found in Dontchev [8] and Alt [1], [2]. We refer in particular to the recent publication by Alt [3], where a mesh-independence principle was shown for numerical approximation of (4.5). We shall verify that the second order condition (SSC) implies strong regularity of the generalized equation at $\overline{w} = (\overline{y}, \overline{p}, \overline{v}, \overline{u})$ in certain subsets $\widehat{V}_{ad} \subset V_{ad}$, $\widehat{U}_{ad} \subset U_{ad}$. Then Theorem 4.1 yields the quadratic convergence of the generalized Newton method in these subsets.

5. Strong regularity. To investigate the strong regularity of the generalized equation (4.5) at $\bar{w}$, we have to consider the perturbed generalized equation (4.7). Once again, we are able to interpret this equation as the optimality system of a linear-quadratic control problem. This problem is not necessarily convex, therefore we study the behavior of the following auxiliary linear-quadratic problem associated with the perturbation e:

($\widehat{\mathrm{QP}_e}$) Minimize

$$\begin{aligned} J_e(y,v,u) = {} & \int_\Omega (\bar\varphi_y + \gamma_\Omega)\, y(T)\, dx + \int_Q (\bar f_y + \gamma_Q)\, y\, dxdt + \int_Q (\bar f_v + \gamma_v)\, v\, dxdt \\ & + \int_\Sigma (\bar g_y + \gamma_\Sigma)\, y\, dSdt + \int_\Sigma (\bar g_u + \gamma_u)\, u\, dSdt + \frac{1}{2}\int_\Omega \bar\varphi_{yy}(y(T) - \bar y(T))^2 dx \\ & + \tfrac{1}{2}\int_Q \begin{pmatrix} y - \bar y \\ v - \bar v \end{pmatrix}^\top D^2 \bar H^Q \begin{pmatrix} y - \bar y \\ v - \bar v \end{pmatrix} dxdt + \tfrac{1}{2}\int_\Sigma \begin{pmatrix} y - \bar y \\ u - \bar u \end{pmatrix}^\top D^2 \bar H^\Sigma \begin{pmatrix} y - \bar y \\ u - \bar u \end{pmatrix} dSdt \end{aligned} \tag{5.1}$$

subject to

$$\begin{aligned} y_t + A\,y + d(\bar y, \bar v) + \bar d_y\,(y - \bar y) + \bar d_v\,(v - \bar v) &= e_Q && \text{in } Q, \\ \partial_\nu y + b(\bar y, \bar u) + \bar b_y\,(y - \bar y) + \bar b_u\,(u - \bar u) &= e_\Sigma && \text{on } \Sigma, \\ y(0) &= y_0 && \text{in } \Omega \end{aligned} \tag{5.2}$$

and to the constraints on the control

$$\begin{array}{rcl} v \in \widehat{V}_{ad} & = & \{v \in V_{ad} \,|\, v(x,t) = \bar{v}(x,t) \text{ on } Q(\sigma)\}, \\ u \in \widehat{U}_{ad} & = & \{u \in U_{ad} \,|\, u(x,t) = \bar{u}(x,t) \text{ on } \Sigma(\sigma)\}. \end{array} \tag{5.3}$$

In this setting, the perturbation vector $e = (e_Q, e_\Sigma, 0, \gamma_Q, \gamma_\Sigma, \gamma_\Omega, \gamma_v, \gamma_u)$ belongs to E. The hat in $(\widehat{\mathrm{QP}}_e)$ indicates that v and u are taken equal to $\bar{v}$ and $\bar{u}$ on the strongly active sets $Q(\sigma)$ and $\Sigma(\sigma)$, respectively.

Remark. The generalized equation (4.7) is equivalent to the optimality system of the problem (QP_e) obtained from $(\widehat{\mathrm{QP}}_e)$ on substituting V_{ad} for $\widehat{V}_{ad}$ and U_{ad} for $\widehat{U}_{ad}$, respectively.

In the space of perturbations E we need another norm

$$\begin{array}{rcl} \|e\|_2 & = & \|e_Q\|_{L^2(Q)} + \|e_\Sigma\|_{L^2(\Sigma)} + \|\gamma_Q\|_{L^2(Q)} + \|\gamma_\Sigma\|_{L^2(\Sigma)} \\ & & + \|\gamma_\Omega\|_{L^2(\Omega)} + \|\gamma_v\|_{L^2(Q)} + \|\gamma_u\|_{L^2(\Sigma)}. \end{array}$$

Moreover, in W we shall also use the norm

$$\|(y,p,v,u)\|_2 = \|y\|_{W(0,T)} + \|p\|_{W(0,T)} + \|v\|_{L^2(Q)} + \|u\|_{L^2(\Sigma)}.$$

The next results follow from the author's paper [33].

LEMMA 5.1. *Suppose that the second order sufficient optimality condition* (SSC) *is satisfied at* $(\bar{y}, \bar{v}, \bar{u})$ *with associated adjoint state* $\bar{p}$. *Then for each* $e \in E$, *the problem* $(\widehat{\mathrm{QP}}_e)$ *has a unique solution* (y_e, v_e, u_e) *with associated adjoint state* p_e. *Let* (y_i, v_i, u_i) *and* $p_i, i = 1, 2$, *be the solutions and adjoint states to* $e_i \in E, i = 1, 2$. *There is a constant* $l_2 > 0$, *not depending on* e_i, *such that*

$$\|(y_1,p_1,v_1,u_1) - (y_2,p_2,v_2,u_2)\|_2 \le l_2 \|e_1 - e_2\|_2 \tag{5.4}$$

holds $\forall\, e_i \in E, i = 1, 2$.

By continuity, (5.4) extends to perturbations e_i of L^2. It was shown in [33] that the second order condition (SSC) implies the following strong *Legendre–Clebsch* condition:

$$\begin{array}{lll} H^Q_{vv}(x,t,\bar{y}(x,t),\bar{p}(x,t),\bar{v}(x,t)) \ge \delta & a.e. \text{ on } & Q, \\ H^\Sigma_{uu}(x,t,\bar{y}(x,t),\bar{p}(x,t),\bar{u}(x,t)) \ge \delta & a.e. \text{ on } & \Sigma. \end{array} \tag{LC}$$

THEOREM 5.2. *Let the assumptions of Lemma* 5.1 *be satisfied. Then there is a constant* $l_\infty > 0$, *not depending on* e_i, *such that*

$$\|(y_1,p_1,v_1,u_1) - (y_2,p_2,v_2,u_2)\|_W \le l_\infty \|e_1 - e_2\|_E \tag{5.5}$$

holds for (y_i, v_i, u_i, p_i), *and* $e_i, i = 1, 2$, *introduced in Lemma* 5.1.

This theorem follows from [33, Thm. 5.2]. (Notice that $v_i = \bar{v}$ and $u_i = \bar{u}$ on $Q(\sigma)$ and $\Sigma(\sigma)$, respectively. This can be expressed by taking $u_a := u_b := \bar{u}$ and $v_a := v_b := \bar{v}$ on these sets. Then [33, Thm. 5.2] is easy to apply.)

Unfortunately, (5.5) holds only for $\widehat{V}_{ad}$ and $\widehat{U}_{ad}$. We are not able to prove (5.5) in V_{ad}, U_{ad}. In this case, J_e might be nonconvex and (QP_e) may have multiple solutions, if solvable at all. However, formulating Theorem 5.2 in the context of our generalized equation, we already have obtained the following result on strong regularity.

THEOREM 5.3. *Suppose that* $\bar{w} = (\bar{y}, \bar{p}, \bar{v}, \bar{u})$ *satisfies the first order optimality system* (2.1), (3.2)–(3.3), (3.7)–(3.9) *together with the second order sufficient condition*

(SSC). *Then the generalized equation* (4.5) *is strongly regular at* $\bar{w}$*, provided that the control sets* $\widehat{V}_{ad}, \widehat{U}_{ad}$ *are substituted for* V_{ad}, U_{ad} *in the definition of* $T(w)$.

Remark. The last assumption means that the cones $N^Q(v)$ and $N^\Sigma(u)$ are defined on using $\widehat{V}_{ad}$ and $\widehat{U}_{ad}$, respectively.

To complete the discussion of the Newton method, the following questions have yet to be answered: How can we solve the generalized equation (4.6) in $\widehat{V}_{ad}$ and $\widehat{U}_{ad}$, and how do we get rid of the artificial restriction $v = \bar{v}$ on $Q(\sigma)$ and $u = \bar{u}$ on $\Sigma(\sigma)$?

We shall show that the SQP method, restricted to a sufficiently small neighborhood around $\bar{v}$ and $\bar{u}$, will solve both the problems: If the region is small enough, then the SQP method delivers a unique solution $w_n = (y_n, p_n, v_n, u_n)$, where $v_n = \bar{v}, u_n = \bar{u}$ is automatically satisfied on $Q(\sigma), \Sigma(\sigma)$. Moreover, this w_n is a solution of the generalized equation (4.5), that is, a solution of the optimality system for (P).

6. The linear-quadratic subproblems (QP_n). The presentation of the SQP method is still quite formal. We do not know whether the quadratic subproblem (QP_n) defined by (3.21)–(3.23) is solvable at all. Moreover, if solutions exist, we are not able to show their uniqueness. There might exist multiple stationary solutions, i.e., solutions satisfying the optimality system for (QP_n). Notice that the objective J_n of (QP_n) is only convex on a subspace. Owing to this, we have to restrict (QP_n) to a sufficiently small neighborhood around the reference solution $(\bar{v}, \bar{u})$. This region is defined by

$$\begin{aligned} V_{ad}^\varrho &= \{v \in V_{ad} \,|\, \|v - \bar{v}\|_{L^\infty(Q)} \le \varrho\}, \\ U_{ad}^\varrho &= \{u \in U_{ad} \,|\, \|u - \bar{u}\|_{L^\infty(\Sigma)} \le \varrho\}, \end{aligned}$$

where $\varrho > 0$ is a sufficiently small radius. To avoid the unknown reference solution $(\bar{v}, \bar{u})$ in the definition of the neighborhood, we shall later replace this neighborhood by a ball around the initial iterate (v_1, u_1) .

Let us denote by (QP_n^ϱ) the problem (QP_n) restricted to V_{ad}^ϱ, U_{ad}^ϱ and by ($\widehat{\mathrm{QP}}_n$) the same problem restricted to $\widehat{V}_{ad}$, $\widehat{U}_{ad}$, respectively. To analyze ($\widehat{\mathrm{QP}}_n$) in a first step, we need some auxiliary results.

LEMMA 6.1. *For all* $K > 0$ *there is a constant* $c_L = c_L(K)$ *such that*

$$\mathcal{E} \le c_L(K) \|w_n - \bar{w}\|_W \tag{6.1}$$

holds $\forall\, w_n \in W$ *with* $\|w_n - \bar{w}\|_W \le K$*, where the expression* $\mathcal{E}$ *is defined by*

$$\begin{aligned} \mathcal{E} = \max \{ &\|f_v^n - \bar{f}_v\|_{L^\infty(Q)}, \|f_y^n - \bar{f}_y\|_{L^\infty(Q)}, \|g_v^n - \bar{g}_u\|_{L^\infty(\Sigma)}, \|g_y^n - \bar{g}_y\|_{L^\infty(\Sigma)}, \\ &\| d_y^n - \bar{d}_y\|_{L^\infty(Q)}, \|d_v^n - \bar{d}_v\|_{L^\infty(Q)}, \|b_y^n - \bar{b}_y\|_{L^\infty(\Sigma)}, \|b_u^n - \bar{b}_u\|_{L^\infty(\Sigma)}, \|\varphi_y^n - \bar{\varphi}_y\|_{C(\bar{\Omega})}, \\ &\| \varphi_{yy}^n - \bar{\varphi}_{yy}\|_{C(\bar{\Omega})}, \|D^2 H^{Q,n} - D^2 \bar{H}^Q\|_{L^\infty(Q)}, \|D^2 H^{\Sigma,n} - D^2 \bar{H}^\Sigma\|_{L^\infty(Q)} \}. \end{aligned}$$

Proof. The estimate follows from the assumptions (A2)–(A4) imposed on the functions f, g, φ, b, d in sections 2 and 3. For instance, the mean value theorem yields

$$\begin{aligned} \|f_v^n - \bar{f}_v\|_{L^\infty(Q)} &= \sup_{(x,t)\in Q} \operatorname{ess} |f_{vy}(y^\vartheta, v^\vartheta)(y_n - \bar{y}) + f_{vv}(v^\vartheta, v^\vartheta)(v_n - \bar{v})| \\ &\le c(K) \sup_{(x,t)\in Q} \operatorname{ess} (|y_n - \bar{y}| + |v_n - \bar{v}|) \end{aligned}$$

by (3.5), where $y^\vartheta = \bar{y} + \vartheta(y_n - \bar{y})$, $v^\vartheta = \bar{v} + \vartheta(v_n - \bar{v})$, and $\vartheta = \vartheta(x,t)$ belongs to $(0,1)$. (Consider, for example, the estimation

$$\begin{aligned} |f_{vy}(y^\vartheta, v^\vartheta)| &\le |f_{yv}(0,0)| + |f_{vy}(y^\vartheta, v^\vartheta) - f_{vy}(0,0)| \le c_1 + c(K)\,(|y^\vartheta| + |v^\vartheta|) \\ &\le c_1 + c(K) \cdot K, \end{aligned}$$

which follows from (3.5).) The other terms in $\mathcal{E}$ are handled analogously. ∎

We shall denote the quadratic part of the functional J_n by

$$\begin{aligned} B_n[(y_1,v_1,u_1),(y_2,v_2,u_2)] \;=\; & \int_\Omega \varphi^n_{yy}\; y_1(T)y_2(T)\,dx + \int_Q (y_1,v_1)D^2H^{Q,n}(y_2,v_2)^\top\,dxdt \\ & + \int_\Sigma (y_1,u_1)D^2H^{\Sigma,n}(y_2,u_2)^\top\,dSdt \end{aligned} \tag{6.2}$$

and write for short $B_n[(y,v,u),(y,v,u)] = B_n[y,v,u]^2$.

LEMMA 6.2. *Suppose that the second order sufficient optimality condition* (SSC) *is satisfied. Then there is $\varrho_1 > 0$ with the following property: If $\|w_n - \bar w\|_W \le \varrho_1$, then*

$$B_n[y,v,u]^2 \ge \frac{\delta}{2}\|(y,v,u)\|^2_H \tag{6.3}$$

holds $\forall\,(y,v,u) \in H$ satisfying $v = 0$ on $Q(\sigma)$, $u = 0$ on $I^u(\sigma)$ together with

$$\begin{aligned} y_t + A\,y + d^n_y\; y + d^n_v\; v &= 0, \\ \partial_\nu y + b^n_y\; y + b^n_u\; u &= 0, \\ y(0) &= 0. \end{aligned} \tag{6.4}$$

Proof. Let z denote the weak solution of the parabolic equation obtained from (6.4) on substituting $\bar d_y, \bar d_v, \bar b_y, \bar b_u$ for d^n_y, d^n_v, b^n_y, b^n_u, respectively. Then

$$\begin{aligned} (y-z)_t + A\,(y-z) + \bar d_y\,(y-z) &= (\bar d_y - d^n_y)\,y + (\bar d_v - d^n_v)\,v, \\ \partial_\nu(y-z) + \bar b_y\,(y-z) &= (\bar b_y - b^n_y)\,y + (\bar b_u - b^n_u)\,u, \\ (y-z)(0) &= 0. \end{aligned}$$

We have $\bar d_y \ge c_0, \bar b_y \ge c_0$. The differences on the right-hand sides can be estimated by Lemma 6.1, where $K = \|\bar w\|_W + \varrho_1$; hence parabolic L^2-regularity yields

$$\begin{aligned} \|y-z\|_{W(0,T)} &\le c\,(\|\bar d_y - d^n_y\|_{L^\infty(Q)}\|y\|_{L^2(Q)} + \|\bar d_y - d^n_v\|_{L^\infty(Q)}\|v\|_{L^2(Q)} \\ &\quad + \|\bar b_y - b^n_y\|_{L^\infty(\Sigma)}\|y\|_{L^2(\Sigma)} + \|\bar b_u - b^n_u\|\|u\|_{L^2(\Sigma)}) \\ &\le c\,\varrho_1(\|y\|_{W(0,T)} + \|v\|_{L^2(Q)} + \|u\|_{L^2(\Sigma)}) \le c\,\varrho_1\|(y,v,u)\|_H. \end{aligned} \tag{6.5}$$

Substituting $y = z + (y-z)$ in B_n,

$$\begin{aligned} B_n[y,v,u]^2 &= B_n[z+(y-z),v,u]^2 \\ &= B[z,v,u]^2 + (B_n - B)[z,v,u]^2 + 2B_n[(z,v,u),(y-z,0,0)] \\ &\quad + B_n[y-z,0,0]^2 \end{aligned}$$

is obtained. (SSC) applies to the first expression B, while the second is estimated by Lemma 6.1. In the remaining two parts, we use the uniform boundedness of all coefficients. Therefore, by (6.5)

$$\begin{aligned} B_n[y,v,u]^2 &\ge \delta\,\|(z,v,u)\|^2_H - c\,\varrho_1\|(z,v,u)\|^2_H - c\,\|(z,v,u)\|_H\|y-z\|_{W(0,T)} \\ &\quad - c\,\|y-z\|^2_{W(0,T)} \\ &\ge \frac{3}{4}\delta\,\|(z,v,u)\|^2_H - c\varrho_1\|(z,v,u)\|_H\|(y,v,u)\|_H - c\varrho_1^2\|(y,v,u)\|^2_H \end{aligned}$$

if ϱ_1 is sufficiently small. Next we resubstitute $z = y + (z - y)$ and apply (6.5) again. In this way, the desired estimate (6.3) is easily verified for sufficiently small $\varrho_1 > 0$. □

COROLLARY 6.3. *If $\|w_n - \bar{w}\|_W \le \varrho_1$ and* (SSC) *is satisfied at $\bar{w}$, then $(\widehat{\mathrm{QP}}_n)$ has a unique optimal pair of controls $(\hat{v}, \hat{u})$ with associated state $\hat{y}$.*

Proof. The functional J_n to be minimized in $(\widehat{\mathrm{QP}}_n)$ has the form (see (3.21))

$$J_n(y, v, u) = a_n(y, v, u) + \frac{1}{2} B_n[y - y_n, v - v_n, u - u_n]^2,$$

where a_n is a linear integral functional. J_n is uniformly convex on the feasible region of $(\widehat{QP}_n)$. By Lemma 6.2, the sets $\widehat{V}_{ad}$ and $\widehat{U}_{ad}$ are weakly compact in $L^2(Q)$ and $L^2(\Sigma)$, respectively. Therefore, the corollary follows from standard arguments. □

Let us return to the discussion of the relation between Newton method and SQP method. In what follows, we shall denote by $\hat{w}_n = (\hat{y}_n, \hat{p}_n, \hat{v}_n, \hat{u}_n)$ the sequence of iterates generated by the SQP method performed in $\widehat{V}_{ad}$, $\widehat{U}_{ad}$ (provided that this sequence is well defined). The iterates of the generalized Newton method are denoted by w_n. Consider now both methods initiating from the same element $w_n = \hat{w}_n$.

If $\|w_n - \bar{w}\|_W \le \varrho_1$, then Corollary 6.3 shows the existence of a unique solution $(\hat{y}_{n+1}, \hat{v}_{n+1}, \hat{u}_{n+1})$ of $(\widehat{\mathrm{QP}}_n)$ having the associated adjoint state $\hat{p}_{n+1}$. The element $\hat{w}_{n+1}$ solves the optimality system corresponding to $(\widehat{\mathrm{QP}}_n)$. By convexity (Lemma 6.2), any other solution of this system solves $(\widehat{\mathrm{QP}}_n)$; hence it is equal to $\hat{w}_{n+1}$. On the other hand, the optimality system is equivalent to the generalized equation (4.6) at w_n (based on the sets $\widehat{V}_{ad}, \widehat{U}_{ad}$). For $\|w_n - \bar{w}\|_W \le r_{\mathcal{N}}$, one step of the generalized Newton method delivers the unique solution w_{n+1} of (4.6). As w_{n+1} solves the optimality system for $(\widehat{\mathrm{QP}}_n)$, it has to coincide with $\hat{w}_{n+1}$. Suppose further that $\|w_n - \bar{w}\|_W \le \min\{r_{\mathcal{N}}, \varrho_1\}$. Then Theorem 4.1 implies that $w_{n+1} = \hat{w}_{n+1}$ remains in $\mathcal{B}_{min\{r_{\mathcal{N}}, \varrho_1\}}(\overline{w})$, so that $\|\hat{w}_{n+1} - \bar{w}\|_W \le \min\{r_{\mathcal{N}}, \varrho_1\}$. Consequently, we are able to perform the next step in both the methods. Moreover, in $\widehat{V}_{ad}, \widehat{U}_{ad}$ each step of the Newton method is equivalent to solving $(\widehat{\mathrm{QP}}_n)$, which always has a unique solution. In other words, the Newton method and the SQP method are identical in $\widehat{V}_{ad}, \widehat{U}_{ad}$.

THEOREM 6.4. *Let $\bar{w} = (\bar{y}, \bar{p}, \bar{v}, \bar{u})$ satisfy the first order optimality system* (2.1), (3.2)–(3.3), (3.7)–(3.9) *together with the second order sufficient optimality conditions* (SSC). *Suppose that $w_1 = (y_1, p_1, v_1, u_1) \in W$ is given such that $\|w_1 - \bar{w}\|_W \le min\{\varrho_1, r_{\mathcal{N}}\}$, $v_1 \in \widehat{V}_{ad}$, and $u_1 \in \widehat{U}_{ad}$. Then in $\widehat{V}_{ad}, \widehat{U}_{ad}$, the generalized Newton method is equivalent to the SQP method: The solution of the generalized equation* (4.6) *is given by the unique solution of $(\widehat{\mathrm{QP}}_n)$ along with the associated adjoint state.*

The result follows from Theorem 5.3 (strong regularity) and the considerations above.

Remark. It is easy to verify that $\hat{w}_n$, the solution of $(\widehat{\mathrm{QP}}_n)$, obeys the optimality system for (P) in the original sets V_{ad}, U_{ad} (cf. also Corollary 6.9).

Next, we discuss the optimality system for $(\widehat{\mathrm{QP}}_n)$ and $(\mathrm{QP}_n^{\varrho})$. Let us denote the associated Hamilton functions by $\tilde{H}$ to distinguish them from H, which belongs to (P):

$$\begin{aligned}
\tilde{H}^Q(x,t,y,p,v) &= f_y^n(y - y_n) + f_v^n(v - v_n) - p\,(d^n + d_y^n(y - y_n) + d_v^n(v - v_n)) \\
&\quad + \tfrac{1}{2}(y - y_n, v - v_n) D^2 H^{Q,n}(y - y_n, v - v_n)^\top, \\
\tilde{H}^\Sigma(x,t,y,p,u) &= g_y^n(y - y_n) + g_u^n(u - u_n) - p\,(b^n + b_y^n(y - y_n) + b_u^n(u - u_n)) \\
&\quad + \tfrac{1}{2}(y - y_n, u - u_n) D^2 H^{\Sigma,n}(y - y_n, u - u_n)^\top,
\end{aligned}$$

where y, v, p, u are real numbers and (x,t) appears in the quantities depending on n. Notice that these Hamiltonians coincide for $(\widehat{\mathrm{QP}}_n), (\mathrm{QP}_n^\varrho)$, and (QP_n), since these problems differ only in the underlying sets of admissible controls. We consider the problems defined at $w_n = (y_n, p_n, v_n, u_n)$. In what follows, we denote solutions of the optimality system corresponding to (QP_n^ϱ) by (y^+, v^+, u^+). The optimality system for (QP_n^ϱ) consists of

$$\int_Q \tilde{H}_v^Q(y^+, p^+, v^+)(v - v^+)\, dxdt \geq 0 \quad \forall v \in V_{ad}^\varrho, \tag{6.6}$$

$$\int_\Sigma \tilde{H}_u^\Sigma(y^+, p^+, u^+)(u - u^+)\, dSdt \geq 0 \quad \forall u \in U_{ad}^\varrho, \tag{6.7}$$

where the associated adjoint state p^+ is defined by

$$\begin{aligned} -p_t^+ + A\,p^+ &= \tilde{H}_y^Q = f_y^n + H_{yy}^{Q,n}(y^+ - y_n) + H_{yv}^{Q,n}(v^+ - v_n) - d_y^n p^+, \\ p(T) &= \varphi_y^n + \varphi_{yy}^n(y^+(T) - y(T)), \\ \partial_\nu p &= \tilde{H}_y^\Sigma = g_y^n + H_{yy}^{\Sigma,n}(y^+ - y_n) + H_{yv}^{\Sigma,n}(u^+ - u_n) - b_y^n p^+. \end{aligned} \tag{6.8}$$

The state equation (3.22) for y^+ and the constraints $v^+ \in V_{ad}^\varrho, u^+ \in U_{ad}^\varrho$ are included in the optimality system too. The optimality system of $(\widehat{\mathrm{QP}}_n)$ has the same principal form as (6.6)–(6.8) and is obtained on replacing (y^+, p^+, v^+, u^+) by $(\hat{y}_{n+1}, \hat{p}_{n+1}, \hat{v}_{n+1}, \hat{u}_{n+1})$. Moreover, $\widehat{V}_{ad}$, $\widehat{U}_{ad}$ is to be substituted for V_{ad}^ϱ, U_{ad}^ϱ there.

In the further analysis, we shall perform the following steps. First we prove by a sequence of results that the solution $(\hat{v}_n, \hat{u}_n)$ of $(\widehat{\mathrm{QP}}_n)$ satisfies the optimality system of (QP_n^ϱ) for sufficiently small ϱ. Moreover, we prove that (QP_n^ϱ) has at least one optimal pair, if w_n is sufficiently close to $\bar{w}$. Finally, relying on (SSC), we verify uniqueness for the optimality system of (QP_n^ϱ). Therefore, $(\hat{v}_n, \hat{u}_n)$ can be obtained as the unique global solution of (QP_n^ϱ). Notice that (QP_n^ϱ) might be nonconvex; hence the optimality of $(\hat{v}_n, \hat{u}_n)$ does not follow directly from fulfilling the optimality system.

LEMMA 6.5. *There is $\varrho_2 > 0$ with the following property: If $\varrho \leq \varrho_2$, $w_n \in W$ fulfills $\|w_n - \bar{w}\|_W \leq \varrho_2$, and (y^+, v^+, u^+) satisfies the constraints of (QP_n^ϱ) with associated adjoint state p^+, then*

$$\textit{sign}\ \tilde{H}_v^Q(y^+, p^+, v^+)(x,t) = \textit{sign}\ H_v^Q(\bar{y}, \bar{p}, \bar{v})(x,t) \quad \textit{a.e. on}\ \ Q(\sigma), \tag{6.9}$$

$$\textit{sign}\ \tilde{H}_u^\Sigma(y^+, p^+, u^+)(x,t) = \textit{sign}\ H_u^\Sigma(\bar{y}, \bar{p}, \bar{u})(x,t) \quad \textit{a.e. on}\ \ \Sigma(\sigma), \tag{6.10}$$

$$|\tilde{H}_v^Q(y^+, p^+, v^+)(x,t)| \geq \frac{\sigma}{2} \quad \textit{a.e. on}\ \ Q(\sigma), \tag{6.11}$$

$$|\tilde{H}_u^\Sigma(y^+, p^+, u^+)(x,t)| \geq \frac{\sigma}{2} \quad \textit{a.e. on}\ \ \Sigma(\sigma). \tag{6.12}$$

Proof. Let us discuss $\tilde{H}_v^Q$; the proof is analogous for $\tilde{H}_u^\Sigma$. We have

$$\begin{aligned} \tilde{H}_v^Q &= f_v^n + H_{yv}^{Q,n}(y^+ - y_n) + H_{vv}^{Q,n}(v^+ - v_n) - p^+ d_v^n \\ &= \bar{f}_v - \bar{p}\bar{d}_v + \{f_v^n - \bar{f}_v + (f_{yv}^n - p_n d_{yv}^n)(y^+ - y_n) \\ &\quad + (f_{vv}^n - p_n d_{vv}^n)(v^+ - v_n) + (\bar{p}\bar{d}_v - p^+ d_v^n)\} = \bar{H}_v^Q + \{\ldots\} \geq \sigma - |\{\ldots\}| \end{aligned}$$

a.e. on $Q(\sigma)$. Lemma 6.1 applies to estimate $|\{\ldots\}| \leq c \cdot \varrho_2$, where c does not depend on w_n, y^+, p^+, u^+, v^+, provided that we are able to prove that $\|p^+ - \bar{p}\|_{C(\bar{Q})} \leq c\ \varrho_2$ and

$\|y^+ - \bar{y}\|_{C(\bar{Q})} \le c\ \varrho_2$ holds with an associated constant c. Let us sketch the estimation of $y^+ - \bar{y} =: y$. This function satisfies

$$\begin{aligned} y_t + A\,y + d_y^n y &= -d_v^n(v^+ - \bar{v}) + (d_y^n - d_y^\vartheta)(y_n - \bar{y}) + (d_v^n - d_v^\vartheta)(v_n - \bar{v}) \\ \partial_\nu y + b_y^n y &= -d_u^n(u^+ - \bar{u}) + (b_y^n - b_y^\vartheta)(y_n - \bar{y}) + (b_u^n - b_u^\vartheta)(u_n - \bar{u}) \\ y(0) &= 0, \end{aligned}$$

where $d_y^\vartheta = d_y(\bar{y} + \vartheta(y_n - \bar{y}), \bar{v} + \vartheta(v_n - \bar{v}))$, $\vartheta = \vartheta(x,t) \in (0,1)$, and the other quantities are defined accordingly. We have $\max\{\|v^+ - \bar{v}\|_{L^\infty(Q)}, \|u^+ - \bar{u}\|_{L^\infty(\Sigma)}\} \le \varrho, \max\{\|y_n - \bar{y}\|_{C(\bar{Q})}, \|u_n - \bar{u}\|_{L^\infty(\Sigma)}, \|v_n - \bar{v}\|_{L^\infty(Q)}\} \le \varrho_2$. Thus the right-hand sides of the PDE and its boundary condition are estimated by $c \cdot \varrho_2$. The estimate for $\|y^+ - \bar{y}\|$ follows from Theorem 2.2. The difference $p^+ - \bar{p}$ is handled in the same way. □

COROLLARY 6.6. *If* $\max\{\|w_n - \bar{w}\|_W, \varrho\} \le \varrho_2$, *then the relations*

$$\begin{aligned} v^+(x,t) &= \bar{v}(x,t) \quad \text{a.e. on } Q(\sigma), \\ u^+(x,t) &= \bar{u}(x,t) \quad \text{a.e. on } \Sigma(\sigma) \end{aligned}$$

hold for all controls (v^+, u^+) *of* (QP_n^ϱ), *satisfying together with the associated state* y^+ *and the adjoint state* p^+ *the optimality systems* (6.6)–(6.8), (3.22).

Proof. On $Q(\sigma)$ we have $\bar{v}(x,t) = v_b$, where $\bar{H}_v^Q(x,t) \le -\sigma$, and $\bar{v}(x,t) = v_a$, where $\bar{H}_v^Q(x,t) \ge \sigma$. Therefore, $v \in V_{ad}^\varrho$ means $v(x,t) \in [v_b - \varrho, v_b]$ or $v(x,t) \in [v_a, v_a + \varrho]$, respectively. Lemma 6.5 yields $\tilde{H}_v^Q \le -\sigma/2$ or $\tilde{H}_v^Q \ge \sigma/2$ on $Q(\sigma)$, hence the variational inequality (6.6) gives $v^+ = v_b$ or $v^+ = v_a$, respectively. In this way, we have shown $v^+ = \bar{v}$ on $Q(\sigma)$; u^+ is handled analogously. □

COROLLARY 6.7. *Let the assumptions of Theorem* 6.4 *be satisfied and suppose that* $\|w_1 - \bar{w}\|_W \le \varrho := \min\{r_{\mathcal{N}}, \varrho_1, \varrho_2\}$. *Then* $\|\hat{w}_n - \bar{w}\|_W \le \varrho$ *holds* $\forall\ n \in N$. *In particular,* $\hat{v}_n \in V_{ad}^\varrho, \hat{u}_n \in U_{ad}^\varrho$.

This is obtained by Theorem 4.1 and the convergence estimate (4.9).

COROLLARY 6.8. *Under the assumptions of Corollary* 6.7, *the sign-conditions* (6.9)–(6.12) *hold true for* $(y^+, p^+, v^+, u^+) := (\hat{y}_n, \hat{p}_n, \hat{v}_n, \hat{u}_n)$.

(Corollary 6.7 yields $\hat{v}_n \in V_{ad}^{\varrho_2}, \hat{u}_n \in U_{ad}^{\varrho_2}$, hence the result follows from Lemma 6.5.)

COROLLARY 6.9. *Under the assumptions of Corollary* 6.7, *the solution* $(\hat{v}_n, \hat{u}_n)$ *of* $(\widehat{QP}_n)$ *satisfies the optimality system of* (QP_n), *too.*

Proof. The optimality systems for $(\widehat{\mathrm{QP}}_n)$ and (QP_n) differ only in the variational inequalities. From the optimality system of $(\widehat{\mathrm{QP}}_n)$ we know that

$$\int_Q \tilde{H}_v^Q(\hat{y}_n, \hat{p}_n, \hat{v}_n)(v - \hat{v}_n)\,dxdt \ge 0 \quad \forall v \in \widehat{V}_{ad}. \tag{6.13}$$

On $Q(\sigma), \hat{v}_n = \bar{v} = v_a$, if $\bar{H}_v^Q \ge \sigma$ and $\hat{v}_n = \bar{v} = v_b$, if $\bar{H}_v^Q \le -\sigma$. Lemma 6.5 and Corollary 6.8 yield $\tilde{H}_v^Q(\hat{y}_n, \hat{p}_n, \hat{v}_n) \ge \sigma/2$ or $\tilde{H}_v^Q(\hat{y}_n, \hat{p}_n, \hat{v}_n) \le -\sigma/2$, respectively. Therefore, $\tilde{H}_v^Q(\hat{y}_n, \hat{v}_n, \hat{p}_n)(v - \hat{v}_n) \ge 0$ holds on $Q(\sigma)$ for all real numbers $v \in [v_a, v_b]$. On the complement $Q \setminus Q(\sigma)$, the controls of $\widehat{V}_{ad}$ are not restricted to be equal to $\bar{v}$; hence in (6.13) v was arbitrary in $[u_a, u_b]$. This yields

$$\int_Q \tilde{H}_v^Q(v - \hat{v}_n)\,dxdt = \int_{Q\setminus Q(\sigma)} \tilde{H}_v^Q(v - \hat{v}_n)\,dxdt + \int_{Q(\sigma)} \tilde{H}_v^Q(v - \hat{v}_n)\,dxdt \ge 0 \qquad \forall v \in V_{ad},$$

where the nonnegativity of the first term follows from (6.13). The variational inequality for $\hat{u}_n$ is discussed in the same way. ∎

COROLLARY 6.10. *Let the assumptions of Corollary* 6.7 *be fulfilled. Then* $(\hat{v}_n, \hat{u}_n)$, *the solution of* $(\widehat{\mathrm{QP}}_n)$, *satisfies the optimality system for* (QP_n^ϱ).

Proof. By Corollary 6.9, $(\hat{v}_n, \hat{u}_n)$ satisfies the variational inequality (6.13) $\forall\, v \in V_{ad}, u \in U_{ad}$, in particular $\forall\, v \in V_{ad}^\varrho, u \in U_{ad}^\varrho$. Moreover, $\hat{v}_n \in V_{ad}^\varrho, \hat{u}_n \in U_{ad}^\varrho$ is granted by Corollary 6.9. ∎

LEMMA 6.11. *Assume that* $\bar{w} = (\bar{y}, \bar{p}, \bar{v}, \bar{u})$ *satisfies the second order condition* (SSC). *If* $\varrho_3 > 0$ *is taken sufficiently small, and* $\|w_n - \bar{w}\|_W \le \varrho_3$, *then for all* $\varrho > 0$ *the problem* (QP_n^ϱ) *has at least one pair of (globally) optimal controls* (v, u).

Proof. If $\|w_n - \bar{w}\|_W \le \varrho_3$ and $\varrho_3 > 0$ is sufficiently small, then

$$H_{vv}^Q(x,t,y_n(x,t),p_n(x,t),v_n(x,t)) \ge \frac{\delta}{2} \quad \text{a.e. on } Q, \tag{6.14}$$

$$H_{uu}^\Sigma(x,t,y_n(x,t),p_n(x,t),u_n(x,t)) \ge \frac{\delta}{2} \quad \text{a.e. on } \Sigma \tag{6.15}$$

follow from (LC), $\|y_n - \bar{y}\|_{C(\bar{Q})} + \|p_n - \bar{p}\|_{C(\bar{Q})} + \|v_n - \bar{v}\|_{L^\infty(Q)} + \|u_n - \bar{u}\|_{L^\infty(\Sigma)} \le \varrho_3$, and the Lipschitz properties of H_{vv}^Q, H_{vv}^Σ. Notice that w_n belongs to a set of diameter $K := \|\bar{w}\|_W + \varrho_3$, hence the Lipschitz estimates (3.5) and (3.6) apply. Therefore, (QP_n^ϱ) has the following properties. It is a linear-quadratic problem with linear equation of state. In the objective, the controls appear linearly and convex-quadratically (with convexity following from (6.14)–(6.15)). The control-state mapping $(v, u) \mapsto y$ is compact from $L^2(Q) \times L^2(\Sigma)$ to Y. Moreover, $V_{ad}^\varrho, U_{ad}^\varrho$ are nonempty weakly compact sets of L^2. Now the existence of at least one optimal pair of controls follows by standard arguments. Here, it is essential that the quadratic control-part of J_n is weakly l.s.c. with respect to the controls and that products of the type $y \cdot v$ or $y \cdot u$ lead to sequences of the type "strongly convergent times weakly convergent sequence," so that $y_n \to y$ and $v_n \rightharpoonup v$ implies $y_n v_n \rightharpoonup yv$. ∎

Remark. Alternatively, this result can be deduced also from the fact that $(\hat{y}_n, \hat{v}_n, \hat{u}_n)$ satisfies together with $\hat{p}_n$ the first and second order necessary conditions for (QP_n^ϱ) and that the optimality system of (QP_n^ϱ) is uniquely solvable (cf. Thm. 6.12).

THEOREM 6.12. *Let* $\bar{w} = (\bar{y}, \bar{p}, \bar{v}, \bar{u})$ *fulfill the first order necessary conditions* (2.1), (3.2)–(3.3), (3.7)–(3.9) *together with the second order sufficient optimality condition* (SSC). *If* $w_n = (y_n, p_n, v_n, u_n) \in W$ *is given such that* $\max\{\, \|w_n - \bar{w}\|_W, \varrho\} \le \min\{r_{\mathcal{N}}, \varrho_1, \varrho_2, \varrho_3\}$, *then the solution* $(\hat{v}_n, \hat{u}_n)$ *of* $(\widehat{\mathrm{QP}}_n)$ *is (globally) optimal for* (QP_n^ϱ). *Together with* $\hat{y}_n, \hat{p}_n$, *it delivers the unique solution of the optimality system of* (QP_n^ϱ).

Proof. Denote by (v^+, u^+) the solution of (QP_n^ϱ), which exists according to Lemma 6.11. Therefore, $(y^+, p^+, v^+, u^+) = w^+$ has to satisfy the associated optimality system. On the other hand, also $\hat{w}_n = (\hat{y}_n, \hat{p}_n, \hat{v}_n, \hat{u}_n)$ fulfills this optimality system by Corollary 6.10. We show that the solution of the optimality system is unique, and then the theorem is proven.

Let us assume that another $\hat{w} = (\hat{y}, \hat{p}, \hat{v}, \hat{u})$ obeys the optimality system too. Inserting $(\hat{v}, \hat{u})$ in the variational inequalities for (v^+, u^+), while (v^+, u^+) is inserted in the corresponding ones for $(\hat{v}, \hat{u})$, we arrive at

$$\begin{aligned} &\int_Q \{\tilde{H}_v^Q(y^+, p^+, v^+)(\hat{v} - v^+) + \tilde{H}_v^Q(\hat{y}, \hat{p}, \hat{v})(v^+ - \hat{v})\}\, dxdt \\ &\quad + \int_\Sigma \{\tilde{H}_u^\Sigma(y^+, p^+, u^+)(\hat{u} - u^+) + \tilde{H}_u^\Sigma(\hat{y}, \hat{p}, \hat{u})(u^+ - \hat{u})\}\, dSdt \ge 0. \end{aligned} \tag{6.16}$$

The expressions under the integral over Q in (6.16) have the form

$$f_v^n(\hat v-v^+)+H_{yv}^{Q,n}(y^+-y_n)(\hat v-v^+)+H_{vv}^{Q,n}(v^+-v_n)(\hat v-v^+)-p^+d_v^n(\hat v-v^+)$$
$$+f_v^n(v^+-\hat v)+H_{yv}^{Q,n}(\hat y-y_n)(v^+-\hat v)+H_{vv}^{Q,n}(\hat v-v_n)(v^+-\hat v)-\hat p\,d_v^n(v^+-\hat v);$$

the other terms look similar. Simplifying (6.16), we get, after setting $y=\hat y-y^+$, $v=\hat v-v^+$, $u=\hat u-u^+$, $p=\hat p-p^+$,

$$\begin{aligned}0 \;\le\; & -\int_Q \{H_{yv}^{Q,n}yv+H_{vv}^{Q,n}v^2+p\;d_v^n\;v\}\,dxdt\\ & -\int_\Sigma \{H_{yu}^{\Sigma,n}yu+H_{uu}^{\Sigma,n}u^2+p\;b_u^n\;u\}\,dSdt.\end{aligned}\tag{6.17}$$

The difference $p=\hat p-p^+$ obeys

$$\begin{aligned}-p_t+Ap &= H_{yy}^{Q,n}y+H_{yv}^{Q,n}v-d_y^np,\\ \partial_\nu p &= H_{yy}^{\Sigma,n}y+H_{yu}^{\Sigma,n}u-b_y^np,\\ p(T) &= \varphi_{yy}^n y(T).\end{aligned}\tag{6.18}$$

Multiplying the PDE in (6.18) by y and integrating over Q we find, after an integration by parts,

$$\begin{aligned}&-\int_\Omega p(T)y(T)dx+\int_0^T(y_t,p)_{H^1(\Omega)',H^1(\Omega)}dt+\int_Q\langle\mathcal{A}\nabla p,\nabla y\rangle dxdt\\ =\;&\int_Q(H_{yy}^{Q,n}y^2+H_{yv}^{Q,n}yv-d_y^n\;p\,y)\,dxdt+\int_\Sigma(H_{yy}^{\Sigma,n}y^2+H_{yu}^{\Sigma,n}yu-b_y^n\;p\;y)\,dSdt.\end{aligned}\tag{6.19}$$

This description of the procedure was formal, as the definition of the weak solution of (6.18) requires the test function y to be zero at $t=T$. To make (6.19) precise we have to use the information that $p\in W(0,T), y\in W(0,T)$ along with the integration by parts formula

$$\int_0^T(p_t,y)_{H^1(\Omega)',H^1(\Omega)}dt=\int_\Omega(p(T)y(T)-p(0)y(0))dx-\int_0^T(y_t,p)_{H^1(\Omega)',H^1(\Omega)}dt.$$

Next, we invoke the state equation for $y=\hat y-y^+$ and the condition for $p(T)$ to obtain from (6.19)

$$\begin{aligned}-\int_\Omega\varphi_{yy}^n y(T)^2dx \;&-\; \int_Q(H_{yy}^{Q,n}y^2+H_{yv}^{Q,n}yv)\,dxdt\\ -\int_\Sigma(H_{yy}^{\Sigma,n}y^2\;&+\;H_{yu}^{\Sigma,n}yu)\,dSdt=\int_Q d_v^n\;v\;p\;dxdt+\int_\Sigma d_u^n\;u\;p\;dSdt.\end{aligned}\tag{6.20}$$

Adding (6.20) to (6.17) yields

$$0\le-\int_\Omega\varphi_{yy}^n y(T)^2dx-\int_Q(y,v)D^2H^{Q,n}(y,v)^\top dxdt-\int_\Sigma(y,u)D^2H^{\Sigma,n}(y,u)^\top dSdt;$$

that is, $0\le -B^n[y,v,u]^2$. As $\max\{\,\|w_n-\bar w\|_W,\varrho\}\le\varrho_2$, Corollary 6.6 yields $v=0$ on $Q(\sigma)$ and $u=0$ on $\Sigma(\sigma)$. Therefore, Lemma 6.2 applies to conclude $\delta/2\,\|(y,v,u)\|_H^2\le 0$, i.e., $v=0, u=0$. In other words, $\hat v=v^+$, $\hat u=u^+$, completing the proof. □

Now we are able to formulate the main result of this paper.

THEOREM 6.13. *Let $\bar{w} = (\bar{y}, \bar{p}, \bar{v}, \bar{u})$ satisfy the assumptions of Theorem* 6.12 *and define $\varrho_{\mathcal{N}} = min\{r_{\mathcal{N}}, \varrho_1, \varrho_2, \varrho_3\}$. If $\max\{\varrho, \|w_1 - \bar{w}\|\} \le \varrho_{\mathcal{N}}$, then the sequence $\{w_n\} = \{(y_n, p_n, v_n, u_n)\}$ generated by the SQP method by solving (QP_n^{ϱ}) coincides with the sequence $\hat{w}_n$ obtained by solving $(\widehat{\mathrm{QP}}_n)$. Therefore, w_n converges q-quadratically to $\bar{w}$ according to the convergence estimate* (4.9).

Thanks to this theorem, we are justified to solve $(\mathrm{QP}_n^{\varrho})$ instead of $(\widehat{\mathrm{QP}}_n)$ to obtain the same (unique) solution. This result is still not completely satisfactory, as the unknown element $\bar{w}$ was used to define $(\mathrm{QP}_n^{\varrho})$.

However, an analysis of this section reveals that any convex, closed set $\tilde{V}_{ad}, \tilde{U}_{ad}$ can be taken instead of $V_{ad}^{\varrho}, U_{ad}^{\varrho}$, if the following properties are satisfied:

$\tilde{V}_{ad} \subset V_{ad}^{\varrho_{\mathcal{N}}}, \tilde{U}_{ad} \subset U_{ad}^{\varrho_{\mathcal{N}}}$, and $\tilde{V}_{ad} \supset V_{ad}^{\varrho_0}, \tilde{U}_{ad} \supset U_{ad}^{\varrho_0}$ for some $\varrho_0 > 0$. (The last condition is needed to guarantee $\hat{v}_n = \bar{v}$ on $Q(\sigma), \hat{u}_n = \bar{u}$ on $\Sigma(\sigma)$ and, last but not least, to make the convergence $\hat{v}_n \to \bar{v}, \hat{u}_n \to \bar{u}$ possible.)

Define, for instance, $\varrho_0 = \|\bar{w} - w_1\|_W$, where $\varrho_0 \le \frac{1}{3}\varrho_{\mathcal{N}}$,

$$\tilde{V}_{ad} = \{v \in V_{ad} |\ \|v - v_1\|_{L^\infty(Q)} \le 2\varrho_0\},$$

$$\tilde{U}_{ad} = \{u \in U_{ad} |\ \|u - u_1\|_{L^\infty(\Sigma)} \le 2\varrho_0\},$$

where $\varrho_0 = \|\bar{w} - w_1\|_W$ is the distance of the starting element of the SQP method to $\bar{w}$. Then $V_{ad}^{\varrho_0} \subset \tilde{V}_{ad} \subset V_{ad}^{\varrho_{\mathcal{N}}}$. The same property holds for $\tilde{U}_{ad}$. In that case, the SQP method will deliver the same solution in $\tilde{V}_{ad}, \tilde{Q}_{ad}$ as in $V_{ad}^{\varrho_{\mathcal{N}}}, U_{ad}^{\varrho_{\mathcal{N}}}$. This however, is the solution in $\widehat{V}_{ad}, \widehat{U}_{ad}$.

Remark. The restriction of the admissible sets to V_{ad}^{ϱ}, U_{ad}^{ϱ} might appear artificial, since restrictions of this type are not known from the theory of SQP methods in spaces of finite dimension. However, it is indispensable. In finite dimensions, the set of active constraints is detected after one step, provided that the starting value was chosen sufficiently close to the reference solution. The further analysis can rely on this. Here, we cannot determine the active set in finitely many steps unless we assume this a priori as in the definition of $\widehat{\mathrm{QP}}_n$.

REFERENCES

[1] W. ALT, *The Lagrange-Newton method for infinite-dimensional optimization problems*, Numer. Funct. Anal. Optim., 11 (1990), pp. 201–224.

[2] W. ALT, *The Lagrange Newton method for infinite-dimensional optimization problems*, Control Cybernet., 23 (1994), pp. 87–106.

[3] W. ALT, *Discretization and mesh independence of Newton's method for generalized equations*, in Mathematical Programming with Data Perturbation, A.K. Fiacco, ed., Lecture Notes in Pure and Appl. Math. 195, Marcel Dekker, New York, pp. 1–30.

[4] W. ALT, R. SONTAG, AND F. TRÖLTZSCH, *An SQP method for optimal control of a weakly singular Hammerstein integral equation*, Appl. Math. Optim., 33 (1996), pp. 227–252.

[5] W. ALT AND K. MALANOWSKI, *The Lagrange–Newton method for nonlinear optimal control problems*, Comput. Optim. Appl., 2 (1993), pp. 77–100.

[6] W. ALT AND K. MALANOWSKI, *The Lagrange–Newton method for state-constrained optimal control problems*, Comput. Optim. Appl., 4 (1995), pp. 217–239.

[7] E. CASAS, *Pontryagin's principle for state-constrained boundary control problems of semilinear parabolic equations*, SIAM J. Control Optim., 35 (1997), pp. 1297–1327.

[8] A.L. DONTCHEV, *Local analysis of a Newton–type method based on partial linearization*, in Proceedings of the AMS-SIAM Summer Seminar in Applied Mathematics on Mathematics of Numerical Analysis: Real Number Algorithms, J. Renegar, M. Shub, and S. Smale, eds., AMS Lectures in Appl. Math. 32, 1996, pp. 295–306.

[9] A.L. Dontchev and W.W. Hager, *Lipschitzian stability in nonlinear control and optimization*, SIAM J. Control Optim., 31 (1993), pp. 569–603.
[10] A.L. Dontchev, W.W. Hager, A.B. Poore, and B. Yang, *Optimality, stability, and convergence in optimal control*, Appl. Math. Optim., 31 (1995), pp. 297–326.
[11] H. Goldberg and F. Tröltzsch, *On a Lagrange–Newton method for a nonlinear parabolic boundary control problem*, Optim. Methods Softw., 8 (1998), pp. 225-247.
[12] H. Goldberg and F. Tröltzsch, *On a SQP-multigrid technique for nonlinear parabolic boundary control problems*, in Optimal Control: Theory, Algorithms, and Applications, W.W Hager and P.M. Pardalos, eds., Kluwer Academic Publishers B.V., Norwell, MA, 1998, pp. 154–177.
[13] M. Heinkenschloss, *The numerical solution of a control problem governed by a phase field model*, Optim. Methods Softw., 7 (1997), pp. 211–263.
[14] M. Heinkenschloss and E. W. Sachs, *Numerical solution of a constrained control problem for a phase field model*, in Control and Estimation of Distributed Parameter Systems, Internat. Ser. Numer. Math. 118, Birkhäuser Verlag, Basel, 1994, pp. 171–188.
[15] M. Heinkenschloss and F. Tröltzsch, *Analysis of the Lagrange-SQP-Newton method for the control of a phase field equation*, Control Cybernet., 28 (1999), pp. 177–211.
[16] K. Ito and K. Kunisch, *Augmented Lagrangian-SQP methods for nonlinear optimal control problems of tracking type*, SIAM J. Control Optim., 34 (1996), pp. 874–891.
[17] K. Ito and K. Kunisch, *Augmented Lagrangian-SQP methods in Hilbert spaces and application to control in the coefficients problems*, SIAM J. Optim., 6 (1996), pp. 96–125.
[18] N.H. Josephy, *Newton's Method for Generalized Equations*, Tech. Summary Report 1965, Mathematics Research Center, University of Wisconsin, Madison, WI, 1979.
[19] C.T. Kelley and E.W. Sachs, *Fast algorithms for compact fixed point problems with inexact function evaluations*, SIAM J. Sci. Comput., 12 (1991), pp. 725–742.
[20] C.T. Kelley and E.W. Sachs, *Multilevel algorithms for constrained compact fixed point problems*, SIAM J. Sci. Comput., 15 (1994), pp. 645–667.
[21] C.T. Kelley and E.W. Sachs, *Solution of optimal control problems by a pointwise projected Newton method*, SIAM J. Control Optim., 33 (1995), pp. 1731–1757.
[22] K. Kunisch, S. Volkwein, *Augmented Lagrangian-SQP techniques and their approximations*, Contemp. Math., 209 (1997), pp. 147–159.
[23] S.F. Kupfer and E.W. Sachs, *Numerical solution of a nonlinear parabolic control problem by a reduced SQP method*, Comput. Optim. Appl., 1 (1992), pp. 113–135.
[24] O.A. Ladyženskaya, V.A. Solonnikov, and N.N. Ural'ceva, *Linear and Quasilinear Equations of Parabolic Type*, Transl. Math. Monogr. 23, AMS, Providence, RI, 1968.
[25] J.L. Lions, *Contrôle optimal des systèmes gouvernés par des equations aux deŕivées partielles*, Dunod, Gauthier-Villars, Paris, 1968.
[26] J.L. Lions and E. Magenes, *Problèmes aux limites non homogènes et applications*, Vol. 1–3, Dunod, Paris, 1968.
[27] K. Machielsen, *Numerical Solution of Optimal Control Problems with State Constraints by Sequential Quadratic Programming in Function Space*, CWI Tract 53, CWI, Amsterdam, 1987.
[28] J.P. Raymond and H. Zidani, *Hamiltonian Pontryagin's principles for control problems governed by semilinear parabolic equations*, Appl. Math. Optim., 39 (1999), pp. 143–177.
[29] S.M. Robinson, *Strongly regular generalized equations*, Math. Oper. Res., 5 (1980), pp. 43–62.
[30] E.J.P.G. Schmidt, *Boundary control for the heat equation with nonlinear boundary condition*, J. Differential Equations, 78 (1989), pp. 89–121.
[31] F. Tröltzsch, *An SQP method for the optimal control of a nonlinear heat equation*, Control Cybernet., 23(1/2) (1994), pp. 267–288.
[32] F. Tröltzsch, *Convergence of an SQP-method for a class of nonlinear parabolic boundary control problems*, in Control and Estimation of Distributed Parameter Systems. Nonlinear Phenomena, W. Desch, F. Kappel, K. Kunisch, eds., Internat. Ser. Numer. Math. 118, Birkhäuser-Verlag, Basel, 1994, pp. 343–358.
[33] F. Tröltzsch, *Lipschitz Stability of Solutions to Linear-Quadratic Parabolic Control Problems with Respect to Perturbations*, Preprint-Series Fac. of Math., TU Chemnitz, Report 97–12, Dynam. Contin., Discrete Impul. Systems, accepted.
[34] A. Unger, *Hinreichende Optimalitätsbedingungen 2. Ordnung und Konvergenz des SQP-Verfahrens für semilineare elliptische Randsteuerprobleme*, Thesis, Technical University of Chemnitz, 1997.

SIAM J. CONTROL OPTIM.
Vol. 38, No. 1, pp. 313–335

ON THE STRONG STABILIZABILITY OF MIMO n-DIMENSIONAL LINEAR SYSTEMS*

JIANG QIAN YING†

Abstract. A plant is *strongly stabilizable* if there exists a *stable* compensator to stabilize it. Based on some theorems in complex analysis of several variables proved in this paper, we present necessary conditions for the strong stabilizability of complex and real n-D multi-input multi-output (MIMO) shift-invariant linear plants. For the real case, the condition is a generalization of the *parity interlacing property* of Youla, Bongiorno, and Lu [*Automatica J. IFAC*, 10 (1974), pp. 159–173] for the strong stabilizability of a real one-dimensional MIMO plant.

These conditions are also sufficient for the cases of n-D plants with a single output (MISO) or with a single input (SIMO). For general n-D MIMO plants, we do not know if the conditions are sufficient or not. A useful sufficient, but not necessary, condition for the strong stabilizability of a class of n-D ($n \geq 2$) MIMO plants is given.

Key words. multidimensional system, feedback stabilization, strong stabilizability, sign condition

AMS subject classifications. 93D15, 93D25, 93B27, 93B25, 32A, 32D15

PII. S0363012998335814

1. Introduction.

Let $\bar{U}^n = \{z \in \mathbf{C}^n \mid |z_1| \leq 1, \ldots, |z_n| \leq 1\}$ be the closed unit polydisc in $\mathbf{C}^n$. In this paper a polynomial in z is said to be *Hurwitz* if it is free from 0 in $\bar{U}^n$. A rational function with a Hurwitz denominator is regular (analytic) over $\bar{U}^n$ and will be said to be *stable*.

An n-dimensional (n-D) multi-input multi-output (MIMO) linear shift invariant plant with l inputs and m outputs can be described by a transfer matrix with entries of rational functions in $z = (z_1, \ldots, z_n)$:

$$P(z) = \begin{bmatrix} p_{11}(z) & \cdots & p_{1l}(z) \\ \vdots & \ldots & \vdots \\ p_{m1}(z) & \cdots & p_{ml}(z) \end{bmatrix}. \tag{1.1}$$

The system is called *real* if the entries are real rational functions and *complex* if they are complex rational functions. See [4] for more materials on n-D systems theory.

$P(z)$ is *stable* by definition if all its entries are stable. For an unstable plant $P(z)$, one may try to use a stabilizing compensator $C(z)$ in some feedback configuration to obtain a stable feedback system. In a standard feedback configuration [23, 22, 15], the stability of the feedback system is equivalent to the stability of the following system:

$$H(P,C) = \begin{bmatrix} I - P(I+CP)^{-1}C & -P(I+CP)^{-1} \\ (I+CP)^{-1}C & (I+CP)^{-1} \end{bmatrix} \tag{1.2}$$

*Received by the editors March 16, 1998; accepted for publication (in revised form) February 8, 1999; published electronically December 21, 1999.

http://www.siam.org/journals/sicon/38-1/33581.html

†Division of Regional Policy, Faculty of Regional Studies, Gifu University, 1-1 Yanagido, Gifu 501, Japan (ying@cc.gifu-u.ac.jp).

$$= \begin{bmatrix} (I+PC)^{-1} & -(I+PC)^{-1}P \\ C(I+PC)^{-1} & I - C(I+PC)^{-1}P \end{bmatrix}. \tag{1.3}$$

P is said to be *stabilizable* if such a C (complex or real) exists to make $H(P,C)$ stable. Stabilizability conditions and stabilizing compensator construction methods have been given in [22, 16, 15]. For real stabilizable systems, the compensators can always be constructed real.

In the case that the compensator $C(z)$ itself can be chosen stable, P is said to be *strongly stabilizable*. Furthermore, for a real system P, if C can be chosen real and stable, we say that P is *real strongly stabilizable*.

For an MIMO one-dimensional (1-D) linear system described by a real rational transfer matrix, Youla, Bongiorno, and Lu [27] in the early seventies gave a constructive condition for the existence of a stable real compensator. It can be easily derived from the works of [27], [20], and [25] that 1-D systems are always strongly stabilizable in the sense that a *complex* stable compensator always exists. See Appendix A of this paper for details.

Youla, Bongiorno, and Lu's work at the same time solved the problem of simultaneously stabilizing two 1-D plants with a single (not necessarily stable) compensator. In general, k 1-D plants can be simultaneously stabilized by a single compensator if certain other $k-1$ plants can be stabilized by a single *stable* compensator. However, the general simultaneous stabilization problem is rather complicated, as indicated by a recent result due to Blondel which says that for some *three* 1-D single-input single-output (SISO) plants, it is not rationally decidable whether or not they can be simultaneously stabilized by a single compensator [2, 3].

On the other hand, the extension of Youla, Bongiorno, and Lu's result to n-D systems or even more general linear systems over a commutative ring has also been a long standing open problem [23, Section 8.3]. For the problem of strong stabilization of n-D linear systems, recently a topological condition for strong stabilizability of an n-D SISO complex system was given by Shankar [20] and a computable equivalent was given by Ying [25]. By introducing a concept of "sign" of real functions on complex varieties, Ying gave a necessary and sufficient condition for real strong stabilizability of a real n-D SISO system [25]. But in the literature nothing significant has been known concerning the strong stabilizability of MIMO n-D systems.

In this paper we present new results concerning strong stabilizability of MIMO n-D systems. The main contributions are some mathematical theorems that extend Shankar and Ying's results. Applying these theorems, we are able to give necessary conditions for the strong stabilizability of MIMO systems. For two special classes of systems, i.e., MISO (multi-input single-output) and SIMO (single-input multi-output) systems, these conditions are also sufficient.

The concept of the sign of a real polynomial function on a complex variety, which has been first formulated in [25] and will also be involved in the part of this paper that concerns the real strong stabilizability of systems, along with the key theorems in several complex variables theory, will be stated in the next section. Readers who feel lack of motivation for the pure mathematical materials of section 2 may start directly with section 3 and look back to section 2 when necessary.

In section 3, necessary conditions for strong stabilizability of n-D MIMO systems are presented. Some examples are given at the end. In section 4, the sufficiency of the conditions for MISO and SIMO systems are proved. In section 5, a useful sufficient, but not necessary, condition for the strong stabilizability of a complex MIMO system is given. The difficulty in applying existing algebraic methods for establishing a

sufficient condition for strong stabilizability, and for stable compensator construction, is briefly discussed for general MIMO n-D systems.

2. Main mathematical theorems. Throughout this paper, by saying that a function is analytic on a *closed* set in $\mathbf{C}^n$, we mean that the function is defined and analytic in some open neighborhood of the closed set. See, e.g., [12, 13] for the theory of analytic functions of complex variables.

In this section we give solutions to the following questions:

Let $g(z)$, $\alpha_1(z), \ldots, \alpha_M(z)$ be complex polynomials.

(i) What is the condition for the existence of M stable complex rational functions $h_1(z), \ldots, h_M(z)$ such that

$$g(z) + h_1(z)\alpha_1(z) + \cdots + h_M(z)\alpha_M(z) \neq 0 \quad \text{on } \bar{U}^n? \tag{2.1}$$

(ii) When $g(z)$, $\alpha_1(z), \ldots, \alpha_M(z)$ are real polynomials and the above inequality (2.1) has a solution, what is the condition for $h_1, \ldots, h_M$ to be real?

For the special case when $M = 1$, questions (i) and (ii) were answered by Shankar [20] and Ying [25], respectively, from which necessary and sufficient conditions for the strong stabilizability of complex and real n-D SISO plants were derived. In the following we present answers for the general case.

2.1. Complex stable rational functions. Clearly, a necessary condition for the existence of $h_1, \ldots, h_M$ such that the inequality (2.1) holds is that $g, \alpha_1, \ldots, \alpha_M$ do not have common zero in $\bar{U}^n$. It can be derived from Cartan's Theorems A or B (see, e.g., [12, 13, 21]) that this condition is also sufficient for the existence of $M + 1$ rational functions analytic on $\bar{U}^n$: $f_0(z), f_1(z), \ldots, f_M(z)$, such that

$$f_0(z)g(z) + f_1(z)\alpha_1(z) + \cdots + f_M(z)\alpha_M(z) \neq 0 \quad \text{on } \bar{U}^n.$$

In the following a stronger condition is given such that $f_0(z)$ can be fixed to a nonzero constant. Before this we review some necessary topological concepts.

Recall that the winding number of a cycle (a closed curve) γ in $\mathbf{C}^* = \mathbf{C}\backslash\{0\}$ is defined as

$$W(\gamma) = \frac{1}{2\pi i}\int_\gamma \frac{d\xi}{\xi} = \frac{1}{2\pi i}\int_\gamma d(\log \xi).$$

It is the number of times that γ winds around the origin counterclockwise. A single-valued logarithmic function $\log \xi$ can be defined in some subset Σ of $\mathbf{C}^*$ if and only if Σ does not contain any cycle with a nonzero winding number.

In general, for a subset Σ in some complex space, an analytic function $g : \Sigma \longrightarrow \mathbf{C}^*$ has a single-valued logarithmic function $\log g$ on Σ if and only if g maps any cycle in Σ into a cycle with winding number 0 in $\mathbf{C}^*$, i.e., $W(g(\gamma)) = 0$. This property is equivalent to that g induces a 0 homomorphism from the first homology group of Σ to that of $\mathbf{C}^*$ (see, e.g., [9]). If this is satisfied, we say that g is 0-*homotopic* on Σ.

For some positive real number r, an open polydisc in $\mathbf{C}^n$ of radius r is defined as

$$\Delta(0; r) = \{z = (z_1, \ldots, z_n) \in \mathbf{C}^n \mid |z_1| < r, \ldots, |z_n| < r\}.$$

Let I be the ideal generated by $\alpha_1, \ldots, \alpha_M$ in $\mathbf{C}[z]$, the ring of complex polynomials, and let

$$V(I) \cap \bar{U}^n = \left\{z \in \bar{U}^n \mid f(z) = 0 \quad \forall f(z) \in I\right\}.$$

THEOREM 2.1. *Let $g(z), \alpha_1(z), \ldots, \alpha_M(z)$ be complex polynomials. A necessary and sufficient condition for the existence of analytic functions $h_1, \ldots, h_M$ on an open polydisc $\Omega = \Delta(0; r)$ containing $\bar{U}^n$, such that*

$$g(z) + \sum_{1 \le j \le M} h_j \alpha_j \neq 0 \quad \text{on } \bar{U}^n \tag{2.2}$$

is that $g(z)$ is 0-homotopic on $V(I) \cap \bar{U}^n$; or equivalently, $g(z)$ has a single-valued logarithmic function $\log g(z)$ on $V(I) \cap \bar{U}^n$.

Proof. Let $\mathcal{I}$ be the ideal sheaf generated by $(\alpha_1, \ldots, \alpha_M)$ in the sheaf $\mathcal{O}$ of analytic functions on some open polydisc $\Omega \supset \bar{U}^n$. We have the following exact sequence of sheaves:

$$0 \to \mathcal{I} \to \mathcal{O} \to \mathcal{O}/\mathcal{I} \to 0, \tag{2.3}$$

which induces a long exact sequence of modules over the ring of analytic functions

$$0 \to \Gamma(\mathcal{I}, \Omega) \to \Gamma(\mathcal{O}, \Omega) \xrightarrow{\varphi} \Gamma(\mathcal{O}/\mathcal{I}, \Omega) \to H^1(\mathcal{I}) \to \cdots . \tag{2.4}$$

The assumption that $\log g(z)$ is defined over $V(I)$ means that $\log g(z)$ is a global section of the quotient sheaf $\mathcal{O}/\mathcal{I}$ on Ω, that is, $\log g(z) \in \Gamma(\mathcal{O}/\mathcal{I}, \Omega)$.

Now $\mathcal{I}$ is finitely generated; thus it is a coherent sheaf. By Cartan's Theorem B, we have $H^1(\mathcal{I}) = 0$, which means that φ is surjective and there exists a $G(z) \in \Gamma(\mathcal{O}, \Omega)$ such that $\varphi(G(z)) = \log g(z)$.

Clearly, $\varphi(g(z)) = e^{\log g(z)}$ and $\varphi(e^{G(z)}) = e^{\varphi(G(z))}$ in $\Gamma(\mathcal{O}/\mathcal{I}, \Omega)$. We have

$$\varphi\big(e^{G(z)} - g(z)\big) = 0 \quad \text{in } \Gamma(\mathcal{O}/\mathcal{I}, \Omega); \tag{2.5}$$

therefore

$$e^{G(z)} - g(z) \in \Gamma(\mathcal{I}, \Omega). \tag{2.6}$$

That is, there exist $h_j(z) \in \Gamma(\mathcal{O}, \Omega)$, $j = 1, \ldots, M$, such that

$$e^{G(z)} = g(z) + \sum_{1 \le j \le M} h_j(z)\alpha_j(z) \neq 0 \quad \text{on } \Omega. \tag{2.7}$$

Since analytic functions can be approximated to any precision by polynomials on $\bar{U}^n$, thus also by stable rational functions, the inequality

$$g(z) + \sum_{1 \le j \le M} h_j(z)\alpha_j(z) \neq 0$$

still holds if the h_i's are replaced with their rational approximators. This proves the sufficiency of the condition.

For necessity, let

$$G(z) = \log\left(g(z) + \sum_{1 \le j \le M} h_j(z)\alpha_j(z)\right),$$

the restriction of $G(z)$ on $V(I) \cap \bar{U}^n$ is a logarithm of $g(z)$. This completes the proof. ∎

2.2. Concept of signs of real polynomial functions. Suppose that $g(z)$, $\alpha_1(z), \ldots, \alpha_M(z)$ are real and that there exist real stable rational functions $h_1(z)$, $\ldots, h_M(z)$ such that

$$g(z) + h_1(z)\alpha_1(z) + \cdots + h_M(z)\alpha_M(z) \neq 0 \quad \text{on } \bar{U}^n.$$

This is a real valued continuous mapping on $\bar{U}^n \cap \mathbf{R}^n$ and

$$g(z) + h_1(z)\alpha_1(z) + \cdots + h_M(z)\alpha_M(z) \neq 0 \quad \text{on } \bar{U}^n \cap \mathbf{R}^n. \tag{2.8}$$

The left side of 2.8 must have an invariant sign on the connected set $\bar{U}^n \cap \mathbf{R}^n$. Therefore g has an invariant sign, either $+$ or $-$, over $V(I) \cap \bar{U}^n \cap \mathbf{R}^n$, where I denotes the ideal generated by $\alpha_1, \ldots, \alpha_M$ in $\mathbf{C}[z]$.

Here we extend this condition to a stronger one which is concerned with the complex components which may not possess a real point. Recall that a connected component, or simply a component, is defined as a connected subset which is not contained in any larger connected subset; see, e.g., [9].

In the following we use "-" to denote a complex conjugation operator. From the context there should be no confusion with the "-" in the notation "$\bar{U}^n$" which means the closedness of the polydisc. We first make clear an important aspect of the structure of a connected component X of $V(I) \cap \bar{U}^n$. Let $z_0 \in X$; if $\overline{z_0} \in X$, then for any $z \in X$, $\overline{z} \in X$. In fact, since z_0, $\overline{z_0}$, and z are all in X, there exist a path γ_0 connecting z_0 to $\overline{z_0}$ and a path γ connecting z_0 to z. The path $\overline{\gamma}$ defined as $t \to \overline{\gamma(t)}$ then connects $\overline{z_0}$ to $\overline{z}$; hence we have $\bar{z} \in X$. In this case X is identical to $\overline{X} = \{\overline{z} \mid z \in X\}$ and is said to be *self-conjugate* in this paper.

On the other hand, if $X \neq \bar{X}$, then $X \cap \bar{X} = \phi$. In this case both X and $\bar{X}$ are non-self-conjugate components.

It is evident from the above argument that a connected component containing a real point is self-conjugate. Now suppose that X is a component containing a real point z_0. We reformulate the above sign consistency condition into one which is independent of the real point. We assume $g(z_0) > 0$. Let G_0 be a real number such that

$$e^{G_0} = g(z_0).$$

For an arbitrary point $z \in X$, let γ be a path in X connecting z_0 to z, define

$$G(z) = G_0 + \int_\gamma \frac{dg}{g}.$$

G is an analytic function on X. Let $\overline{\gamma}$ be the path defined as

$$\overline{\gamma}(t) = \overline{\gamma(t)} \quad \text{for } t \in [0, 1].$$

We have

$$\overline{G(z)} = \overline{G_0} + \int_\gamma \frac{d\overline{g(z)}}{\overline{g(z)}} = \overline{G_0} + \int_\gamma \frac{dg(\overline{z})}{g(\overline{z})} = G_0 + \int_{\overline{\gamma}} \frac{dg(z)}{g(z)} = G(\overline{z}).$$

To summarize, $G(z)$ is an analytic function on X such that

$$e^{G(z)} = g(z) \quad \text{and} \quad \overline{G(z)} = G(\overline{z}) \quad \forall z \in X. \tag{2.9}$$

On the other hand, the existence of such a $G(z)$ implies that $g(z) = e^{G(z)} > 0$ for $z \in X \cap \mathbf{R}^n$, because $G(z)$ is real on $\mathbf{R}^n$. The case $g(z_0) < 0$ can be proved similarly.

DEFINITION 2.2. *Let X be a self-conjugate connected component of $V(I) \cap \bar{U}^n$. Then g is said to have a* positive sign *on X if there exists an analytic function G on X such that*

$$e^{G(z)} = g(z) \quad \textit{and} \quad \overline{G(z)} = G(\overline{z}) \quad \forall z \in X;$$

g is said to have a negative sign *on X if $-g$ has a positive sign on X.*

The following proposition asserts that g either has a positive sign or has a negative sign on X.

PROPOSITION 2.3. *Let $g(z), \alpha_1(z), \ldots, \alpha_M(z)$ be real polynomials. Let I denote the ideal generated by $\alpha_1(z), \ldots, \alpha_M(z)$ in $\mathbf{C}[z]$. Assume that $g(z)$ has logarithm on $V(I) \cap \bar{U}^n$. Let X be a self-conjugate component of $V(I) \cap \bar{U}^n$, and let γ_0 be a path in X connecting a point $z_0 \in X$ to its conjugate $\overline{z_0}$. Let G_0 be a complex (possibly real) number such that*

$$e^{G_0} = g(z_0).$$

Then

$$\int_{\gamma_0} \frac{dg}{g} = \overline{G_0} - G_0 + m2\pi i, \tag{2.10}$$

where m is an integer. Moreover, g has a positive sign if m is an even number and g has a negative sign if m is an odd number.

Proof. Let

$$\tilde{G_0} = G_0 + \int_{\gamma_0} \frac{dg}{g};$$

then we have

$$e^{\tilde{G_0}} = g(\overline{z_0}) = e^{\overline{G_0}},$$

and hence

$$\tilde{G_0} = \overline{G_0} + m2\pi i$$

for some integer m. Therefore

$$\int_{\gamma_0} \frac{dg}{g} = \tilde{G_0} - G_0 = \overline{G_0} - G_0 + m2\pi i.$$

If $m = 2k$, for an integer k, we set

$$G_{z_0} = G_0 - k2\pi i.$$

It then follows that

$$G_{z_0} + \int_{\gamma_0} \frac{dg}{g} = G_0 - k2\pi i + \overline{G_0} - G_0 + m2\pi i = \overline{G_0} + k2\pi i = \overline{G_{z_0}}.$$

For any $z \in X$, let γ be a path connecting z_0 to z; from the assumption of the proposition, the integral

$$G(z) = G_{z_0} + \int_{\gamma} \frac{dg}{g}$$

is independent of the path and therefore defines a function $G(z)$ in z on X.

If γ is a path connecting z_0 to z, then $\overline{\gamma}$ is a path connecting $\overline{z_0}$ to $\overline{z}$, and the concatenation $\gamma_0\overline{\gamma}$, then connects z_0 to $\overline{z}$,

$$G(\overline{z}) = G_{z_0} + \int_{\gamma_0\overline{\gamma}} \frac{dg}{g} = \overline{G_{z_0}} + \int_{\overline{\gamma}} \frac{dg}{g} = \overline{G_{z_0} + \int_{\gamma} \frac{dg}{g}} = \overline{G(z)}.$$

In this case g has a positive sign on X.

Else if $m = 2k - 1$ for an integer k, we set

$$G_0' = G_0 - (2k-1)\pi i.$$

We have

$$-g(z_0) = -e^{G_0} = e^{G_0'},$$
$$G_0' + \int_{\gamma_0} \frac{d(-g)}{(-g)} = G_0 - (2k-1)\pi i + \overline{G_0} - G_0 + m2\pi i = \overline{G_0'}.$$

In this case $-g$ has a positive sign and hence g has a negative sign on X. □

Let X be a non-self-conjugate component of $V(I) \cap \bar{U}^n$. Then $\bar{X}$ is also a non-self-conjugate component since $X \cap \bar{X} = \phi$. If some analytic function $G(z)$ is defined on X such that

$$e^{G(z)} = g(z) \quad \forall z \in X,$$

then we can extend it to the union $X \cup \bar{X}$ by letting

$$G(\xi) = \overline{G(\bar{\xi})}, \quad \xi \in \bar{X}.$$

It is clear that $G(z)$ satisfies

$$e^{G(z)} = g(z), \quad \text{and} \quad \overline{G(z)} = G(\bar{z}) \quad \forall z \in X \cup \bar{X}.$$

In this sense, we say that $g(z)$ has a positive sign on $X \cup \bar{X}$. On the other hand, since the above argument also holds exactly for $-g(z)$, we may also say that $g(z)$ simultaneously has a negative sign on $X \cup \bar{X}$.

Therefore we are justified to make the following definition.

DEFINITION 2.4. *$g(z)$ is said to have a positive (negative, respectively) sign on $V(I) \cap \bar{U}^n$ if $g(z)$ has a positive (negative, respectively) sign over the union of the self-conjugate components of $V(I) \cap \bar{U}^n$.*

It is clear that $g(z)$ has a positive (negative, respectively) sign on $V(I) \cap \bar{U}^n$ if there exists an analytic function G on $V(I) \cap \bar{U}^n$ such that

$$e^{G(z)} = g(z)\big(e^{G(z)} = -g(z), \text{ respectively}\big) \quad \text{and} \quad \overline{G(z)} = G(\bar{z}) \quad \forall z \in V(I) \cap \bar{U}^n.$$

In particular, this is true if there is no self-conjugate component at all.

2.3. Real stable rational functions. We are now able to answer the second question proposed at the beginning of this section.

THEOREM 2.5. *Let $g(z), \alpha_1(z), \ldots, \alpha_M(z)$ be real polynomials. Let I denote the ideal generated by $\alpha_1(z), \ldots, \alpha_M(z)$ in $\mathbf{C}[z]$. A necessary and sufficient condition for the existence of real stable rational functions $h_1, \ldots, h_M$, such that*

$$g(z) + \sum_{1 \le j \le M} h_j \alpha_j \neq 0 \quad \text{on } \bar{U}^n \tag{2.11}$$

is that $g(z)$ is 0-homotopic on $V(I) \cap \bar{U}^n$ and that $g(z)$ has an invariant sign on $V(I) \cap \bar{U}^n$.

Proof. For proof of sufficiency, clearly it suffices to treat the case when $g(z)$ has a positive sign on $V(I) \cap \bar{U}^n$, as will be assumed in the following. From the statement at the end of section 2.2, it is seen that an analytic logarithmic function $\log g(z)$ can be constructed such that

$$\overline{\log g(z)} = \log g(\bar{z}) \quad \text{on } V(I) \cap \bar{U}^n. \tag{2.12}$$

Let Ω be an open polydisc containing $\bar{U}^n$ small enough so that $\log g(z)$ can be extended to $V(I) \cap \Omega$, where the above equation still holds. From the proof of Theorem 2.1, we have an analytic function $G(z)$ on Ω such that

$$e^{G(z)} = \log g(z) \quad \forall z \in V(I) \cap \Omega.$$

Then the analytic function defined as

$$W(z) = \big(G(z) + \overline{G(\bar{z})}\,\big)/2, \quad z \in \Omega, \tag{2.13}$$

satisfies

$$\varphi\big(W(z)\big) = \log g(z) \quad \text{in } \Gamma(\mathcal{O}/\mathcal{I}, \Omega),$$

where $\varphi : \Gamma(\mathcal{O}, \Omega) \to \Gamma(\mathcal{O}/\mathcal{I}, \Omega)$ is the canonical map. It follows that there exist $h_j(z) \in \Gamma(\mathcal{O}, \Omega)$, $j = 1, \ldots, M$, such that

$$e^{W(z)} = g(z) + \sum_{1 \le j \le M} h_j(z)\alpha_j(z) \neq 0 \quad \text{on } \bar{U}^n. \tag{2.14}$$

Now $W(z)$ satisfies

$$\overline{W(z)} = W(\bar{z}) \tag{2.15}$$

and can be chosen as a power series with real coefficients convergent on Ω. It is thus clear that $h_1(z), \ldots, h_M(z)$ can be chosen as a power series with real coefficients convergent on Ω and can be approximated by real stable rational functions.

For necessity, suppose that real stable rational functions $h_1(z), \ldots, h_M(z)$ have been given such that

$$u(z) = g(z) + \sum_{1 \le j \le M} h_j(z)\alpha_j(z) \neq 0 \quad \text{on } \Omega.$$

If $u(0) > 0$, let $\log u(0)$ be defined as a real number. Define

$$W(z) = \log u(0) + \int_\gamma \frac{du(z)}{u(z)} \tag{2.16}$$

for any γ connecting 0 to z in $\bar{U}^n$. Now W satisfies

$$\overline{W(z)} = W(\bar{z}), \qquad e^{W(z)} = g(z) \quad \forall z \in V(I) \cap \bar{U}^n;$$

$g(z)$ therefore has a positive sign on $V(I) \cap \bar{U}^n$.

Similarly, it could be shown that $g(z)$ has a negative sign if $u(0) < 0$. □

3. Necessary conditions for strong stabilizability of MIMO systems. In this section we will adopt the *matrix fraction description* (MFD) approach for describing a system. A *left* MFD of P is defined as

$$P(z) = D^{-1}(z)N(z),$$

where D is an $m \times m$ and N an $m \times l$ polynomial matrix:

$$D = \begin{bmatrix} d_{11} & \cdots & d_{1m} \\ \vdots & \ldots & \vdots \\ d_{m1} & \cdots & d_{mm} \end{bmatrix}, \qquad N = \begin{bmatrix} n_{11} & \cdots & n_{1l} \\ \vdots & \ldots & \vdots \\ n_{m1} & \cdots & n_{ml} \end{bmatrix}, \tag{3.1}$$

where d_{jk} and n_{jk} are polynomials in z with real or complex coefficients, corresponding to a real or a complex system, respectively. Let

$$F = [D\ N] = \begin{bmatrix} d_{11} & \cdots & d_{1m} & n_{11} & \cdots & n_{1l} \\ \vdots & \ldots & \vdots & \vdots & \ldots & \vdots \\ d_{m1} & \cdots & d_{mm} & n_{m1} & \cdots & n_{ml} \end{bmatrix}. \tag{3.2}$$

Let $M = \binom{m+l}{l}$. Let $\alpha_1, \alpha_2, \ldots, \alpha_M$ denote the M maximal order minors of F, with $\alpha_1 = \det D$.

An MFD is said to be *minor coprime* if the α_i's have no nonunit common factor. Similarly, a minor coprime *right* MFD is defined as a matrix fraction YX^{-1} such that the maximal order minors of $\left[\begin{smallmatrix} X \\ Y \end{smallmatrix}\right]$ have no common nonunit factor. A minor coprime MFD (either left or right) can always be constructed for a 1-D and a two-dimensional (2-D) transfer matrix, but not generally for an n-D ($n \geq 3$) matrix [28, 14].

In the following we will first treat MIMO systems with a minor coprime MFD because of their simplicity, although they will turn out to be special instances of a more general case treated subsequently.

3.1. Necessary conditions for strong stabilizability of MIMO systems with a minor coprime MFD. It is known that if $D^{-1}N$ is a minor coprime MFD for a plant P, then P is stable if and only if $\det D(z) \neq 0$ for $z \in \bar{U}^n$; see, e.g., [14, proof of Theorem 3.4].

If $D^{-1}N$ is a minor coprime MFD for an unstable plant P, consider a compensator C with minor coprime right MFD $C = YX^{-1}$. C stabilizes P if and only if the determinant of

$$D(z)X(z) + N(z)Y(z)$$

is free from zero in $\bar{U}^n$. See, e.g., [11, p. 77] for the 2-D case and [15, Theorem 2] for the general case.

By the Cauchy–Binet theorem, we have

$$\det[DX + NY] = \sum_{i=1}^{M} \alpha_i \beta_i,$$

where $\beta_1, \ldots, \beta_M$ are the maximal minors of $\begin{bmatrix} X \\ Y \end{bmatrix}$, with $\beta_1 = \det X$.

It follows obviously that a necessary condition for $P = D^{-1}N$ to be stabilizable is that there exist $M = \binom{m+l}{m}$ polynomials $\beta_1(z), \ldots, \beta_M(z)$ such that

$$\sum_{i=1}^{M} \alpha_i(z)\beta_i(z) \neq 0 \quad \text{for } z \in \bar{U}^n. \tag{3.3}$$

This condition is equivalent to that $\alpha_1, \ldots, \alpha_M$ have no common zero in $\bar{U}^n$ [11, 21].

Conversely, if such polynomials exist, a compensator $C = YX^{-1}$ can be constructed such that [23, 10, 11, 15]

$$\det[DX + NY] = \sum_{i=1}^{M} \alpha_i\beta_i \neq 0 \quad \text{on } \bar{U}^n. \tag{3.4}$$

It then follows that for a plant $P = D^{-1}N$ to be stabilizable it is necessary and sufficient that $\alpha_1, \ldots, \alpha_M$ have no common zero in $\bar{U}^n$. From the construction [11] of the compensator, it can be seen that a real compensator can always be found for a real plant which satisfies this condition.

For a 1-D system, minor coprimeness is equivalent to *zero coprimeness* [28] which means that the α_i's have no common zero at all. This implies that a 1-D system described with a minor coprime MFD is always stabilizable.

If the compensator $C = YX^{-1}$ itself can be chosen stable, then as a necessary condition $\beta_1(z)$ in (3.3) must be able to be chosen stable. Therefore we have the following theorem.

THEOREM 3.1. *Assume that P has a minor coprime MFD $P = D^{-1}N$.*

(i) *Let α_j, $j = 1, \ldots, M$, $M = \binom{m+l}{m}$ be the maximal order minors of $[D\ N]$ with $\alpha_1 = \det D$. A necessary condition for $P = D^{-1}N$ to be strongly stabilizable is that there are stable rational functions $h_2, \ldots, h_M$ such that*

$$\det D + \sum_{2 \leq j \leq M} h_j\alpha_j \neq 0 \quad \textit{on } \bar{U}^n.$$

(ii) *If $P = D^{-1}N$ is strongly stabilizable and is real, then a necessary condition for the existence of a real stable compensator is that the above h_j's can be chosen real.*

Recalling Theorems 2.1 and 2.5 gives the following theorem.

THEOREM 3.2. *Assume that P has a minor coprime MFD $P = D^{-1}N$. Let I be the ideal generated by $\alpha_2, \ldots, \alpha_M$ in $\mathbf{C}[z]$.*

(i) *A necessary condition for P to be strongly stabilizable is that $\det D(z)$ is 0-homotopic on $V(I) \cap \bar{U}^n$.*

(ii) *If P is strongly stabilizable and is real, then a necessary condition for the existence of a real stable compensator is that $\det D$ has an invariant sign on $V(I) \cap \bar{U}^n$.*

The following lemmas say that, instead of $V(I)$, it is equivalent to investigate the behavior of $\det D(z)$ on another variety $V(J)$, where J is the ideal generated by $n_{jk}(z)$, $1 \leq j \leq m, 1 \leq k \leq l$, the entries of $N(z)$.

LEMMA 3.3. *Assume $D^{-1}N$ is a minor coprime MFD for a stabilizable plant. Let $\alpha_1 = \det D, \alpha_2, \ldots, \alpha_M$ be the maximal order minors of $[D\ N]$, let I be the ideal generated by $\alpha_2, \ldots, \alpha_M$, and J be the ideal generated by n_{ij}, $1 \leq j \leq m, 1 \leq k \leq l$, the entries of $N(z)$. Then*

$$V(I) \backslash V(\det D) = V(J) \backslash V(\det D).$$

Proof. It is clear that

$$V(J) \subseteq V(I);$$

hence

$$V(J)\backslash V(\det D) \subseteq V(I)\backslash V(\det D).$$

Assume that

$$z_0 \in V(I)\backslash V(\det D).$$

If $z_0 \notin V(J)$, then there is an n_{jk} that is not zero at z_0. Since $\det D(z_0) \neq 0$, there are some $m-1$ columns of $D(z_0)$ independent of the kth column

$$[n_{1k}(z_0) \cdots n_{mk}(z_0)]^t$$

of $N(z_0)$. The corresponding minor of $[D\ N]$ is then not 0 at z_0. Thus $z_0 \notin V(I)$. This contradicts the assumption. Therefore

$$V(I)\backslash V(\det D) \subseteq V(J)\backslash V(\det D). \qquad \square$$

LEMMA 3.4. *With the same notations and assumptions as in Lemma* 3.3*, we have*

$$V(I) \cap \bar{U}^n = V(J) \cap \bar{U}^n.$$

Proof. Since $D^{-1}N$ is assumed to be a minor coprime MFD of a stabilizable plant, the maximum order minors necessarily do not have common zero on $\bar{U}^n$. In our notation, I is the ideal generated by all the minors excluding $\det D$; therefore we have

$$V(I) \cap \bar{U}^n \cap V(\det D) = \phi,$$
$$V(I) \cap \bar{U}^n = V(I) \cap \bar{U}^n\backslash V(\det D) = V(J) \cap \bar{U}^n\backslash V(\det D) \subseteq V(J) \cap \bar{U}^n.$$

On the other hand,

$$V(J) \cap \bar{U}^n \subseteq V(I) \cap \bar{U}^n.$$

This completes the proof. $\square$

The following theorem is obvious from Lemma 3.4 and Theorem 3.2.

THEOREM 3.5. *Suppose that P has a minor coprime MFD $D^{-1}N$. Let J be the ideal generated by the entries of N in $\mathbf{C}[z]$.*

(i) *A necessary condition for P to be strongly stabilizable is that* $\det D(z)$ *is* 0*-homotopic on $V(J) \cap \bar{U}^n$.*

(ii) *If P is strongly stabilizable and is real, then a necessary condition for the existence of a real stable compensator is that* $\det D$ *has an invariant sign on $V(J) \cap \bar{U}^n$.*

This theorem applies to all 1-D and 2-D systems, for which minor coprime MFDs always exist. For the 1-D case, the condition of (i) is trivial, because $V(I)$ consists of a finite number of discrete points. Furthermore, since the only self-conjugate components of $V(I) \cap \bar{U}^n$ are the real points, condition (ii) means exactly that $\det D$ has an invariant sign on $V(I) \cap \bar{U}^n \cap \mathbf{R}^n$. This is equivalent to the well-known *parity interlacing property* of Youla, Bongiorno, and Lu [27] and is sufficient for the existence of a real stable compensator; see Appendix A for more details.

Unfortunately, for 2-D and general n-D systems, we do not know if the conditions of (i) and (ii) in Theorems 3.2 and 3.5 are sufficient or not.

3.2. Necessary conditions for strong stabilizability of general MIMO systems. Let P be an n-D transfer matrix with a (not necessarily minor coprime) left MFD $D^{-1}N$. Let $F = [D\ N]$.

Let $\alpha_1, \alpha_2, \ldots, \alpha_M$ denote the $M = \binom{m+l}{l}$ maximal order minors of F, with $\alpha_1 = \det D$. Let d be the greatest common divisor of the $\alpha_i's$. The polynomials

$$b_1 = \frac{\alpha_1}{d}, \ldots, b_M = \frac{\alpha_M}{d} \tag{3.5}$$

are called *generating polynomials* of P and are independent of the choices of F up to a nonzero constant [14]. The following two propositions establish the stability and stabilizability conditions of an n-D system in terms of the generating polynomials.

PROPOSITION 3.6 (see [14]). *P is stable if and only if the first generating polynomial b_1 is free from 0 in $\bar{U}^n$.*

PROPOSITION 3.7 (see [15, 22]). *P is stabilizable if and only if the generating polynomials do not share a common zero in $\bar{U}^n$.*

In the following we assume the stabilizability of $P(z)$.

Let $e_1, \ldots, e_M$ be the generating polynomials of the compensator C. It is shown in [15] that the first generating polynomial of the resultant feedback system is

$$b_{H1} = r \sum_{j=1}^{M} b_j e_j, \tag{3.6}$$

where r is a nonzero constant.

If $P = D^{-1}N$ is strongly stabilizable, then the compensator C can be chosen stable. This implies that both e_1 and b_{H1} in the above equation are free from 0 in $\bar{U}^n$. Dividing the equation by e_1 and r, we have

$$\frac{b_{H1}}{re_1} = b_1 + \sum_{j=2}^{M} b_j \frac{e_j}{e_1}. \tag{3.7}$$

This leads to the following theorem.

THEOREM 3.8. (i) *Let b_j, $j = 1, \ldots, M$, $M = \binom{m+l}{m}$ be the generating polynomials of $P = D^{-1}N$. A necessary condition for P to be strongly stabilizable is that there exist stable rational functions $h_2, \ldots, h_M$ such that*

$$b_1 + \sum_{2 \le j \le M} h_j b_j \neq 0 \quad \text{on } \bar{U}^n.$$

(ii) *If $P = D^{-1}N$ is strongly stabilizable and is real, then a necessary condition for the existence of a real stable compensator is that the above h_j's can be chosen real.*

In view of Theorems 2.1 and 2.5 of section 2, we have the following theorem.

THEOREM 3.9. *Let $b_1, b_2, \ldots, b_M$ be the generating polynomials of $P = D^{-1}N$, and I' be the ideal generated by $b_2, \ldots, b_M$ in $\mathbf{C}[z]$.*

(i) *A necessary condition for P to be strongly stabilizable is that b_1 is 0-homotopic on $V(I') \cap \bar{U}^n$.*

(ii) *If P is strongly stabilizable and is real, then a necessary condition for the existence of a real stable compensator is that b_1 has an invariant sign on $V(I') \cap \bar{U}^n$.*

If $P = D^{-1}N$ is a *minor coprime* MFD, then the generating polynomials are exactly the maximal order minors of $[D\ N]$. In this case Theorems 3.8 and 3.9 reduce to Theorems 3.1 and 3.2, respectively.

3.3. Examples.

Example 1 (real strongly stabilizable).

$$F = [D\ N] = \begin{bmatrix} z_2^2 & 1 & z_1 & z_2 \\ -1 & z_3^2 & z_3 & 0 \end{bmatrix}.$$

The minors are

$$(\alpha_1, \alpha_2, \alpha_3, \alpha_4, \alpha_5, \alpha_6) = \big(1 + z_2^2 z_3^2, z_1 + z_2^2 z_3, z_2, z_3 - z_1 z_3^2, -z_2 z_3^2, -z_2 z_3\big).$$

Obviously $D^{-1}N$ is minor coprime. In accordance with our previous notations, $I = (\alpha_1, \ldots, \alpha_6)$, $J = (z_1, z_2, z_3)$. We have

$$V(I) \cap \bar{U}^3 = V(J) \cap \bar{U}^3 = \{(0, 0, 0)\}.$$

The necessary conditions of Theorems 3.2 and 3.5 are satisfied. For this plant we have the following stable real compensator $C = YX^{-1}$ with

$$\begin{bmatrix} X \\ Y \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ z_2 & 0 \\ -z_1 & z_3 \end{bmatrix}.$$

Example 2 (strongly stabilizable, but not real strongly stabilizable).

$$F = \big[z_1, z_1^2 + z_2 z_3 - z_2 - 2, z_3\big].$$
$$V(J) \cap \bar{U}^3 = V\big(z_1^2 + z_2 z_3 - z_2 - 2, z_3\big) \cap \bar{U}^3 = \{(-1, -1, 0), (1, -1, 0)\}$$

is a discrete point set, and the condition (i) for strong stabilizability of Theorem 3.5 is trivial. By trial, we found a *complex* stable compensator $C = YX^{-1}$,

$$\begin{bmatrix} X \\ Y \end{bmatrix} = \begin{bmatrix} 0.5z_1 + i \\ -1 \\ z_2 \end{bmatrix}.$$

In fact, it is not difficult to check that the polynomial

$$(0.5z_1 + i)z_1 - \big(z_1^2 + z_2 z_3 - z_2 - 2\big) + z_2 z_3 = -0.5z_1^2 + iz_1 + z_2 + 2$$

is free from 0 in $\bar{U}^2$ and hence in $\bar{U}^3$. However, z_1 has opposite signs at the two discrete points of $\bar{U}^3 \cap V(J)$, thus violating condition (ii) in Theorem 3.5. There is no real stable compensator. Using the Gröbner basis method [5, 11, 24], we can find a real compensator $\left[\begin{smallmatrix} z_1 \\ -1 \\ z_2 \end{smallmatrix}\right]$, which is, however, not stable.

Example 3 (not minor coprime MFD).

$$F = \begin{bmatrix} d(z_1, z_2, z_3) & 0 & z_1 z_2 \\ 0 & d(z_1, z_2, z_3) & 1 \end{bmatrix},$$
$$d = z_1^2 + z_2^2 + z_3^2 - 1.$$

The minors are

$$\alpha_1 = d^2, \qquad \alpha_2 = d, \qquad \alpha_3 = -dz_1 z_2.$$

$D^{-1}N$ is not minor coprime. The generating polynomials are

$$(b_1, b_2, b_3) = (d, 1, -z_1 z_2).$$

We have

$$1 \cdot b_1 + 5 \cdot b_2 = z_1^2 + z_2^2 + z_3^2 + 4,$$

a Hurwitz polynomial. Using the method of [10, 15], one can find a compensator $C = YX^{-1}$,

$$\begin{bmatrix} X \\ Y \end{bmatrix} = \begin{bmatrix} d+5 & -z_1 z_2 \\ 0 & d \\ 0 & 5d \end{bmatrix}.$$

$YX^{-1} = [0\ 5]$ is a right MFD of a stable system.

Example 4 (stabilizable, but not strongly stabilizable).

$$P = \frac{1 - 4z_1 z_2}{z_1} \begin{bmatrix} z_2 + 1 & z_2 \\ z_1 & z_1 \end{bmatrix},$$
$$F = \begin{bmatrix} z_1 & 0 & (z_2+1)f & z_2 f \\ 0 & z_1 & z_1 f & z_1 f \end{bmatrix},$$

where $f = 1 - 4z_1 z_2$. The 2×2 minors are

$$(\alpha_1, \alpha_2, \alpha_3, \alpha_4, \alpha_5, \alpha_6) = \big(z_1^2, z_1^2 f, z_1^2 f, -z_1(z_2+1)f, -z_1 z_2 f, z_1 f^2\big).$$

The generating polynomials are

$$(b_1, b_2, b_3, b_4, b_5, b_6) = \big(z_1, z_1 f, z_1 f, -(z_2+1)f, -z_2 f, f^2\big).$$

The plant is stabilizable because

$$V(b_1, b_6) \cap \bar{U}^2 = \phi.$$

But

$$V(b_2, \ldots, b_6) \cap \bar{U}^2 = V(1 - 4z_1 z_2) \cap \bar{U}^2$$

has a cycle

$$\gamma : t \longrightarrow \left(\frac{1}{2} e^{i2\pi t}, \frac{1}{2} e^{-i2\pi t}\right), \quad t \in [0, 1]$$

which is mapped by z_1 to a cycle $\{\frac{1}{2} e^{i2\pi t},\ 0 \le t \le 1\}$ in $\mathbf{C}^*$, which has winding number 1 around the origin. Thus the system is not strongly stabilizable.

4. MISO and SIMO systems. For two special classes of n-D systems, the MISO and SIMO systems, the necessary conditions for strong stabilizability given in Theorems 3.8 and 3.9 are actually sufficient. In the following theorem, the conditions are reformed for convenience of exposition. Their equivalence with that given in Theorem 3.9 is not difficult to verify.

THEOREM 4.1. *Let $P(z)$ be either an m-input/1-output or a 1-input/m-output system, described by the following corresponding transfer matrices:*

$$P(z) = d^{-1}(z)\big[n_1(z)\cdots n_m(z)\big] \quad \textit{and} \tag{4.1}$$

$$P(z) = \begin{bmatrix} n_1(z) \\ \vdots \\ n_m(z) \end{bmatrix} d^{-1}(z). \tag{4.2}$$

Let J be the ideal generated by $n_1(z), \ldots, n_m(z)$. Then

(i) *P is strongly stabilizable (by a complex compensator) if and only if $d(z)$ is 0-homotopic on $V(J) \cap \bar{U}^n$;*

(ii) *if P is a strongly stabilizable real plant, then a real stable compensator exists if and only if $d(z)$ has an invariant sign on the self-conjugate components of $V(J) \cap \bar{U}^n$.*

Proof. We give a proof for the SIMO case. Necessity: P has a left MFD

$$P = D^{-1}N = \begin{bmatrix} d & 0 & \cdots & 0 \\ 0 & d & & \vdots \\ \vdots & & \ddots & 0 \\ 0 & \cdots & 0 & d \end{bmatrix}^{-1} \begin{bmatrix} n_1 \\ \vdots \\ n_m \end{bmatrix}.$$

The maximal order minors of $[D\ N]$ are

$$d^m, d^{m-1}n_1, \ldots, d^{m-1}n_m,$$

and the generating polynomials are

$$d, n_1, \ldots, n_m.$$

The necessary part of the conditions is then obvious by Theorem 3.9. Sufficiency: By Theorems 2.1 and 2.5, there exist $y(z)(\neq 0 \text{ on } \bar{U}^n)$, $x_1(z), \ldots, x_m(z)$, real under condition (ii), such that

$$y(z)d(z) + x_1(z)n_1(z) + \cdots + x_m(z)n_m(z) \neq 0 \quad \text{on } \bar{U}^n.$$

Let

$$C(z) = y^{-1}X = y^{-1}(z)\big[x_1(z)\cdots x_m(z)\big].$$

It suffices to show that $H(P, C)$ in (1.2) is stable. $H(P, C)$ is stable if and only if

$$\begin{aligned} H'(P,C) &= \begin{bmatrix} P(I+CP)^{-1}C & P(I+CP)^{-1} \\ (I+CP)^{-1}C & (I+CP)^{-1} \end{bmatrix} \\ &= \begin{bmatrix} N(yd+XN)^{-1}X & N(yd+XN)^{-1}y \\ d(yd+XN)^{-1}X & d(yd+XN)^{-1}y \end{bmatrix} \\ &= \begin{bmatrix} N \\ d \end{bmatrix} (yd + x_1n_1 + \cdots + x_mn_m)^{-1} [Xy] \end{aligned} \tag{4.3}$$

is stable. Actually it is stable from the conditions. Therefore $C(z)$ is a stable compensator, real under condition (ii).

The proof for the MISO case is similar by using the alternate expression (1.3) in section 1 for $H(P, C)$. □

5. On sufficient conditions for strong stabilizability of MIMO systems. Based on the fact that all stable complex (or real) rational functions in one variable form a Euclidean domain, the necessary conditions in Theorem 3.5 have been constructively proven to be sufficient for 1-D systems, too; see Appendix A and [23, section 4.4]. The construction of a stable compensator for a 1-D system relies essentially on the Euclidean division algorithm in the domains of one-variable stable rational functions. But the domains of stable rational functions in two or more variables are no longer Euclidean and the method of [23] cannot be applied. Although we have not met any counterexample which disproves the sufficiency of the necessary conditions in Theorems 3.1, 3.2, 3.5, 3.8, and 3.9, we have not been able to prove the sufficiency.

In this section, by extending the analytic method used in section 2.1, we give a condition which is sufficient, but not necessary, for the strong stabilizability of a certain class of complex MIMO n-D plants. We also discuss some difficulties in adopting existing algebraic methods for the construction of stable compensators for MIMO n-D systems.

5.1. A sufficient condition for complex strong stabilizability.

THEOREM 5.1. *Let $D(z)$ and $N(z)$ be two complex polynomial $m \times m$ matrices. A sufficient condition for the existence of a complex polynomial $m \times m$ matrix $X(z)$ such that*

$$\det\big[D(z) + N(z)X(z)\big] \neq 0 \quad \text{on } \bar{U}^n \tag{5.1}$$

is that $D(z)$ has a single-valued logarithmic matrix function $\log D(z)$ on $V(\det N(z)) \cap \bar{U}^n$.

Proof. Let $\mathcal{O}$ be the sheaf of germs of analytic functions on some open polydisc $\Omega \supset \bar{U}^n$. Let $\mathcal{O}^{m\times m}$ be a free sheaf of modules of rank $m \times m$ over $\mathcal{O}$; let $\mathcal{N} = N \cdot \mathcal{O}^{m\times m}$ be a subsheaf of $\mathcal{O}^{m\times m}$ obtained by matrix multiplication with N, which is finitely generated. We have the following exact sequence of sheaves:

$$0 \to \mathcal{N} \to \mathcal{O}^{m\times m} \to \mathcal{O}^{m\times m}/\mathcal{N} \to 0, \tag{5.2}$$

which induces a long exact sequence of modules over the ring of analytic functions

$$0 \to \Gamma(\mathcal{N}, \Omega) \to \Gamma(\mathcal{O}^{m\times m}, \Omega) \xrightarrow{\varphi} \Gamma(\mathcal{O}^{m\times m}/\mathcal{N}, \Omega) \to H^1(\mathcal{N}) \to \cdots . \tag{5.3}$$

Since $\mathcal{N}$ is finitely generated, it is a coherent analytic sheaf. Therefore we have

$$H^1(\mathcal{N}) = 0, \tag{5.4}$$

and φ is a surjective mapping.

By assumption $\log D(z)$ is defined over $V(\det N)$. This can be analytically extended to an open neighborhood U_0 containing $V(\det N)$. Let

$$U_1 = \Omega \backslash V(\det N).$$

Consider the $m \times m$ zero matrix $\mathbf{0}$ as a section of $\mathcal{O}^{m\times m}$ over U_1. On $U_0 \cap U_1$, we have

$$\log D(z) - \mathbf{0} = N(z)\big(N(z)^{-1} \log D(z)\big).$$

This implies that $\log D(z)$ can actually be viewed as a global section of the quotient sheaf $\mathcal{O}^{m\times m}/\mathcal{N}$ on Ω, which satisfies

$$e^{\log D(z)} = D(z) \quad \text{for } z \in V(\det N) \cap \Omega. \tag{5.5}$$

Since φ is surjective, there exists a global section $G(z)$ of $\mathcal{O}^{m\times m}$ on Ω, such that

$$\varphi\big(G(z)\big) = \log D(z) \quad \text{for } z \in V(\det N) \cap \Omega. \tag{5.6}$$

Clearly, $\varphi(D(z)) = e^{\log D(z)}$ and $\varphi(e^{G(z)}) = e^{\varphi(G(z))}$ in $\Gamma(\mathcal{O}^{m\times m}/\mathcal{N}, \Omega)$, we have

$$\varphi\big(e^{G(z)} - D(z)\big) = 0 \quad \text{in } \Gamma\big(\mathcal{O}^{m\times m}/\mathcal{N}, \Omega\big); \tag{5.7}$$

therefore

$$e^{G(z)} - D(z) \in \Gamma(\mathcal{N}, \Omega). \tag{5.8}$$

That is, there exists an $H(z) \in \Gamma(\mathcal{O}^{m\times m}, \Omega)$ such that

$$e^{G(z)} = D(z) + N(z) \cdot H(z) \quad \text{on } \Omega. \tag{5.9}$$

Since $\det e^{G(z)} \neq 0$, approximating $H(z)$ by a complex polynomial matrix $X(z)$, we have

$$\det\big[D(z) + N(z) \cdot X(z)\big] \neq 0 \quad \text{on } \bar{U}^n.$$

This completes the proof. □

As an immediate consequence of this theorem, we have the following sufficient condition for strong stabilizability.

COROLLARY 5.2. *Let $D(z)^{-1}N(z)$ be a minor coprime MFD of a complex $m \times l$ plant $P(z)$, $l \geq m$. If there is some $m \times m$ submatrix $N_0(z)$ of $N(z)$, such that $D(z)$ has a single-valued logarithmic matrix function $\log D(z)$ on $V(\det N_0(z)) \cap \bar{U}^n$, then $P(z)$ is strongly stabilizable.*

In Example 1 of section 3.3,

$$[D\ N] = \begin{bmatrix} z_2^2 & 1 & z_1 & z_2 \\ -1 & z_3^2 & z_3 & 0 \end{bmatrix}.$$

It can be verified (e.g., by using the theory of operational calculus [7]; see Appendix B of this paper) that $D(z) = \left[\begin{smallmatrix} z_2^2 & 1 \\ -1 & z_3^2 \end{smallmatrix}\right]$ has a single-valued analytic logarithmic function over $V(\det N) \cap \bar{U}^3 = V(-z_2 z_3) \cap \bar{U}^3$.

If $D(z) = \operatorname{diag}(d_1(z), \ldots, d_m(z))$, a diagonal matrix, then its logarithmic function can be defined as

$$\log D(z) = \operatorname{diag}\big(\log d_1(z), \ldots, \log d_m(z)\big).$$

This yields the following simpler criterion for a strong stabilizability test for a class of plants that have a special form of minor coprime MFD.

COROLLARY 5.3. *Let $D(z)^{-1}N(z)$ be a minor coprime MFD of a complex $m \times l$ plant $P(z)$, $l \geq m$ with $D(z) = \operatorname{diag}\big(d_1(z), \ldots, d_m(z)\big)$, a diagonal matrix. If there is some $m \times m$ submatrix $N_0(z)$ of $N(z)$, such that, for each $i = 1, \ldots, m$, $d_i(z)$ has a single-valued logarithmic function $\log d_i(z)$ on $V(\det N_0(z)) \cap \bar{U}^n$; or equivalently, each $d_i(z)$ is 0-homotopic on $V(\det N_0(z)) \cap \bar{U}^n$, then $P(z)$ is strongly stabilizable.*

Example 5.

$$[D\ N] = \begin{bmatrix} 1-4z_1z_2 & 0 & z_2+1 & z_2 \\ 0 & 1 & z_1 & z_1 \end{bmatrix}.$$

$\det N = z_1$. The condition of Corollary 5.3 is satisfied. We actually have a stable compensator $C = YX^{-1}$ with

$$\begin{bmatrix} X \\ Y \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ 0 & 0 \\ 4z_1 & 0 \end{bmatrix}.$$

However, both conditions of the above corollaries are not necessary, as demonstrated by the following example.

Example 6.

$$F = [D\ N] = \begin{bmatrix} z_1 & 0 & 1 & 0 \\ 0 & z_2 & 0 & 1-4z_1z_2 \end{bmatrix}.$$

We have $\det D(z) + \frac{1}{4}\det N(z) = z_1z_2 + \frac{1}{4}(1-4z_1z_2) = \frac{1}{4}$. But it is not possible to define a single-valued $\log D(z)$ on $V(\det N(z)) \cap \bar{U}^2$ (see Appendix B). Nevertheless, for this plant, we have the following stable real compensator $C = YX^{-1}$ with

$$\begin{bmatrix} X \\ Y \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ 3z_1 & 1 \\ --1 & 0 \end{bmatrix}.$$

Actually, $\det\left([D\ N]\left[\begin{smallmatrix} X \\ Y \end{smallmatrix}\right]\right) = \det[DX+NY] = \det\left[\begin{smallmatrix} 4z_1 & 1 \\ -1+4z_1z_2 & z_2 \end{smallmatrix}\right] = 1$. Note that while $\det\left[\begin{smallmatrix} 4z_1 & 1 \\ -1+4z_1z_2 & z_2 \end{smallmatrix}\right] = 1$, and has a logarithmic function 0 on $\bar{U}^2$, $\left[\begin{smallmatrix} 4z_1 & 1 \\ -1+4z_1z_2 & z_2 \end{smallmatrix}\right]$ does not have a logarithmic function on $\bar{U}^2$, even though $\bar{U}^2$ is simply connected. In fact, it can be shown that $\left[\begin{smallmatrix} 4z_1 & 1 \\ 0 & z_2 \end{smallmatrix}\right]$ does not have a logarithmic function on $V(-1+4z_1z_2)\cap\bar{U}^2$ (see Appendix B). Theorem 5.1 actually requires the resulting matrix have a logarithmic function. This is stronger than what we need for strong stabilizability, as is illustrated by the above example.

5.2. Difficulties in an algebraic approach. Surely it is desirable to establish a method that not only gives a sufficient *and* necessary condition for strong stabilizability, but also constructs a stable compensator when it exists. However, even when an n-D MIMO system *is* strongly stabilizable, existing algebraic methods for constructing a compensator do not ensure the resulting compensator be stable itself. For instance, the method of [10], applied to Example 1 of section 3.3, goes as follows:

$$\alpha_1 + z_2z_3\alpha_6 = \det D + z_2z_3 \det N = 1.$$

Let D^{adj} and N^{adj} denote the adjoint matrices of D and N, respectively. Let

$$\begin{bmatrix} X \\ Y \end{bmatrix} = \begin{bmatrix} D^{adj} \\ \mathbf{0} \end{bmatrix} + z_2z_3 \begin{bmatrix} \mathbf{0} \\ N^{adj} \end{bmatrix} = \begin{bmatrix} z_3^2 & -1 \\ 1 & z_2^2 \\ 0 & -z_2^2z_3 \\ -z_2z_3^2 & z_1z_2z_3 \end{bmatrix},$$

$[DX + NY] = DD^{adj} + z_2 z_3 N N^{adj} = (\alpha_1 + z_2 z_3 \alpha_6)\left[\begin{smallmatrix} 1 & 0 \\ 0 & 1 \end{smallmatrix}\right] = \left[\begin{smallmatrix} 1 & 0 \\ 0 & 1 \end{smallmatrix}\right]$. $X^{-1}Y$ is the resulting compensator, which is itself unstable.

A possible approach by making use of the existing algebraic method for constructing a stable compensator for 1-D system is as follows. Let

$$D^{-1}N = D(z_1, z_2)^{-1} N(z_1, z_2)$$

be a 2-D system. Let

$$\alpha_1 = \det D, \quad \alpha_2, \ldots, \alpha_M$$

be the maximum order minors of [D N]. Assume there exist polynomials $\beta_2, \ldots, \beta_M$, such that

$$\det D + \beta_2 \alpha_2 + \cdots + \beta_M \alpha_M = u(z_1, z_2) \neq 0 \quad \forall (z_1, z_2) \in \bar{U}^2.$$

Viewing the entries of D and N, and $u(z_1, z_2)$ as elements of $K(z_1)[z_2]$, $K = \mathbf{C}$ or $\mathbf{R}$, the ring of polynomials in a single variable z_2 over the field of rational functions $K(z_1)$. Since $K(z_1)[z_2]$ is a Euclidean domain, the algorithm of [23] can be used to give a matrix $Y'(z_1, z_2)$ such that

$$\det \left[D(z_1, z_2) + N(z_1, z_2) Y'(z_1, z_2) \right] = u(z_1, z_2).$$

There $Y'(z_1, z_2)$ has entries belonging to $K(z_1)[z_2]$ and can be written as

$$Y'(z_1, z_2) = \frac{1}{d(z_1)} Y(z_1, z_2),$$

with $d(z_1)$ in $K[z_1]$ and the entries of $Y(z_1, z_2)$ in $K[z_1, z_2]$.

If $d(z_1)$ is a 1-D Hurwitz polynomial, then Y' is already a stable compensator. In general, $d(z_1)$ may have zeros in $\bar{U}^1$. In this case, if we have a method that can remove the zeros of $d(z_1)$ in $\bar{U}^1$ from $Y'(z_1, z_2)$, while keeping the resulting determinant $u(z_1, z_2)$ to be nonzero over $\bar{U}^2$, then we will be able to construct a stable compensator. Unfortunately, after much effort, we have not been able to find such a method.

6. Conclusion. We conclude this paper with open problems for future research.

1. For a finitely generated ideal J, the set $V(J) \cap \bar{U}^n$ can be viewed as a semi-algebraic set and its homology groups can be computed by the *cylindrical algebraic decomposition* [6, 1, 19]. For the case that the ideal J is generated by one polynomial, computational procedures for testing the criterion in Theorem 2.1 and the sign consistency condition in Theorem 2.5 have been given based on the cylindrical algebraic decomposition [25, 26]. The extension to the general case (J be generated by a finite number of polynomials) is straightforward.

Unfortunately, our method is not constructive in that it does not give a solution to inequality (2.1) in section 2. The development of an algorithmic method to solve inequality (2.1) remains open.

2. Though it is also possible to explore further advanced analysis method to fill the gap between the necessary conditions and the sufficient one presented in this paper, it is desirable to have an algebraic constructive method, as that for a 1-D MIMO system.

We believe that an adequate preparation has been made by our work for the final solution of the problem of the strong stabilizability of n-D MIMO systems, which is

most likely to be achieved by a fundamentally new algebraic approach that considers standard representations of MIMO n-D systems, analogous in function to that for conventional 1-D systems, such as the Smith–McMillan forms, which have played an essential role in stable compensator construction.

On the other hand, the application of the Euclidean algorithm has been indispensable to the construction of the Smith–McMillan forms, as well as to the construction of stable compensators for 1-D systems. Therefore a novel promising algebraic framework to address the above issue in linear MIMO n-D systems should be one in which adequate algorithms can be applied. In the recent behavioral approach to the study of n-D systems [18, 8, 29, 17], it seems relatively easy to apply the Gröbner basis algorithms, which are in a sense alternatives of Euclidean algorithm for dealing with algorithmic problems in multivariate polynomials and modules. However, we are not sure if this approach can be easily adapted for solving our problems.

To conclude, it is an interesting topic for future research to establish a new algebraic framework to solve old open problems, as the one remains unsolved in this paper.

Appendix A. Strong stabilizability of 1-D systems. Let $P(z)$ be a 1-D system with a minor coprime left MFD

$$P(z) = D^{-1}N = \begin{bmatrix} d_{11} & \cdots & d_{1m} \\ \vdots & \cdots & \vdots \\ d_{m1} & \cdots & d_{mm} \end{bmatrix}^{-1} \begin{bmatrix} n_{11} & \cdots & n_{1l} \\ \vdots & \cdots & \vdots \\ n_{m1} & \cdots & n_{ml} \end{bmatrix}.$$

Let $b(z)$ be the greatest common divisor of $n_{ij}(z), 1 \le i \le m, 1 \le j \le l$. Then $b(z)$ is the single generator of the ideal $J = (n_{ij}(z), 1 \le i \le m, 1 \le j \le l) \subset \mathbf{C}[z]$. $b(z)$ is equal to the *smallest invariant factor* of $P(z)$, which is defined as the smallest of the numerators when $P(z)$ is written in the Smith–McMillan form [23, p. 401]. Note that b can be chosen real if P is real.

Let $a(z) = \det D(z)$. The minor coprimeness of $D(z)^{-1}N(z)$ implies that $a(z)$ and $b(z)$ have no common zero. Since $V(b)$ is discrete, by Theorem 2.1, there exists a *complex* stable rational function $r(z)$ such that

$$a(z) + b(z)r(z) \neq 0 \quad \text{for } z \in \bar{U}^1.$$

The following lemma is easy to prove.

LEMMA A.1. (i) *Let*

$$S_{\boldsymbol{C}} = \{f(z)/g(z) : f(z),\ g(z) \in \boldsymbol{C}[z],\ g(z) \neq 0 \text{ on } \bar{U}^1\} \subset \boldsymbol{C}(z),$$

the ring of all stable complex rational functions. For each $s \in S_{\boldsymbol{C}}$, define the degree $\deg s$= *the number of zeros of s in $\bar{U}^1$. Then with this degree $S_{\boldsymbol{C}}$ is a Euclidean domain.*

(ii) *Let*

$$S_{\boldsymbol{R}} = S_{\boldsymbol{C}} \cap \boldsymbol{R}(z),$$

the ring of real stable rational functions. Then with the degree induced from that defined in $S_{\boldsymbol{C}}$, $S_{\boldsymbol{R}}$ is a Euclidean domain.

Based on the Euclidean property of $S_{\boldsymbol{C}}$, a complex stable rational matrix $R(z)$ can be constructed such that [23, section 4.4]

$$\det\big[D(z) + N(z)R(z)\big] = a(z) + b(z)r(z).$$

This means that P is stabilized by the stable compensator $R(z)$.

If $P = D^{-1}N$ is a real plant, the condition that $\det D$ has an invariant sign on $V(J) = V(b)$ guarantees the existence of a real stable rational function $r(z)$, such that

$$a(z) + b(z)r(z) \neq 0 \quad \text{for } z \in \bar{U}^1.$$

Using the Euclidean property of $S_{\mathbf{R}}$, one can construct a real stable rational matrix R such that

$$\det\big[D(z) + N(z)R(z)\big] = a(z) + b(z)r(z) \neq 0 \quad \text{for } z \in \bar{U}^1.$$

It is easy to see that $a(z) = \det D(z)$ is the least common denominator of the entries of $P(z)$ [23, p. 90] and that the zeros of the plant $P(z)$ are exactly the zeros of $b(z)$ [23, 27].

The condition that $a(z)$ has an invariant sign at the real zeros of $b(z)$ in $\bar{U}^1$ is equivalent to that the numbers of poles of $P(z)$ lying between two adjacent zeros on the interval $[-1, 1]$ are always even. This is the *parity interlacing property* on the interval $[-1, 1]$. By transforming the unit disc to the right halfplane, one obtains the same property on the right half real line [27, 25].

Appendix B. Operational calculus on matrix. For some z, consider $D(z)$ as a linear operator on a finite-dimensional linear space. Let $\sigma(D(z)) \subset \mathbf{C}^*$ denote the set of its eigenvalues. As an inverse of the exponential map, an analytic logarithmic function of $D(z)$ can be defined by the following integral:

$$\log D(z) = \frac{1}{2\pi i}\int_B \log(\lambda)\big(\lambda I - D(z)\big)^{-1} d\lambda, \tag{B.1}$$

where B is the boundary of some domain which contains the closure of some open set containing $\sigma(D(z))$ and consists of a finite number of closed rectifiable Jordan curves. Clearly, $\log D(z)$ is single valued if the domain with boundary B does not contain the origin in $\mathbf{C}^*$.

In Example 1, $D(z) = \left[\begin{smallmatrix} z_2^2 & 1 \\ -1 & z_3^2 \end{smallmatrix}\right]$ and $N(z) = \left[\begin{smallmatrix} z_1 & z_2 \\ z_3 & 0 \end{smallmatrix}\right]$. The eigenvalues of $D(z)$ are

$$\lambda = \frac{1}{2}\left(z_2^2 + z_3^3 \pm \sqrt{\big(z_2^2 - z_3^3\big)^2 - 4}\right). \tag{B.2}$$

It is easy to see that the union of the sets of eigenvalues $\cup_{z \in V(z_2 z_3) \cap \bar{U}^3} \sigma(D(z))$ does not intersect with the real axis, and can be contained in a simply connected open domain not containing the origin in $\mathbf{C}^*$. With B being the boundary of this domain, $\log D(z)$ defined in (B.1) is a single-valued analytic logarithmic function in z because λ in (B.2) varies analytically in z.

On the other hand, if $\log D(z)$ is defined as a single-valued matrix function, then by the *spectral mapping theorem* [7, p. 569], the eigenvalues of $D(z)$ have single-valued logarithms which are the eigenvalues of $\log D(z)$.

In Example 6, the eigenvalues of $D(z)$ are z_1 and z_2, which are not 0-homotopic on $V(\det N(z)) \cap \bar{U}^2 = V(1 - 4z_1 z_2) \cap \bar{U}^2$. Therefore $D(z)$ does not have a single-valued logarithm on $V(\det N(z)) \cap \bar{U}^2$. For the same reason, $\left[\begin{smallmatrix} 4z_1 & 1 \\ -1+4z_1 z_2 & z_2 \end{smallmatrix}\right]$ does not have a logarithm on $V(1 - 4z_1 z_2) \cap \bar{U}^2$.

Acknowledgments. I am grateful to L. Xu, S. Shankar, and N. K. Bose for their valuable discussions during this work. I am indebted to T. Mandai for important

discussions concerning section 5.1 of this paper. I would like to thank Z. Lin for his valuable suggestions and help during the writing of this paper. I would also like to thank the editor and reviewers for their comments which have greatly improved the presentation of this paper.

REFERENCES

[1] D. S. Arnon, G. E. Collins, and S. McCallum, *Cylindrical algebraic decomposition* I: *The basic algorithm*, SIAM J. Comput., 13 (1984), pp. 865–877; *Cylindrical algebraic decomposition* II: *An adjacency algorithm for the plane*, pp. 878–889.

[2] V. Blondel and M. Gevers, *Simultaneous stabilizability of three linear systems is rationally undecidable*, Math. Control Signals Systems, 6 (1993), pp. 135–145.

[3] V. Blondel, M. Gevers, R. Mortini, and R. Rupp, *Simultaneous stabilization of three or more plants: Conditions on the positive real axis do not suffice*, SIAM J. Control Optim., 32 (1994), pp. 572–590.

[4] N. K. Bose, J. P. Guiver, E. W. Kamen, H. M. Valenzuela, and B. Buchberger, *Multidimensional Systems Theory: Progress, Directions and Open Problems in Multidimensional Systems*, Math. Appl. 16, D. Reidel, Dordrecht, The Netherlands, Boston, MA, 1985.

[5] B. Buchberger, *Gröbner bases: An algorithmic method in polynomial ideal theory*, in Multidimensional Systems Theory: Progress, Directions and Open Problems in Multidimensional Systems, N. K. Bose, J. P. Guiver, E. W. Kamen, H. M. Valenzuela, and B. Buchberger, eds., D. Reidel, Dordrecht, The Netherlands, Boston, MA, 1985, pp. 184–232.

[6] G. E. Collins, *Quantifier elimination for real closed fields by cylindrical algebraic decomposition*, in Automation Theory and Formal Languages, Lecture Notes in Comput. Sci. 33, Springer-Verlag, Berlin, 1975, pp. 134–183.

[7] N. Dunford and J. T. Schwartz, *Linear Operators, Part* I, Interscience, New York, 1958.

[8] E. Fornasini and M. E. Valcher, *Stability and stabilizability of* 2*D behaviors*, in Proceedings of the 36th IEEE Conference on Decision and Control, San Diego, CA, 1997, pp. 2590–2595.

[9] W. Fulton, *Algebraic Topology: A First Course*, Springer-Verlag, New York, 1995.

[10] J. P. Guiver, *The equation AX = b over the ring C[z,w]*, in Multidimensional Systems Theory: Progress, Directions and Open Problems in Multidimensional Systems , N. K. Bose, J. P. Guiver, E. W. Kamen, H. M. Valenzuela, and B. Buchberger, eds., D. Reidel, Dordrecht, The Netherlands, Boston, MA, 1985.

[11] J. P. Guiver and N. K. Bose, *Causal and weakly causal* 2*-D filters with applications in stabilization*, in Multidimensional Systems Theory: Progress, Directions and Open Problems in Multidimensional Systems, N. K. Bose, J. P. Guiver, E. W. Kamen, H. M. Valenzuela, and B. Buchberger, eds., D. Reidel, Dordrecht, The Netherlands, Boston, MA, 1985, pp. 52–100.

[12] R. C. Gunning and H. Rossi, *Analytic Functions of Several Complex Variables*, Prentice–Hall, Englewood Cliffs, NJ, 1965.

[13] L. Hörmander, *An Introduction to Complex Analysis in Several Variables*, 3rd ed., North–Holland, Amsterdam, 1990.

[14] Z. Lin, *On matrix fraction descriptions of multivariable Linear n-D systems*, IEEE Trans. Circuits Systems, 35 (1988), pp. 1317–1322.

[15] Z. Lin, *Feedback stabilizability of MIMO n-D linear systems*, Multidimens. Systems Signal Process., 9 (1998), pp. 149–172.

[16] K. Mori and K. Abe, *Feedback stabilization over commutative rings: Two-stage feedback stabilization approach*, in Proceedings of the 36th IEEE Conference on Decision and Control, San Diego, CA, 1997, pp. 324–332.

[17] H. Pillai and S. Shankar, *A behavioral approach to control of distributed systems*, SIAM J. Control Optim., 37 (1999), pp. 388–408.

[18] P. Rocha and J. C. Willems, *Controllability of* 2*D Systems*, IEEE Trans. Automat. Control, 36 (1991), pp. 413–423.

[19] J. T. Schwartz and M. Sharir, *On the piano mover's problem* II: *General techniques for computing topological properties of real algebraic manifolds*, Adv. Appl. Math., 4 (1983), pp. 298–351.

[20] S. Shankar, *An obstruction to the simultaneous stabilization of two n-D plants*, Acta Appl. Math., 36 (1994), pp. 289–301.

[21] S. Shankar and V. R. Sule, *Algebraic geometric aspects of feedback stabilization*, SIAM J. Control Optim., 30 (1992), pp. 11–30.

[22] V. R. Sule, *Feedback stabilization over commutative rings: The matrix case*, SIAM J. Control

Optim., 32 (1994), pp. 1675–1695.
[23] M. Vidyasagar, *Control System Synthesis: A Factorization Approach*, MIT Press, Cambridge, MA, 1985.
[24] L. Xu, O. Saito, and K. Abe, *Output feedback stabilizability and stabilization algorithms for 2-D systems*, Multidimens. Systems Signal Process., 5 (1994), pp. 41–60.
[25] J. Q. Ying, *Conditions for strong stabilizabilities of n-d imensional systems*, Multidimens. Systems Signal Process., 9 (1998), pp. 125–148.
[26] J. Q. Ying and L. Xu, *Procedures for testing for the strong stabilizabilities of n-D linear systems*, in Proceedings of the 36th IEEE Conference on Decision and Control, San Diego, CA, 1997, pp. 337–338.
[27] D. C. Youla, J. J. Bongiorno, Jr., and C. N. Lu, *Single loop feedback stabilization of linear multivariable plants*, Automatica J. IFAC, 10 (1974), pp. 159–173.
[28] D. C. Youla and G. Gnavi, *Notes on n-dimensional system theory*, IEEE Trans. Circuits Systems, CAS-26 (1979), pp. 105–111.
[29] S. Zampieri, *Causal input/output representation of 2D systems in the behavioral approach*, SIAM J. Control Optim., 36 (1998), pp. 1133–1146.

SIAM J. CONTROL OPTIM.
Vol. 38, No. 2, pp. 337–383

BOUNDARY CONTROLLABILITY OF THERMOELASTIC PLATES VIA THE FREE BOUNDARY CONDITIONS*

GEORGE AVALOS† AND IRENA LASIECKA‡

Abstract. Controllability properties of a partial differential equation (PDE) model describing a thermoelastic plate are studied. The PDE is composed of a Kirchoff plate equation coupled to a heat equation on a bounded domain, with the coupling taking place on the interior and boundary of the domain. The coupling in this PDE is parameterized by $\alpha > 0$. Boundary control is exerted through the (two) free boundary conditions of the plate equation and through the Robin boundary condition of the temperature. These controls have the physical interpretation of inserted forces and moments and prescribed temperature, respectively, all of which act on the edges of the plate. The main result here is that under such boundary control, and with initial data in the basic space of well-posedness, one can simultaneously control the displacement of the plate *exactly* and the temperature *approximately*. Moreover, the thermal control may be taken to be arbitrarily smooth in time and space, and the thermal control region may be any nonempty subset of the boundary. This controllability holds for arbitrary values of the coupling parameter α, with the optimal controllability time in line with that seen for uncoupled Kirchoff plates.

Key words. partial differential equations, exact-approximate controllability

AMS subject classification. 35B37

PII. S0363012998339836

1. Introduction.

1.1. Statement of the problem. Throughout, Ω will be a bounded open subset of $\mathbb{R}^2$ with sufficiently smooth boundary $\Gamma = \overline{\Gamma_0} \cup \overline{\Gamma_1}$, with both Γ_0 and Γ_1 being open, with Γ_0 being possibly empty, and satisfying $\overline{\Gamma_0} \cap \overline{\Gamma_1} = \varnothing$. Furthermore, Γ_2 will be any open and nonempty subset of Γ_1. With this geometry, we shall consider here the following thermoelastic system on finite time $(0, T)$:

$$
\left\{
\begin{array}{l}
\left\{ \begin{array}{l} \omega_{tt} - \gamma \Delta \omega_{tt} + \Delta^2 \omega + \alpha \Delta \theta = 0 \\ \beta \theta_t - \eta \Delta \theta + \sigma \theta - \alpha \Delta \omega_t = 0 \end{array} \right. \quad \text{on } (0,T) \times \Omega; \\
\\
\omega = \dfrac{\partial \omega}{\partial \nu} = 0 \ \text{ on } (0,T) \times \Gamma_0; \\
\\
\left\{ \begin{array}{l} \Delta \omega + (1-\mu) B_1 \omega + \alpha \theta = u_1 \\ \dfrac{\partial \Delta \omega}{\partial \nu} + (1-\mu) \dfrac{\partial B_2 \omega}{\partial \tau} - \gamma \dfrac{\partial \omega_{tt}}{\partial \nu} + \alpha \dfrac{\partial \theta}{\partial \nu} = u_2 \end{array} \right. \quad \text{on } (0,T) \times \Gamma_1; \\
\\
\dfrac{\partial \theta}{\partial \nu} + \lambda \theta = \left\{ \begin{array}{ll} u_3 & \text{on } (0,T) \times \Gamma_2, \\ 0 & \text{on } (0,T) \times \Gamma \backslash \Gamma_2, \end{array} \right. \qquad \lambda \geq 0; \\
\\
\omega(t=0) = \omega_0,\ \omega_t(t=0) = \omega_1,\ \theta(t=0) = \theta_0 \quad \text{on } \Omega.
\end{array}
\right. \tag{1.1}
$$

*Received by the editors June 5, 1998; accepted for publication (in revised form) April 27, 1999; published electronically January 11, 2000.

http://www.siam.org/journals/sicon/38-2/33983.html

†Department of Mathematics and Statistics, Texas Tech University, Lubbock, TX 79409 (avalos@math.ttu.edu). The research of this author was partially supported by NSF grant DMS-9710981.

‡Department of Mathematics, Thornton Hall, University of Virginia, Charlottesville, VA 22903 (il2v@amsun.apma.virginia.edu). The research of this author was partially supported by NSF grant DMS-9804822 and Army Research Office grant DAAH04-98-1-0059.

Here, α, β, η, and σ are positive constants. The positive constant γ is proportional to the thickness of the plate and assumed to be small with $0 < \gamma \leq M$. The boundary operators B_i are given by

$$B_1\omega \equiv 2\nu_1\nu_2\frac{\partial^2\omega}{\partial x\partial y} - \nu_1^2\frac{\partial^2\omega}{\partial y^2} - \nu_2^2\frac{\partial^2\omega}{\partial x^2} \text{ and } B_2\omega \equiv (\nu_1^2 - \nu_2^2)\frac{\partial^2\omega}{\partial x\partial y} + \nu_1\nu_2\left(\frac{\partial^2\omega}{\partial y^2} - \frac{\partial^2\omega}{\partial x^2}\right). \tag{1.2}$$

The constant $\mu \in (0, \frac{1}{2})$ is the familiar Poisson ratio, and $\nu = [\nu_1, \nu_2]$ denotes the outward unit normal to the boundary. Here and throughout we shall make the following geometric assumption on the (uncontrolled) portion of the boundary Γ_0:

$$\text{with } \overline{h}(x,y) \equiv [x - x_0, y - y_0], \exists\ \{x_0, y_0\} \in \mathbb{R}^2 \text{ such that } \overline{h}(x,y)\cdot\nu \leq 0 \text{ on } \Gamma_0. \tag{1.3}$$

The PDE model (1.1), with boundary functions $u_1 = u_2 = 0$ and $u_3 = 0$, mathematically describes an uncontrolled Kirchoff plate subjected to a thermal damping, with the displacement of the plate represented by the function $\omega(t,x,y)$ and the temperature given by the function $\theta(t,x,y)$ (see [11] for a derivation of this model). The given control variables $u_1(t,x)$ and $u_2(t,x)$ are defined on the portion of the boundary $(0,T)\times\Gamma_1$; the control $u_3(t,x)$ is defined on $(0,T)\times\Gamma_2$.

Making the denotation

$$H^k_{\Gamma_0}(\Omega) \equiv \left\{\varpi \in H^k(\Omega) : \left.\frac{\partial^j\varpi}{\partial\nu^j}\right|_{\Gamma_0} = 0 \text{ for } j = 0, ..., k-1\right\}, \tag{1.4}$$

we will throughout take the initial data $[\omega_0, \omega_1, \theta_0]$ to be in $H^2_{\Gamma_0}(\Omega)\times H^1_{\Gamma_0}(\Omega)\times L^2(\Omega)$. For initial data in these spaces and controls $u_1 = u_2 = 0$ and $u_3 = 0$, one can show the well-posedness of (1.1) with the corresponding solution $[\omega, \omega_t, \theta]$ being in $C([0,T]; H^2_{\Gamma_0}(\Omega)\times H^1_{\Gamma_0}(\Omega)\times L^2(\Omega))$ (see, e.g., [11] and [2]). In this paper, we will study controllability properties of solutions of (1.1) under the influence of boundary control functions in the following spaces:

$$[u_1, u_2, u_3] \in L^2(0,T; L^2(\Gamma_1)\times H^{-1}(\Gamma_1))\times C^r(\Sigma_{2,T}), \text{ where } r > 0 \text{ and } \Sigma_{2,T} = (0,T)\times\Gamma_2. \tag{1.5}$$

For arbitrary $[u_1, u_2, u_3]$ of such smoothness, the corresponding solution $[\omega, \omega_t, \theta]$ will be in the "large" space $C([0,T]; [D(\mathcal{A}^*_\gamma)]')$ (see the definition of $D(\mathcal{A}^*_\gamma)$ in (1.49)). In particular, we intend to address, on the finite time interval $[0,T]$, the question of *exact-approximate controllability* (this term being originally coined in [6]). That is to say, for given data $[\omega_0, \omega_1, \theta_0]$ (initial) and $\left[\omega_0^T, \omega_1^T, \theta_0^T\right]$ (terminal) in $H^2_{\Gamma_0}(\Omega)\times H^1_{\Gamma_0}(\Omega)\times L^2(\Omega)$, and arbitrary $\epsilon > 0$, is there a suitable control triple $[u_1, u_2, u_3] \in L^2(0,T; L^2(\Gamma_1)\times H^{-1}(\Gamma_1))\times C^r(\Sigma_{2,T})$ such that the corresponding solution $[\omega, \omega_t, \theta]$ of (1.1) satisfies the following steering property at terminal time T:

$$[\omega(T), \omega_t(T)] = \left[\omega_0^T, \omega_1^T\right] \text{ and } \left\|\theta(T) - \theta_0^T\right\|_{L^2(\Omega)} \leq \epsilon?$$

In this regard, we post our main result here for which we need the number

$$T^* \equiv 2\sqrt{\gamma}\cdot\max\left\{\sqrt{\frac{2}{1-\mu}}\max_{[x,y]\in\overline{\Omega}}\left|\overline{h}(x,y)\right|, \sup_{[x,y]\in\Omega} d\left([x,y],\Gamma_2\right)\right\}, \tag{1.6}$$

where, above, $d([x,y],\Gamma_2)$ denotes the distance between $[x,y]$ and Γ_2.

THEOREM 1.1. *Let assumptions* (1.3) *and* (1.6) *stand. Then for* $T > T^*$, *the following controllability property holds true: For given initial data* $[\omega_0, \omega_1, \theta_0]$ *and terminal data* $\left[\omega_0^T, \omega_1^T, \theta_0^T\right]$ *in the space* $H^2_{\Gamma_0}(\Omega) \times H^1_{\Gamma_0}(\Omega) \times L^2(\Omega)$, *and arbitrary* $\epsilon > 0$, *one can find control functions* $[u_1^*, u_2^*, u_3^*] \in L^2(0,T; L^2(\Gamma_1) \times H^{-1}(\Gamma_1)) \times C^r(\Sigma_{2,T})$ *(where arbitrary* $r \geq 0$*) such that the corresponding solution* $[\omega^*, \omega_t^*, \theta^*]$ *to* (1.1) *satisfies at terminal time* T,

$$[\omega^*(T), \omega_t^*(T)] = \left[\omega_0^T, \omega_1^T\right],$$

$$\left\|\theta^*(T) - \theta_0^T\right\|_{L^2(\Omega)} < \epsilon.$$

Theorem 1.1 is almost a corollary from the following controllability result for the mechanical variable only, which comprises the bulk of our effort here.

THEOREM 1.2. *With the coupling parameter* α *in* (1.1) *being arbitrary, and* (1.3), (1.6) *in place, then for* $T > T^*$, *the following property holds true: For all initial data* $[\omega_0, \omega_1, \theta_0] \in H^2_{\Gamma_0}(\Omega) \times H^1_{\Gamma_0}(\Omega) \times L^2(\Omega)$ *and terminal data* $\left[\omega_0^T, \omega_1^T\right] \in H^2_{\Gamma_0}(\Omega) \times H^1_{\Gamma_0}(\Omega)$, *there exists* $[u_1, u_2, u_3] \in L^2(0,T; L^2(\Gamma_1) \times H^{-1}(\Gamma_1)) \times H^s(\Sigma_{2,T})$, *where arbitrary* $s \geq 0$, *such that the corresponding solution* $[\omega, \omega_t, \theta]$ *to* (1.1) *satisfies* $[\omega(T), \omega_t(T)] = \left[\omega_0^T, \omega_1^T\right]$.

Remark 1.3. Note that the point $[x_0, y_0]$ can be selected in such a way so that $2 \max_{[x,y] \in \overline{\Omega}} \left|\overline{h}(x,y)\right| \leq diam\,(\Omega)$, and so, ultimately, T^* in (1.6) can be rechosen as $T^* = 2\sqrt{\gamma}\ diam\,(\Omega)$.

Remark 1.4. Note that in our statement of controllability, no geometric conditions are imposed on the controlled region of the boundary Γ_1 , only on the (possibly void) boundary portion Γ_0.

1.2. Literature. To date, the only work dealing with the *boundary* control of thermoelastic plates, in dimension greater than one, had been that of J. Lagnese in [12] (indeed, this present paper is principally motivated by [12]). In this paper, Lagnese shows that if the coupling parameter α is small enough and the boundary Γ is "star shaped," then the boundary controlled system (1.1) is (partially) exactly controllable with respect to the mechanical variables $[\omega, \omega_t]$. Also in [22], a boundary-controlled system of thermoelastic waves is studied, with a coupling parameter α likewise present therein, and a result of partial exact controllability for this PDE is cited (again, controllability with respect to the hyperbolic component). This controllability result is quoted in [22] to be valid for all sizes of α; however, in the erratum [23], the author of [22] has acknowledged a flaw in the controllability proof, the correction of which will necessitate a smallness criterion on α. Ultimately, then, the paper [22] produces a controllability result if the coupling parameter is small enough, a result in the style of [12]. The chief contribution of the present paper is to remove restrictions on the size of α (see Theorem 1.2 above). For a one-dimensional version of (1.1), S. Hansen and B. Zhang in [8], via a moment problem approach, show the system's exact null controllability with boundary control in either the plate or the thermal component.

Other controllability results for the thermoelastic system, which do not assume any "smallness" condition on the coupling parameters, involve the implementation of *distributed/internal* controls subject to clamped or hinged boundary conditions. These results include that in [6], in which interior control is placed in the Kirchoff plate component subject to *clamped* boundary conditions.

With such control, one obtains exact controllability for the plate $[\omega, \omega_t]$ and approximate controllability for the temperature θ (i.e., exact-approximate controllability). In addition, the work in [19] deals with obtaining a result of null controllability for a linear system of thermoelasticity, in which both the hyperbolic and the parabolic components can be driven to zero by means of interior control placed in the hyperbolic (wave) component.

Another result of internal control for the thermoelastic PDE (1.1) is in [5], wherein interior control is placed in the heat equation only (i.e., $\beta\theta_t - \eta\Delta\theta + \sigma\theta - \alpha\Delta\omega_t = u$) so as to obtain exact controllability for *both* components ω and θ. The novelty of this result is that this (total) exact controllability obtains for all values of the rotational inertia parameter $\gamma \geq 0$: in the limiting case $\gamma = 0$, one is then presented with a result of exact controllability for a PDE modeled by the generator of an analytic semigroup (see [18]). This controllability holds for all values of α.

Again, the main contribution of this paper is that we consider *boundary* controls acting via the higher order *free* mechanical boundary conditions, and we do not assume any size restriction on the coupling parameter α. Moreover, we do not impose any geometric "star-shaped" conditions on the controlled portion of the geometry.

At this point, we attempt to compare the degree of difficulty in obtaining controllability results for thermoelastic plates under mechanical interior control with lower-order mechanical boundary conditions enforced (such as *clamped* or *hinged*), versus that involved in the present study, where, again, boundary control is exerted upon the second and third order free boundary conditions. This comparison is appropriate, since the novelty of our work is touted to be (mechanical) exact controllability for the PDE (1.1), whatever α may be; and excluding the paper [5], the only other available controllability results for thermoelastic systems, which require no size constraints on α, concerned thermoelastic systems under (distributed) interior mechanical control and with lower mechanical boundary conditions in place.

An underlying strategy in control theoretic studies of thermoelastic plates has been to exploit, if possible, previously known controllability results for (uncoupled) Kirchoff plates. To this end, one attempts to treat the thermoelastic system as a sort of perturbation of the Kirchoff plate. It is well known that if the underlying controllability map can be decomposed into the sum of a compact map and a surjective controllability map, corresponding to a (simpler) subcomponent of the PDE system, then the exact controllability of the original problem is equivalent to its approximate controllability. This favorable scenario occurs in equations of thermoelasticity with either *clamped* or *hinged* boundary conditions and *interior*, *distributed* controls (see, e.g., [20]). Indeed, the part of the simpler component is played by the classical and much-studied Kirchoff plate, for which many results on exact controllability are already available in the literature. Taking the boundary conditions to be clamped or hinged allows for a known structural decomposition of the thermoelastic system into a group (associated with the Kirchoff plate) and a compact perturbation. Combining this decomposition with the boundedness of interior control actions immediately yields the desired decomposition of the original controllability map into the sum of a surjective controllability map (corresponding to the Kirchoff plate) and a compact perturbation. This popular strategy was used in [6], where an *exact-approximate* controllability result was established for the thermoelastic system with clamped homogeneous boundary conditions and internal controls.

The situation is drastically different in the present paper, involving the case of *boundary controls*. Here, in this case of *free* mechanical boundary conditions, the

corresponding controllability operator *cannot* be taken to be a compact perturbation of the controllability map for the (uncoupled) boundary-controlled Kirchoff plate. In the first place, the associated input→state space map, defined explicitly in (1.42), is an inherently *unbounded* operator with respect to the natural energy space (see [17] for recent sharp regularity results for corresponding solutions, which are still, however, below the level of energy). Moreover, in the present case of *free* boundary conditions, there is a decomposition of the underlying thermoelastic semigroup, but it is into the sum of a Kirchoff plate semigroup and an *unbounded*—not compact—operator (see [16]). This complication is due to the fact that the Lopatinski conditions are not satisfied for the Kirchoff model under free boundary conditions, and to the intrinsic nature of the coupling between the mechanical and thermal variables within the free boundary conditions. These two complications above, again an artifact of the "free case," explain why there have been so few results regarding the boundary control of thermoelastic plates and why a "decoupling" of the thermoelastic PDE into a sole Kirchoff plate can only go so far.

Our goal here is to dispense with this smallness assumption and, in addition, show that a control can be constructed that provides exact controllability of the mechanical variables and approximate controllability of the thermal component. We note that the thermal control u_3 present in (1.1)—wholly absent in [12]—plays no part at all in the removal of the size restriction on α; it is in place only to exploit, in a compactness-uniqueness argument, recently obtained approximate controllability properties of the thermoelastic plate under the action of boundary control in the free mechanical boundary conditions (see [10]). At this point in time, the thermoelastic system cannot be shown to be approximately controllable with control in the free boundary conditions only (and no thermal control). Therefore, the presence of the thermal boundary control here is not an artificiality; it appears to be necessary for approximate controllability. (We do not know if the future will bring a unique continuation result for the thermoelastic plate in the absence of the thermal component.) However, the result of Theorem 1.1 says that the thermal control may be taken to be very smooth and with arbitrarily small support Γ_2. Again, this benign situation is a consequence of our employing thermal control at the compactness-uniqueness level only; it plays no part whatsoever in generating the main observability estimate (estimate (2.5) of Theorem 2.1), this being free of any size restrictions on α.

The strategy adopted in this paper consists of the following steps. Initially, a suitable transformation of variables is made and applied to (1.1); subsequently, a multiplier method is invoked with respect to the transformed equation. The mulitiplers employed here are the differential multipliers used in the study of exact controllability for the Kirchoff plate model (inspired by [11]), together with the nonlocal (ΨDO) multipliers used in the study of thermoelastic plates in [3] and [4]. The controllability time T^* in Theorem 1.1 ultimately depends in part upon the radial vector field associated with the differential Kirchoff multipliers (see Lemma 2.5 below). This multiplier method allows the attainment of preliminary estimates for the energy of the system. However, these estimates are "polluted" by certain boundary terms that are not majorized by the energy. To cope with these, we use the sharp trace estimates established in [15] for Kirchoff plates. The use of this PDE result introduces lower order terms into the energy estimate, which are eventually eliminated with the help of a new unique continuation result in [10]. It is *only* at the level of invoking this uniqueness result that the thermal control u_3 on Γ_2 must be introduced. The controllability time T^* in (1.6) is optimal.

1.3. Operator theoretic formulation and analysis.

1.3.1. Preliminary definitions. In obtaining our controllability result Theorem 1.1, it will be useful to consider the PDE system (1.1) as an abstract evolution equation in a certain Hilbert space, to which end we introduce the following definitions and notation.

- With $H^k_{\Gamma_0}(\Omega)$ as defined in (1.4), we define $\mathring{\mathbf{A}}$: $L^2(\Omega) \supset D\left(\mathring{\mathbf{A}}\right) \to L^2(\Omega)$ to be $\mathring{\mathbf{A}} = \Delta^2$, with domain

$$D(\mathring{\mathbf{A}}) = \Big\{\omega \in H^4(\Omega) \cap H^2_{\Gamma_0}(\Omega) : \Delta\omega + (1-\mu)B_1\omega = 0 \text{ on } \Gamma_1 \text{ and } \frac{\partial \Delta\omega}{\partial \nu} + (1-\mu)\frac{\partial B_2\omega}{\partial \tau} = 0 \text{ on } \Gamma_1\Big\}. \tag{1.7}$$

- $\mathring{\mathbf{A}}$ is then positive definite and self-adjoint, and consequently from [7] we have the characterizations

$$\begin{aligned} D\left(\mathring{\mathbf{A}}^{\frac{1}{4}}\right) &= H^1_{\Gamma_0}(\Omega), \\ D\left(\mathring{\mathbf{A}}^{\frac{1}{2}}\right) &= H^2_{\Gamma_0}(\Omega), \\ D\left(\mathring{\mathbf{A}}^{\frac{3}{4}}\right) &= \left\{\omega \in H^3(\Omega) \cap H^2_{\Gamma_0}(\Omega) : \Delta\omega + (1-\mu)B_1\omega = 0 \text{ on } \Gamma_1\right\}. \end{aligned} \tag{1.8}$$

Note that without loss of generality, we are here taking Γ_0 to be nonempty in order to have the equivalence of the $H^2(\Omega)$ norm with that induced by the $D(\mathring{\mathbf{A}}^{\frac{1}{2}})$. In the case that $\Gamma_0 = \emptyset$, we would simply modify $D(\mathring{\mathbf{A}})$ by enforcing $\frac{\partial \Delta\omega}{\partial \nu} + (1-\mu)\frac{\partial B_2\omega}{\partial \tau}|_{\Gamma_1} = \omega|_{\Gamma_1}$ (instead of $\frac{\partial \Delta\omega}{\partial \nu} + (1-\mu)\frac{\partial B_2\omega}{\partial \tau}|_{\Gamma_1} = 0$ in (1.7)). This modification would not change the problem.

Moreover, using Green's formula in [11], we have that for ω, $\widehat{\omega}$ "smooth enough,"

$$\begin{aligned} \int_\Omega (\Delta^2\omega)\widehat{\omega}d\Omega = a(\omega,\widehat{\omega}) + \int_\Gamma \left[\frac{\partial \Delta\omega}{\partial \nu} + (1-\mu)\frac{\partial B_2\omega}{\partial \tau}\right]\widehat{\omega}d\Gamma \\ - \int_\Gamma \left[\Delta\omega + (1-\mu)B_1\omega\right]\frac{\partial \widehat{\omega}}{\partial \nu}d\Gamma, \end{aligned} \tag{1.9}$$

where $a(\cdot,\cdot)$ is defined by

$$a(\omega,\widehat{\omega}) \equiv \int_\Omega \left[\omega_{xx}\widehat{\omega}_{xx} + \omega_{yy}\widehat{\omega}_{yy} + \mu\left(\omega_{xx}\widehat{\omega}_{yy} + \omega_{yy}\widehat{\omega}_{xx}\right) + 2(1-\mu)\omega_{xy}\widehat{\omega}_{xy}\right]d\Omega. \tag{1.10}$$

In particular, this formula and the second characterization in (1.8) give that for all ω, $\widehat{\omega} \in D(\mathring{\mathbf{A}}^{\frac{1}{2}})$,

$$\begin{aligned} \left\langle \mathring{\mathbf{A}}\omega, \widehat{\omega}\right\rangle_{\left[D\left(\mathring{\mathbf{A}}^{\frac{1}{2}}\right)\right]' \times D\left(\mathring{\mathbf{A}}^{\frac{1}{2}}\right)} &= \left(\mathring{\mathbf{A}}^{\frac{1}{2}}\omega, \mathring{\mathbf{A}}^{\frac{1}{2}}\widehat{\omega}\right)_{L^2(\Omega)} = a(\omega,\widehat{\omega})_{L^2(\Omega)}, \\ \|\omega\|^2_{D\left(\mathring{\mathbf{A}}^{\frac{1}{2}}\right)} &= \left\|\mathring{\mathbf{A}}^{\frac{1}{2}}\omega\right\|^2_{L^2(\Omega)} = a(\omega,\omega). \end{aligned} \tag{1.11}$$

- We define $A_D : L^2(\Omega) \supset D(A_D) \to L^2(\Omega)$ to be $A_D = -\Delta$, with Dirichlet boundary conditions, viz.,

$$D(A_D) = H^2(\Omega) \cap H^1_0(\Omega). \tag{1.12}$$

A_D is also positive definite, self-adjoint, and by [7]

$$D\left(A_D^{\frac{1}{2}}\right) = H_0^1(\Omega). \tag{1.13}$$

- We denote the operator $A_R : L^2(\Omega) \supset D(A_R) \to L^2(\Omega)$ by the following second order elliptic operator:

$$A_R = -\Delta + \frac{\sigma}{\eta}\mathbf{I}, \text{ with } D(A_R) = \left\{\vartheta \in H^2(\Omega) : \frac{\partial \vartheta}{\partial \nu} + \lambda\vartheta = 0\right\}. \tag{1.14}$$

A_R is self-adjoint, positive definite on $L^2(\Omega)$, with its fractional powers therefore being well defined. In particular, we have again by [7] that for $s \in \left[0, \frac{3}{4}\right)$,

$$\begin{aligned} &D(A_R^s) = H^{2s}(\Omega), \\ &\left(\vartheta, \widetilde{\vartheta}\right)_{H^1(\Omega)} = \left(A_R^{\frac{1}{2}}\vartheta, A_R^{\frac{1}{2}}\widetilde{\vartheta}\right)_{L^2(\Omega)} \\ &= \left(\nabla\vartheta, \nabla\widetilde{\vartheta}\right)_{L^2(\Omega)} + \lambda\left(\vartheta, \widetilde{\vartheta}\right)_{L^2(\Gamma)} + \frac{\sigma}{\eta}\left(\vartheta, \widetilde{\vartheta}\right)_{L^2(\Omega)}. \end{aligned} \tag{1.15}$$

- We denote the operator $A_N : L^2(\Omega) \supset D(A_N) \to L^2(\Omega)$ by the following second order elliptic operator:

$$A_N = -\Delta, \text{ with } D(A_N) = \left\{\vartheta \in H^2(\Omega) : \vartheta|_{\Gamma_0} = \left.\frac{\partial \vartheta}{\partial \nu}\right|_{\Gamma_1} = 0\right\}. \tag{1.16}$$

Once again by [7], we have for $s \in \left(\frac{1}{4}, \frac{3}{4}\right)$

$$D(A_N^s) = \left\{\vartheta \in H^{2s}(\Omega) \text{ such that } \vartheta|_{\Gamma_0} = 0\right\}. \tag{1.17}$$

- (γ_0, γ_1) will denote the classical Sobolev trace maps, which yield for $f \in C^\infty(\overline{\Omega})$

$$\gamma_0 f = f|_\Gamma\,;\, \gamma_1 f = \left.\frac{\partial f}{\partial \nu}\right|_\Gamma. \tag{1.18}$$

- We define the elliptic operators G_1, G_2, and D as follows:

$$G_1 h = v \Longleftrightarrow \begin{cases} \Delta^2 v = 0 \quad \text{on } \Omega, \\ v = \dfrac{\partial v}{\partial \nu} = 0 \quad \text{on } \Gamma_0, \\ \begin{cases} \Delta v + (1-\mu)B_1 v = h \\ \dfrac{\partial \Delta v}{\partial \nu} + (1-\mu)\dfrac{\partial B_2 v}{\partial \tau} = 0 \end{cases} \quad \text{on } \Gamma_1, \end{cases} \tag{1.19}$$

$$G_2 h = v \Longleftrightarrow \begin{cases} \Delta^2 v = 0 \quad \text{on } \Omega, \\ v = \dfrac{\partial v}{\partial \nu} = 0 \quad \text{on } \Gamma_0, \\ \begin{cases} \Delta v + (1-\mu)B_1 v = 0 \\ \dfrac{\partial \Delta v}{\partial \nu} + (1-\mu)\dfrac{\partial B_2 v}{\partial \tau} = h \end{cases} \quad \text{on } \Gamma_1, \end{cases}$$

$$Dh = v \Longleftrightarrow \begin{cases} \Delta v = 0 & \text{on } \Omega, \\ v|_{\Gamma} = h & \text{on } \Gamma; \end{cases} \quad Rh = v \Longleftrightarrow \begin{cases} \left(-\Delta + \frac{\sigma}{\eta}\mathbf{I}\right) v = 0 & \text{on } \Omega, \\ \dfrac{\partial v}{\partial \nu} + \lambda v = h & \text{on } \Gamma_2, \\ \dfrac{\partial v}{\partial \nu} + \lambda v = 0 & \text{on } \Gamma \backslash \Gamma_2. \end{cases} \tag{1.20}$$

The classic regularity results of [21, p. 152] then provide that for all q real,

$$\begin{cases} D \in \mathcal{L}\left(H^q(\Gamma), H^{q+\frac{1}{2}}(\Omega)\right), \\ R \in \mathcal{L}\left(H_0^q(\Gamma_2), H^{q+\frac{3}{2}}(\Omega)\right), \\ G_1 \in \mathcal{L}\left(H_0^q(\Gamma_1), H^{q+\frac{5}{2}}(\Omega)\right), \\ G_2 \in \mathcal{L}\left(H_0^{q-\frac{1}{2}}(\Gamma_1), H^{q+3}(\Omega)\right). \end{cases} \tag{1.21}$$

Denoting the topological dual of H^q as $[H^q]'$ (pivotal with respect to the L^2-inner product), then with the elliptic operators A_R and R as defined above, one can show that for $q \geq -\frac{1}{2}$, the (Banach space) adjoint $R^* A_R \in \mathcal{L}(D(A_R^{\frac{1}{2}}), [H^q(\Gamma_2)]')$ satisfies

$$R^* A_R \vartheta = \vartheta|_{\Gamma_2} \quad \text{for all } \vartheta \in D\left(A_R^{\frac{1}{2}}\right). \tag{1.22}$$

Moreover, with the operators $\mathring{\mathbf{A}}$ and G_i as defined above, one can readily show with the use of Green's formula (1.9) that $\forall\ \varpi \in D(\mathring{\mathbf{A}}^{\frac{1}{2}})$ the (Banach space) adjoints $G_i^* \mathring{\mathbf{A}} \in \mathcal{L}(D(\mathring{\mathbf{A}}^{\frac{1}{2}}), H^{i-\frac{1}{2}}(\Gamma_1))$ satisfy for $i = 1, 2$,

$$G_i^* \mathring{\mathbf{A}} \varpi = \begin{cases} (-1)^{i-1}\ \gamma_{2-i}\varpi|_{\Gamma_1} & \text{on } \Gamma_1, \\ 0 & \text{on } \Gamma_0. \end{cases} \tag{1.23}$$

- With A_N given by (1.16), we define the operator $P_\gamma : D(P_\gamma) \subset L^2(\Omega) \to L^2(\Omega)$ by

$$P_\gamma \equiv \mathbf{I}\ + \gamma A_N. \tag{1.24}$$

(i) With the parameter $\gamma > 0$, we define a space $H^1_{\Gamma_0,\gamma}(\Omega)$ equivalent to $H^1_{\Gamma_0}(\Omega)$ with inner product

$$(\omega_1, \omega_2)_{H^1_{\Gamma_0,\gamma}(\Omega)} \equiv (\omega_1, \omega_2)_{L^2(\Omega)} + \gamma\, (\nabla\omega_1, \nabla\omega_2)_{L^2(\Omega)} \quad \forall\, \omega_1, \omega_2 \in H^1_{\Gamma_0}(\Omega) \tag{1.25}$$

and with its dual denoted as $H^{-1}_{\Gamma_0,\gamma}(\Omega)$. After recalling that $H^1_{\Gamma_0}(\Omega) = D(A_N^{\frac{1}{2}})$ (by (1.17)), two extensions by continuity will then yield that

$$\begin{aligned} &P_\gamma \in \mathcal{L}\left(H^1_{\Gamma_0,\gamma}(\Omega), H^{-1}_{\Gamma_0,\gamma}(\Omega)\right) \\ &\text{with } \langle P_\gamma \omega_1, \omega_2 \rangle_{H^{-1}_{\Gamma_0,\gamma}(\Omega) \times H^1_{\Gamma_0,\gamma}(\Omega)} = (\omega_1, \omega_2)_{H^1_{\Gamma_0,\gamma}(\Omega)}. \end{aligned} \tag{1.26}$$

Furthermore, the obvious $H^1_{\Gamma_0,\gamma}(\Omega)$-ellipticity of P_γ and Lax–Milgram give us that $P_\gamma \in \mathcal{L}(H^1_{\Gamma_0,\gamma}(\Omega), H^{-1}_{\Gamma_0,\gamma}(\Omega))$is boundedly invertible, with

$$P_\gamma^{-1} \in \mathcal{L}\left(H^{-1}_{\Gamma_0,\gamma}(\Omega), H^1_{\Gamma_0,\gamma}(\Omega)\right). \tag{1.27}$$

Moreover, because P_γ is positive definite and self-adjoint as an operator $P_\gamma : L^2(\Omega) \supset D(P_\gamma) \to L^2(\Omega)$, the square root $P_\gamma^{\frac{1}{2}}$ is consequently well defined with $D(P_\gamma^{\frac{1}{2}}) = H^1_{\Gamma_0,\gamma}(\Omega)$, by (1.17). It then follows from (1.25) and (1.26) that for ω and $\widehat{\omega} \in H^1_{\Gamma_0,\gamma}(\Omega)$,

$$\left\| P_\gamma^{\frac{1}{2}} \omega \right\|^2_{L^2(\Omega)} = \|\omega\|^2_{L^2(\Omega)} + \gamma \|\nabla \omega\|^2_{L^2(\Omega)} = \|\omega\|^2_{H^1_{\Gamma_0,\gamma}(\Omega)}, \tag{1.28}$$

$$\left(P_\gamma^{\frac{1}{2}} \omega, P_\gamma^{\frac{1}{2}} \widehat{\omega} \right)_{L^2(\Omega)} = (\omega, \widehat{\omega})_{H^1_{\Gamma_0,\gamma}(\Omega)}. \tag{1.29}$$

(ii) Finally, by Green's formula we have for ω, $\widehat{\omega} \in D(\mathring{\mathbf{A}}^{\frac{1}{2}})$,

$$\begin{aligned} &\gamma \left\langle \left(\Delta + \mathring{\mathbf{A}} G_2 \gamma_1\right) \omega, \widehat{\omega} \right\rangle_{H^{-1}_{\Gamma_0,\gamma}(\Omega) \times H^1_{\Gamma_0,\gamma}(\Omega)} \\ &= -\gamma \left(\nabla \omega, \nabla \widehat{\omega}\right)_{L^2(\Omega)} + \gamma \left(\frac{\partial \omega}{\partial \nu}, \widehat{\omega} \right)_{L^2(\Gamma_1)} + \gamma \left(\gamma_1 \omega, G_2^* \mathring{\mathbf{A}} \widehat{\omega} \right)_{L^2(\Gamma_1)} \\ &= -\gamma \left(\nabla \omega, \nabla \widehat{\omega}\right)_{L^2(\Omega)} = -\gamma \left\langle A_N \omega, \widehat{\omega} \right\rangle_{H^{-1}_{\Gamma_0,\gamma}(\Omega) \times H^1_{\Gamma_0,\gamma}(\Omega)} \end{aligned} \tag{1.30}$$

after using (1.23). We thus obtain after two extensions by continuity to $H^1_{\Gamma_0,\gamma}(\Omega)$ that

$$P_\gamma = \mathbf{I} - \gamma \left(\Delta + \mathring{\mathbf{A}} G_2 \gamma_1\right) \text{ as elements of } \mathcal{L}\left(H^1_{\Gamma_0,\gamma}(\Omega), H^{-1}_{\Gamma_0,\gamma}(\Omega) \right). \tag{1.31}$$

In obtaining the equality above, we have used implicitly the fact that for every $\varpi^* \in H^{-1}_{\Gamma_0,\gamma}(\Omega)$ and $\varpi \in D(\mathring{\mathbf{A}}^{\frac{1}{2}})$,

$$\langle \varpi^*, \varpi \rangle_{H^{-1}_{\Gamma_0,\gamma}(\Omega) \times H^1_{\Gamma_0,\gamma}(\Omega)} = \langle \varpi^*, \varpi \rangle_{\left[D\left(\mathring{\mathbf{A}}^{\frac{1}{2}} \right) \right]' \times D\left(\mathring{\mathbf{A}}^{\frac{1}{2}} \right)} \tag{1.32}$$

- We denote the Hilbert space $\mathbf{H}_\gamma$ to be

$$\mathbf{H}_\gamma \equiv D\left(\mathring{\mathbf{A}}^{\frac{1}{2}} \right) \times H^1_{\Gamma_0,\gamma}(\Omega) \times L^2(\Omega), \tag{1.33}$$

with the inner product

$$\begin{aligned} \left(\begin{bmatrix} \omega_1 \\ \omega_2 \\ \theta \end{bmatrix}, \begin{bmatrix} \widehat{\omega}_1 \\ \widehat{\omega}_2 \\ \widehat{\theta} \end{bmatrix} \right)_{\mathbf{H}_\gamma} &= \left(\mathring{\mathbf{A}}^{\frac{1}{2}} \omega_1, \mathring{\mathbf{A}}^{\frac{1}{2}} \widehat{\omega}_1 \right)_{L^2(\Omega)} \\ &\quad + \left(P_\gamma^{\frac{1}{2}} \omega_2, P_\gamma^{\frac{1}{2}} \widehat{\omega}_2 \right)_{L^2(\Omega)} + \beta \left(\theta, \widehat{\theta} \right)_{L^2(\Omega)}. \end{aligned} \tag{1.34}$$

- With the above definitions, and making the denotation

$$(\clubsuit) \equiv A_R - \frac{\sigma}{\eta} - \mathring{\mathbf{A}} G_1 \gamma_0 + \lambda \mathring{\mathbf{A}} G_2 \gamma_0, \tag{1.35}$$

we then set $\mathcal{A}_\gamma : \mathbf{H}_\gamma \supset D(\mathcal{A}_\gamma) \to \mathbf{H}_\gamma$ to be

$$\begin{aligned} \mathcal{A}_\gamma &\equiv \begin{pmatrix} \mathbf{I} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & P_\gamma^{-1} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{I} \end{pmatrix} \begin{pmatrix} 0 & \mathbf{I} & 0 \\ -\mathring{\mathbf{A}} & 0 & \alpha(\clubsuit) \\ 0 & -\dfrac{\alpha}{\beta} A_D (\mathbf{I} - D\gamma_0) & -\dfrac{\eta}{\beta} A_R \end{pmatrix} \\ &\text{with } D(\mathcal{A}_\gamma) = \Big\{ [\omega_0, \omega_1, \theta_0] \in D\left(\mathring{\mathbf{A}}^{\frac{1}{2}} \right) \times D\left(\mathring{\mathbf{A}}^{\frac{1}{2}} \right) \times D(A_R)) \\ &\qquad \text{such that } \mathring{\mathbf{A}} \omega_0 + \alpha \mathring{\mathbf{A}} G_1 \gamma_0 \theta_0 \in H^{-1}_{\Gamma_0,\gamma}(\Omega) \Big\}. \end{aligned} \tag{1.36}$$

- We make the following denotations for the space of controllability:

$$\begin{aligned} U_s &= L^2(\Gamma_1) \times H^{-1}(\Gamma_1) \times H^s(\Gamma_2), \\ \mathcal{U}_s &= L^2(0,T; L^2(\Gamma_1) \times H^{-1}(\Gamma_1)) \times H^s((0,T) \times \Gamma_2), \end{aligned} \tag{1.37}$$

where $s \geq 0$. We define the control operator $\mathcal{B}$ on U_s by having for every $\overline{u} = [u_1, u_2, u_3] \in U_s$,

$$\mathcal{B}\overline{u} = \begin{bmatrix} 0 \\ P_\gamma^{-1}[\mathring{\mathbf{A}} G_1 u_1 + \mathring{\mathbf{A}} G_2 u_2] \\ \dfrac{\eta}{\beta} A_R R u_3 \end{bmatrix}. \tag{1.38}$$

Note that a priori the mapping $\mathcal{B}$ only makes sense as an element of $\mathcal{L}(U_s, [D(\mathcal{A}_\gamma^*)]')$, where $\mathbf{H}_\gamma \subset [D(\mathcal{A}_\gamma^*)]'$. Indeed, for fixed $\overline{u} = [u_1, u_2, u_3] \in U_s$ one has, upon using the expression for the inverse $\mathcal{A}_\gamma^{-1}$ given in (4.2) below, and the definition of the elliptic operators G_1, G_2, and R in (1.19) and (1.20) above, that

$$\begin{aligned} \mathcal{B}\overline{u} &= \mathcal{A}_\gamma \mathcal{A}_\gamma^{-1} \begin{bmatrix} 0 \\ P_\gamma^{-1}[\mathring{\mathbf{A}} G_1 u_1 + \mathring{\mathbf{A}} G_2 u_2] \\ \dfrac{\eta}{\beta} A_R R u_3 \end{bmatrix} \\ &= \mathcal{A}_\gamma \begin{bmatrix} -G_1 u_1 - G_2 u_2 - \alpha \mathring{\mathbf{A}}^{-1}(\clubsuit) R u_3 \\ 0 \\ -R u_3 \end{bmatrix} \in \left[D(\mathcal{A}_\gamma^*)\right]', \end{aligned} \tag{1.39}$$

where ($\clubsuit$) is as defined in (1.35).

- By duality, we have

$$\begin{aligned} U_s^* &= L^2(\Gamma_1) \times H^1(\Gamma_1) \times [H^s(\Gamma_2)]', \\ \mathcal{U}_s^* &= L^2(0,T; L^2(\Gamma_1) \times H^1(\Gamma_1)) \times \left[H^s(0,T; L^2(\Gamma_2))\right]', \end{aligned} \tag{1.40}$$

and $\mathcal{B}^* \in \mathcal{L}\left(D(\mathcal{A}_\gamma^*), U_s^*\right)$.

1.3.2. Abstract operator formulation. If we take the initial data $[\omega_0, \omega_1, \theta_0]$ to be in $\mathbf{H}_\gamma$, and control $\overline{u} \in \mathcal{U}_s$, where $\mathcal{U}_s$ is as defined in (1.37), then considering the operator definitions above, the coupled system (1.1) can be rewritten a fortiori as the operator theoretic model

$$\frac{d}{dt}\begin{bmatrix} \omega(t) \\ \omega_t(t) \\ \theta(t) \end{bmatrix} = \mathcal{A}_\gamma \begin{bmatrix} \omega(t) \\ \omega_t(t) \\ \theta(t) \end{bmatrix} + \mathcal{B}\overline{u}(t), \qquad \begin{bmatrix} \omega(0) \\ \omega_t(0) \\ \theta(0) \end{bmatrix} = \begin{bmatrix} \omega_0 \\ \omega_1 \\ \theta_0 \end{bmatrix}, \tag{1.41}$$

with this equation having sense in $\left[D(\mathcal{A}_\gamma^*)\right]'$ (a space strictly larger than $\mathbf{H}_\gamma$). Given the operator definitions for $\mathcal{A}_\gamma$ and $\mathcal{B}$ above, the solution $[\omega, \omega_t, \theta]$ to the ODE (1.41) (and so to the PDE (1.1)) is given by

$$\begin{bmatrix} \omega(\cdot) \\ \omega_t(\cdot) \\ \theta(\cdot) \end{bmatrix} = e^{\mathcal{A}_\gamma(\cdot)} \begin{bmatrix} \omega_0 \\ \omega_1 \\ \theta_0 \end{bmatrix} + \int_0^{(\cdot)} e^{\mathcal{A}_\gamma(\cdot - s)} \mathcal{B}\overline{u}(s) ds, \tag{1.42}$$

which by (1.39) and the convolution theorem is an element of $C([0,T];[D(\mathcal{A}_\gamma^*)]')$. With this representation of the solution $[\omega,\omega_t,\theta]$ in mind, we define the *input* $\to$ *terminal state* map $\mathcal{L}_T \in \mathcal{L}(\mathcal{U}_s, \left[D(\mathcal{A}_\gamma^*)\right]')$ as

$$\mathcal{L}_T\overline{u} = \int_0^T e^{\mathcal{A}_\gamma(T-s)}\mathcal{B}\overline{u}(s)ds. \tag{1.43}$$

Taken as an unbounded operator from $\mathcal{U}_s$ into $\mathbf{H}_\gamma$, then $\mathcal{L}_T : D(\mathcal{L}_T) \subset \mathcal{U}_s \to \mathbf{H}_\gamma$ is closed and densely defined, with its domain of definition $D(\mathcal{L}_T)$ given to be

$$D(\mathcal{L}_T) = \{\overline{u} \in \mathcal{U}_s : \mathcal{L}_T\overline{u} \in \mathbf{H}_\gamma\}. \tag{1.44}$$

Its adjoint $\mathcal{L}_T^* : D(\mathcal{L}_T^*) \subset \mathbf{H}_\gamma \to \mathcal{U}_s^*$, where $\mathcal{U}_s^*$ is as given in (1.40), is likewise closed and densely defined, with

$$D(\mathcal{L}_T^*) = \left\{ [\phi_0,\phi_1,\psi_0] \in \mathbf{H}_\gamma : \mathcal{L}_T^* \begin{bmatrix} \phi_0 \\ \phi_1 \\ \psi_0 \end{bmatrix} \in \mathcal{U}_s^* \right\}. \tag{1.45}$$

As we are concerned with obtaining exact controllability of the displacement $[\omega,\omega_t]$ only, we accordingly define the projection operator $\Pi : \mathbf{H}_\gamma \to D(\mathring{\mathbf{A}}^{\frac{1}{2}}) \times H^1_{\Gamma_0,\gamma}(\Omega)$ by

$$\Pi \begin{bmatrix} \varpi_0 \\ \varpi_1 \\ \vartheta_0 \end{bmatrix} = \begin{bmatrix} \varpi_0 \\ \varpi_1 \end{bmatrix}. \tag{1.46}$$

Henceforth, the work here will be concerned with determining the surjectivity of the closed operator $\Pi\mathcal{L}_T$, $D(\Pi\mathcal{L}_T) \subset \mathcal{U}_s \to D(\mathring{\mathbf{A}}^{\frac{1}{2}}) \times H^1_{\Gamma_0,\gamma}(\Omega)$, with

$$\Pi\mathcal{L}_T\overline{u} = \Pi \int_0^T e^{\mathcal{A}_\gamma(T-s)}\mathcal{B}\overline{u}(s)ds, \tag{1.47}$$

and with $D(\Pi\mathcal{L}_T) = D(\mathcal{L}_T)$. Determining the surjectivity of the operator $\Pi\mathcal{L}_T$ for some $T > 0$ becomes our concern here, since it is equivalent to showing the exact controllability of the mechanical component $[\omega,\omega_t]$ to (1.1) (Theorem 1.2). This surjectivity for $\Pi\mathcal{L}_T$ is in turn equivalent to the existence of a certain observability inequality pertaining to the range of the adjoint $\mathcal{L}_T^*\Pi^*$ (the inequality (2.1) below), where $\mathcal{L}_T^*\Pi^* : D(\mathcal{L}_T^*\Pi^*) \subset D(\mathring{\mathbf{A}}^{\frac{1}{2}}) \times H^1_{\Gamma_0,\gamma}(\Omega) \to \mathbf{H}_\gamma$ is likewise a closed densely defined operator (as $\mathcal{L}_T^*$ is), with its domain given by

$$D(\mathcal{L}_T^*\Pi^*) = \left\{ [\phi_0,\phi_1] \in D\left(\mathring{\mathbf{A}}^{\frac{1}{2}}\right) \times H^1_{\Gamma_0}(\Omega) : [\phi_0,\phi_1,0] \in D(\mathcal{L}_T^*) \right\}. \tag{1.48}$$

It is the injectivity condition (2.1) that we intend to directly verify. In order to rewrite this abstract inequality in "PDE form" (i.e., as the inequality (2.2) below), we need the following two propositions, the first of which is proved in the appendix below.

PROPOSITION 1.5. *The Hilbert space adjoint $\mathcal{A}_\gamma^*$ of $\mathcal{A}_\gamma$, as defined in* (1.36)*, is given to be*

$$\mathcal{A}_\gamma^* = \begin{pmatrix} \mathbf{I} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & P_\gamma^{-1} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{I} \end{pmatrix} \begin{pmatrix} \mathbf{0} & -\mathbf{I} & \mathbf{0} \\ \mathring{\mathbf{A}} & \mathbf{0} & -\alpha(\clubsuit) \\ \mathbf{0} & \dfrac{\alpha}{\beta}A_D(\mathbf{I}-D\gamma_0) & -\dfrac{\eta}{\beta}A_R \end{pmatrix},$$

$$\text{with } D(\mathcal{A}_\gamma^*) = \left\{ [\phi_0,\phi_1,\psi_0] \in D\left(\mathring{\mathbf{A}}^{\frac{1}{2}}\right) \times D\left(\mathring{\mathbf{A}}^{\frac{1}{2}}\right) \times D(A_R) \right.$$

$$\left. \text{such that } \mathring{\mathbf{A}}\phi_0 + \alpha\mathring{\mathbf{A}}G_1\gamma_0\psi_0 \in H^{-1}_{\Gamma_0,\gamma}(\Omega) \right\} \tag{1.49}$$

(above, (♣) is the same denotation made in (1.35)*).*

Remark 1.6. Using the semigroup $\{e^{\mathcal{A}_\gamma^* t}\}_{t\geq 0}$ generated by $\mathcal{A}_\gamma^*$, then for terminal data $[\phi_0, \phi_1, \psi_0] \in \mathbf{H}_\gamma$,

$$\begin{bmatrix} \phi(t) \\ \phi_t(t) \\ \psi(t) \end{bmatrix} = e^{\mathcal{A}_\gamma^*(T-t)} \begin{bmatrix} \phi_0 \\ \phi_1 \\ \psi_0 \end{bmatrix} \in C([0,T];\ \mathbf{H}_\gamma) \tag{1.50}$$

is the solution to the following backward problem:

$$\begin{cases} \begin{cases} \phi_{tt} - \gamma\Delta\phi_{tt} + \Delta^2\phi + \alpha\Delta\psi = 0 \\ \beta\psi_t + \eta\Delta\psi - \sigma\psi - \alpha\Delta\phi_t = 0 \end{cases} \text{ on } (0,\infty)\times\Omega, \\ \phi = \dfrac{\partial\phi}{\partial\nu} = 0 \ \text{ on } (0,\infty)\times\Gamma_0, \\ \begin{cases} \Delta\phi + (1-\mu)B_1\phi + \alpha\psi = 0 \\ \dfrac{\partial\Delta\phi}{\partial\nu} + (1-\mu)\dfrac{\partial B_2\phi}{\partial\tau} - \gamma\dfrac{\partial\phi_{tt}}{\partial\nu} + \alpha\dfrac{\partial\psi}{\partial\nu} = 0 \end{cases} \text{ on } (0,\infty)\times\Gamma_1, \\ \dfrac{\partial\psi}{\partial\nu} + \lambda\psi = 0 \ \text{ on } (0,\infty)\times\Gamma,\ \lambda\geq 0, \\ [\phi(T), \phi_t(T), \psi(T)] = [\phi_0, \phi_1, \psi_0]\,. \end{cases} \tag{1.51}$$

Remark 1.7. For terminal data $[\phi_0, \phi_1, \psi_0]$ in $D(\mathcal{A}_\gamma^*)$, the two equations of (1.51) may be written pointwise as

$$P_\gamma\phi_{tt} = -\mathring{\mathbf{A}}\phi - \alpha\mathring{\mathbf{A}}G_1\gamma_0\psi + \alpha\lambda\mathring{\mathbf{A}}G_2\gamma_0\psi - \alpha\Delta\psi \ \text{ in } H^{-1}_{\Gamma_0,\gamma}(\Omega), \tag{1.52}$$

$$\beta\psi_t = -\eta\Delta\psi + \sigma\psi + \alpha\Delta\phi_t \ \text{ in } L^2(\Omega), \tag{1.53}$$

$$[\phi(T), \phi_t(T), \psi(T)] = [\phi_0, \phi_1, \psi_0]\,. \tag{1.54}$$

Remark 1.8. Since $\overline{\Gamma}_0 \cap \overline{\Gamma}_1 = \emptyset$, and Γ is smooth, we can assume throughout that $D(\mathcal{A}_\gamma^*)$ is dense in the graph topology of $D(\mathcal{L}_T^*)$.

Proposition 1.9. *The adjoint $\mathcal{L}_T^* : D(\mathcal{L}_T^*) \subset \mathbf{H}_\gamma \to \mathcal{U}_s^*$ of $\mathcal{L}_T$ is computed to be*

$$\mathcal{L}_T^* \begin{bmatrix} \phi_0 \\ \phi_1 \\ \psi_0 \end{bmatrix} = \begin{bmatrix} \dfrac{\partial\phi_t}{\partial\nu}\Big|_{\Gamma_1}, -\phi_t|_{\Gamma_1}, \eta\,\psi|_{\Gamma_2} \end{bmatrix} \ \textit{for all} \ \begin{bmatrix} \phi_0 \\ \phi_1 \\ \psi_0 \end{bmatrix} \in D(\mathcal{L}_T^*)\,, \tag{1.55}$$

where $[\frac{\partial\phi_t}{\partial\nu}|_{\Gamma_1}, \phi_t|_{\Gamma_1}, \psi|_{\Gamma_2}]$ are boundary "traces" of the solution $[\phi, \phi_t, \psi]$ to the coupled system (1.51).

Proof. By Remark 1.8, it is enough to show the characterization in (1.55) for $[\phi_0, \phi_1, \psi_0] \in D(\mathcal{A}_\gamma^*)$. With this in mind, one has readily the classic representation

$$\mathcal{L}_T^* \begin{bmatrix} \phi_0 \\ \phi_1 \\ \psi_0 \end{bmatrix} = \mathcal{B}^* e^{\mathcal{A}_\gamma^*(T-t)} \begin{bmatrix} \phi_0 \\ \phi_1 \\ \psi_0 \end{bmatrix} \quad \text{for every} \quad \begin{bmatrix} \phi_0 \\ \phi_1 \\ \psi_0 \end{bmatrix} \in D\left(\mathcal{A}_\gamma^*\right), \tag{1.56}$$

where again, $\mathcal{B}^* \in \mathcal{L}\left(D\left(\mathcal{A}_\gamma^*\right), U_s^*\right)$ is the adjoint of $\mathcal{B}$. We must show that the right-hand side of this equality may be written explicitly in "PDE form" as (1.55). To this end, for every $[u_1, u_2, u_3] \in \mathcal{U}_s$ and $[\phi_0, \phi_1, \psi_0] \in D\left(\mathcal{A}_\gamma^*\right)$, we have

(1.57)

$$
\begin{aligned}
&\left\langle \mathcal{L}_T \begin{bmatrix} u_1 \\ u_2 \\ u_3 \end{bmatrix}, \begin{bmatrix} \phi_0 \\ \phi_1 \\ \psi_0 \end{bmatrix} \right\rangle_{[D(\mathcal{A}_\gamma^*)]' \times D(\mathcal{A}_\gamma^*)} \\
&= \left\langle \int_0^T e^{\mathcal{A}_\gamma (T-s)} \mathcal{B} \begin{bmatrix} u_1(s) \\ u_2(s) \\ u_3(s) \end{bmatrix} ds, \begin{bmatrix} \phi_0 \\ \phi_1 \\ \psi_0 \end{bmatrix} \right\rangle_{[D(\mathcal{A}_\gamma^*)]' \times D(\mathcal{A}_\gamma^*)} \\
&= \int_0^T \left\langle e^{\mathcal{A}_\gamma (T-s)} \mathcal{A}_\gamma \mathcal{A}_\gamma^{-1} \mathcal{B} \begin{bmatrix} u_1(s) \\ u_2(s) \\ u_3(s) \end{bmatrix}, \begin{bmatrix} \phi_0 \\ \phi_1 \\ \psi_0 \end{bmatrix} \right\rangle_{[D(\mathcal{A}_\gamma^*)]' \times D(\mathcal{A}_\gamma^*)} ds \\
&= \int_0^T \left(\mathcal{A}_\gamma^{-1} \mathcal{B} \begin{bmatrix} u_1(s) \\ u_2(s) \\ u_3(s) \end{bmatrix}, e^{\mathcal{A}_\gamma^* (T-s)} \mathcal{A}_\gamma^* \begin{bmatrix} \phi_0 \\ \phi_1 \\ \psi_0 \end{bmatrix} \right)_{\mathbf{H}_\gamma} ds \\
&= \int_0^T \left(\begin{bmatrix} -G_1 u_1(s) - G_2 u_2(s) - \alpha \mathring{\mathbf{A}}^{-1} (\clubsuit) R u_3 \\ 0 \\ -R u_3 \end{bmatrix}, \mathcal{A}_\gamma^* e^{\mathcal{A}_\gamma^* (T-s)} \begin{bmatrix} \phi_0 \\ \phi_1 \\ \psi_0 \end{bmatrix} \right)_{\mathbf{H}_\gamma} ds.
\end{aligned}
$$

Noting that

$$
\begin{bmatrix} \phi(t) \\ \phi_t(t) \\ \psi(t) \end{bmatrix} \equiv e^{\mathcal{A}_\gamma^* (T-t)} \begin{bmatrix} \phi_0 \\ \phi_1 \\ \psi_0 \end{bmatrix}
$$

gives the solution to the backward problem (1.51), we then use this relation, the definition of the adjoint $\mathcal{A}_\gamma^*$ in (1.49), and Proposition 4.1 of the appendix to obtain

(1.58)

$$
\begin{aligned}
&\left\langle \mathcal{L}_T \begin{bmatrix} u_1(s) \\ u_2(s) \\ u_3(s) \end{bmatrix}, \begin{bmatrix} \phi_0 \\ \phi_1 \\ \psi_0 \end{bmatrix} \right\rangle_{[D(\mathcal{A}_\gamma^*)]' \times D(\mathcal{A}_\gamma^*)} \\
&= \int_0^T \left(\begin{bmatrix} -G_1 u_1(s) - G_2 u_2(s) - \alpha \mathring{\mathbf{A}}^{-1} (\clubsuit) R u_3 \\ 0 \\ -R u_3 \end{bmatrix}, \right. \\
&\qquad \left. \begin{bmatrix} -\phi_t \\ P_\gamma^{-1} \mathring{\mathbf{A}} \phi - \alpha P_\gamma^{-1} (\clubsuit) \psi \\ \dfrac{\alpha}{\beta} A_D (\mathbf{I} - D\gamma_0) \phi_t - \dfrac{\eta}{\beta} A_R \psi \end{bmatrix} \right)_{\mathbf{H}_\gamma} ds \\
&= \int_0^T \left[\left(\mathring{\mathbf{A}}^{\frac{1}{2}} G_1 u_1, \mathring{\mathbf{A}}^{\frac{1}{2}} \phi_t \right)_{L^2(\Omega)} + \left(\mathring{\mathbf{A}}^{\frac{1}{2}} G_2 u_2, \mathring{\mathbf{A}}^{\frac{1}{2}} \phi_t \right)_{L^2(\Omega)} + \eta \left(R u_3, A_R \psi \right)_{L^2(\Omega)} \right] dt \\
&= \int_0^T \left[\left(u_1, G_1^* \mathring{\mathbf{A}} \phi_t \right)_{L^2(\Gamma_1)} + \left\langle u_2, G_2^* \mathring{\mathbf{A}} \phi_t \right\rangle_{H^{-1}(\Gamma_1) \times H^1(\Gamma_1)} \right] dt + \int_0^T \eta \left(u_3, \psi \right)_{L^2(\Gamma_2)} dt \\
&= \int_0^T \left[\left(u_1, \frac{\partial \phi_t}{\partial \nu} \right)_{L^2(\Gamma_1)} - \left\langle u_2, \phi_t \right\rangle_{H^{-1}(\Gamma_1) \times H^1(\Gamma_1)} \right] dt + \eta \left\langle u_3, \psi \right\rangle_{H^s((0,T)\Gamma_2) \times [H^s((0,T) \times \Gamma_2)]'},
\end{aligned}
$$

thereby completing the proof of Proposition 1.9. □

Immediately, we have Corollary 1.10.

COROLLARY 1.10. *The adjoint operator* $\mathcal{L}_T^*\Pi^* : D(\mathcal{L}_T^*\Pi^*) \subset D(\mathring{\mathbf{A}}^{\frac{1}{2}}) \times H^1_{\Gamma_0,\gamma}(\Omega) \to \mathcal{U}_s^*$ *is given by*

$$\mathcal{L}_T^*\Pi^* \begin{bmatrix} \phi_0 \\ \phi_1 \end{bmatrix} = \left[\left. \frac{\partial \phi_t}{\partial \nu} \right|_{\Gamma_1}, -\phi_t|_{\Gamma_1}, \eta\psi|_{\Gamma_2} \right] \tag{1.59}$$

for all $[\phi_0, \phi_1] \in D(\mathcal{L}_T^*\Pi^*)$, *where* $[\frac{\partial \phi_t}{\partial \nu}|_{\Gamma_1}, \phi_t|_{\Gamma_1}, \psi|_{\Gamma_2}]$ *are boundary traces of the solution* $[\phi, \phi_t, \psi]$ *to the following (backward) system:*

$$\begin{cases} \begin{cases} \phi_{tt} - \gamma\Delta\phi_{tt} + \Delta^2\phi + \alpha\Delta\psi = 0 \\ \beta\psi_t + \eta\Delta\psi - \sigma\psi - \alpha\Delta\phi_t = 0 \end{cases} & \text{on } (0,\infty)\times\Omega, \\ \phi = \dfrac{\partial\phi}{\partial\nu} = 0 \quad \text{on } (0,\infty)\times\Gamma_0, \\ \begin{cases} \Delta\phi + (1-\mu)B_1\phi + \alpha\psi = 0 \\ \dfrac{\partial\Delta\phi}{\partial\nu} + (1-\mu)\dfrac{\partial B_2\phi}{\partial\tau} - \gamma\dfrac{\partial\phi_{tt}}{\partial\nu} + \alpha\dfrac{\partial\psi}{\partial\nu} = 0 \end{cases} & \text{on } (0,\infty)\times\Gamma_1, \\ \dfrac{\partial\psi}{\partial\nu} + \lambda\psi = 0 \quad \text{on } (0,\infty)\times\Gamma,\ \lambda \geq 0, \\ [\phi(T), \phi_t(T), \psi(T)] = [\phi_0, \phi_1, 0]\,. \end{cases} \tag{1.60}$$

We conclude this section with a regularity result for the thermal component of the solution $[\phi, \phi_t, \psi]$ to (1.51) , this being originally derived in [11] and [2] for the forward problem (1.1). Assuming terminal data $[\phi_0, \phi_1, \psi_0] \in D(\mathcal{A}_\gamma^*)$, we have, by using (1.50), the equality

$$\frac{d}{dt}\left\| \begin{bmatrix} \phi(t) \\ \phi_t(t) \\ \psi(t) \end{bmatrix} \right\|^2_{\mathbf{H}_\gamma} = -2\left(\mathcal{A}_\gamma^* \begin{bmatrix} \phi(t) \\ \phi_t(t) \\ \psi(t) \end{bmatrix}, \begin{bmatrix} \phi(t) \\ \phi_t(t) \\ \psi(t) \end{bmatrix} \right)_{\mathbf{H}_\gamma}. \tag{1.61}$$

Integrating this equation from 0 to T, performing computations similar to those performed for the proof of Proposition 1.9, recalling the characterization (1.15), and subsequently invoking a density argument, we have the following proposition.

PROPOSITION 1.11. *With terminal data* $[\phi_0, \phi_1, \psi_0] \in \mathbf{H}_\gamma$, *we have that the component* ψ *of the solution of* (1.51) *is an element of* $L^2(0,\infty; D(A_R^{\frac{1}{2}}))$. *Indeed, we have the following relation valid for all* $T > 0$:

$$\left\| \begin{bmatrix} \phi_0 \\ \phi_1 \\ \psi_0 \end{bmatrix} \right\|^2_{\mathbf{H}_\gamma} - \left\| \begin{bmatrix} \phi(0) \\ \phi_t(0) \\ \psi(0) \end{bmatrix} \right\|^2_{\mathbf{H}_\gamma} = 2\eta \int_0^T \left\| A_R^{\frac{1}{2}}\psi \right\|^2_{L^2(\Omega)} dt. \tag{1.62}$$

2. Proof of Theorem 1.2.

2.1. The necessary inequality. As stated above, showing the partial exact controllability of the system (1.1) for some time $T > 0$ is equivalent to showing the surjectivity of the operator $\Pi\mathcal{L}_T : D(\mathcal{L}_T) \subset \mathcal{U}_s \to D(\mathring{\mathbf{A}}^{\frac{1}{2}}) \times H^1_{\Gamma_0,\gamma}(\Omega)$, where $\Pi\mathcal{L}_T$ is as defined in (1.47) and with $D(\mathcal{L}_T)$ as defined in (1.44). Using the classical functional

analysis (e.g., couple Lemma 3.8.18(i) and Theorem 6.5.10(ii) of [9]), the surjectivity of $\Pi\mathcal{L}_T$ for some time $T > 0$ is tantamount to the existence of a constant $C_T > 0$ such that following inequality is satisfied for all $[\phi_0, \phi_1] \in D(\mathcal{L}_T^*\Pi^*)$, where $D(\mathcal{L}_T^*\Pi^*)$ is as defined in (1.48):

$$\left\| \mathcal{L}_T^* \begin{bmatrix} \phi_0 \\ \phi_1 \\ 0 \end{bmatrix} \right\|_{\mathcal{U}_s^*} \geq C_T \left\| \begin{bmatrix} \phi_0 \\ \phi_1 \end{bmatrix} \right\|_{D\left(\mathring{\mathbf{A}}^{\frac{1}{2}}\right) \times H^1_{\Gamma_0,\gamma}(\Omega)}. \tag{2.1}$$

Corollary 1.10 then gives that this abstract inequality above may be rewritten by having for all $[\phi_0, \phi_1] \in D(\mathcal{L}_T^*\Pi^*)$,

$$\int_0^T \left[\|\phi_t\|^2_{H^1(\Gamma_1)} + \left\| \frac{\partial \phi_t}{\partial \nu} \right\|^2_{L^2(\Gamma_1)} \right] dt + \eta \|\psi\|^2_{[H^s((0,T)\times\Gamma_2)]'} \geq C_T \left\| \begin{bmatrix} \phi_0 \\ \phi_1 \end{bmatrix} \right\|^2_{D\left(\mathring{\mathbf{A}}^{\frac{1}{2}}\right) \times H^1_{\Gamma_0,\gamma}(\Omega)}, \tag{2.2}$$

where $[\frac{\partial \phi_t}{\partial \nu}|_{\Gamma_1}, \phi_t|_{\Gamma_1}, \psi|_{\Gamma_2}]$ are traces of the solution $[\phi, \phi_t, \psi]$ to the backward system (1.60) (this being "adjoint" with respect to (1.1)). So to prove the statement of partial exact controllability of the thermoelastic system (Theorem 1.2), it will hence suffice to establish the inequality (2.2) for $T > 0$ large enough. With this end in mind, we make the following denotation for the mechanical "energy" of the system for $0 \leq t \leq T$:

$$E_\phi(t) = \frac{1}{2} \left[\left\| \mathring{\mathbf{A}}^{\frac{1}{2}} \phi(t) \right\|^2_{L^2(\Omega)} + \left\| P_\gamma^{\frac{1}{2}} \phi_t(t) \right\|^2_{L^2(\Omega)} \right], \tag{2.3}$$

where again $[\phi, \phi_t, \psi]$ solve the backward system (1.60). In addition, we will denote by $\mathsf{l.o.t.}(\phi, \phi_t, \psi)$ ("lower order terms") any sum of terms that obey the following estimate for some constant C_T:

$$\begin{aligned} \mathsf{l.o.t.}(\phi, \phi_t, \psi) \leq C_T \Bigg(& \|\phi\|^2_{L^\infty\left(0,T;H^{\frac{3}{2}+\epsilon}(\Omega)\right)} + \|\phi_t\|^2_{L^\infty\left(0,T;H^{\frac{1}{2}+\epsilon}(\Omega)\right)} \\ & + \|\psi\|^2_{L^2\left(0,T;H^{\frac{1}{2}+\epsilon}(\Omega)\right)} + \|\psi\|^2_{L^\infty\left(0,T;H^{-\frac{1}{2}+\epsilon}(\Omega)\right)} \Bigg). \end{aligned} \tag{2.4}$$

By way of establishing (2.2), the bulk of the work will entail the derivation of the following estimate.

THEOREM 2.1. *For $T > 0$ large enough, the solution $[\phi, \phi_t, \psi]$ to (1.51) with terminal data $[\phi_0, \phi_1, \psi_0] \in D(\mathcal{L}_T^*)$ satisfies the following inequality:*

$$\begin{aligned} & \int_0^T \left[E_\phi(t) + \left\| A_R^{\frac{1}{2}} \psi \right\|^2_{L^2(\Omega)} \right] dt + E_\phi(0) \\ & \leq C_T \left(\|\psi_0\|^2_{L^2(\Omega)} + \int_0^T \|\nabla \phi_t\|^2_{L^2(\Gamma_1)}\, dt + \mathsf{l.o.t.}(\phi, \phi_t, \psi) \right). \end{aligned} \tag{2.5}$$

This theorem will follow from a chain of results. Given the density of $D(\mathcal{A}_\gamma^*)$ in $D(\mathcal{L}_T^*)$ (see Remark 1.8) and the fact that the solution of (1.51) has the representation

$$\begin{bmatrix} \phi(t) \\ \phi_t(t) \\ \psi(t) \end{bmatrix} = e^{\mathcal{A}_\gamma^*(T-t)} \begin{bmatrix} \phi_0 \\ \phi_1 \\ \psi_0 \end{bmatrix}, \tag{2.6}$$

it will be enough to show inequality (2.5) for solutions $[\phi, \phi_t, \psi]$ to (1.51) corresponding to terminal data in $D(\left[\mathcal{A}_\gamma^*\right]^2)$. Taking $[\phi_0, \phi_1, \psi_0] \in D(\left[\mathcal{A}_\gamma^*\right]^2)$, we then have that $[\phi, \phi_t, \psi]$ is an element of $C^2([0,T]; \mathbf{H}_\gamma) \cap C^1([0,T]; D(\mathcal{A}_\gamma^*)) \cap C([0,T]; D(\left[\mathcal{A}_\gamma^*\right]^2))$ and as such has the additional regularity (see [3, Theorem 2] and also [12]):

$$\begin{aligned} &\phi \in C([0,T]; H^4(\Omega)); \ \phi_t \in C([0,T]; H^3(\Omega)); \ \phi_{tt} \in C\left([0,T]; D\left(\mathring{\mathbf{A}}^{\frac{1}{2}}\right)\right), \\ &\psi_t \in C([0,T]; D(A_R)), \\ &\phi - \gamma G_2 \gamma_1 \phi_{tt} + \alpha G_1 \gamma_0 \psi - \alpha\lambda G_2 \gamma_0 \psi \in C([0,T]; D(\mathring{\mathbf{A}})). \end{aligned} \tag{2.7}$$

This extra regularity of $[\phi, \phi_t, \psi]$, corresponding to smooth initial data, will justify the computations to be done below.

2.2. Proof of Theorem 2.1. As mentioned above, the terminal data $[\phi_0, \phi_1, \psi_0]$ will be considered to be in $D(\left[\mathcal{A}_\gamma^*\right]^2)$; accordingly the corresponding solution $[\phi, \phi_t, \psi]$ of (1.51) will be a classical one, with the regularity posted in (2.7). With the end in mind of deriving the estimate (2.2), we start by making the substitution

$$\widehat{\phi}(t) = e^{-\xi t}\phi(t) \text{ and } \widehat{\psi}(t) = e^{-\xi t}\psi(t), \tag{2.8}$$

where parameter $\xi \in \mathbb{R}$ is to be determined. Necessarily then $[\widehat{\phi}, \widehat{\phi}_t, \widehat{\psi}]$ solves the coupled (backward) system

$$\begin{cases} \begin{cases} \left(\xi^2\widehat{\phi} + 2\xi\widehat{\phi}_t + \widehat{\phi}_{tt}\right) - \gamma\Delta\left(\xi^2\widehat{\phi} + 2\xi\widehat{\phi}_t + \widehat{\phi}_{tt}\right) + \Delta^2\widehat{\phi} + \alpha\Delta\widehat{\psi} = 0 \\ \beta\left(\xi\widehat{\psi} + \widehat{\psi}_t\right) + \eta\Delta\widehat{\psi} - \sigma\widehat{\psi} - \alpha\Delta\left(\xi\widehat{\phi} + \widehat{\phi}_t\right) = 0 \end{cases} & \text{on } (0,\infty)\times\Omega, \\ \widehat{\phi} = \dfrac{\partial\widehat{\phi}}{\partial\nu} = 0 \ \text{ on } (0,\infty)\times\Gamma_0, \\ \begin{cases} \Delta\widehat{\phi} + (1-\mu)B_1\widehat{\phi} + \alpha\widehat{\psi} = 0 \\ \dfrac{\partial\Delta\widehat{\phi}}{\partial\nu} + (1-\mu)\dfrac{\partial B_2\widehat{\phi}}{\partial\tau} - \gamma\dfrac{\partial}{\partial\nu}\left(\xi^2\widehat{\phi} + 2\xi\widehat{\phi}_t + \widehat{\phi}_{tt}\right) + \alpha\dfrac{\partial\widehat{\psi}}{\partial\nu} = 0 \end{cases} & \text{on } (0,\infty)\times\Gamma_1, \\ \dfrac{\partial\widehat{\psi}}{\partial\nu} + \lambda\widehat{\psi} = 0 \ \text{ on } (0,\infty)\times\Gamma, \ \lambda \geq 0, \\ \left[\widehat{\phi}(T), \widehat{\phi}_t(T), \widehat{\psi}(T)\right] = \left[e^{-\xi T}\phi_0, -\xi e^{-\xi T}\phi_0 + e^{-\xi T}\phi_1, e^{-\xi T}\psi_0\right]. \end{cases} \tag{2.9}$$

Since $[\phi_0, \phi_1, \psi_0] \in D(\left[\mathcal{A}_\gamma^*\right]^2)$, the extra regularity in (2.7) gives that $[\widehat{\phi}, \widehat{\phi}_t, \widehat{\psi}]$ is a classical (not just weak) solution of (2.9); accordingly, we can rewrite (2.9) abstractly as (see Remark 1.7 and (1.31))

$$\begin{aligned} &\left(\xi^2\widehat{\phi} + 2\xi\widehat{\phi}_t + \widehat{\phi}_{tt}\right) - \gamma\Delta\left(\xi^2\widehat{\phi} + 2\xi\widehat{\phi}_t + \widehat{\phi}_{tt}\right) - \gamma\mathring{\mathbf{A}}G_2\gamma_1\left(\xi^2\widehat{\phi} + 2\xi\widehat{\phi}_t + \widehat{\phi}_{tt}\right) \\ &\qquad + \mathring{\mathbf{A}}\widehat{\phi} + \alpha\mathring{\mathbf{A}}G_1\gamma_0\widehat{\psi} - \alpha\lambda\mathring{\mathbf{A}}G_2\gamma_0\widehat{\psi} + \alpha\Delta\widehat{\psi} = 0 \text{ in } H^{-1}_{\Gamma_0,\gamma}(\Omega), \end{aligned} \tag{2.10}$$

$$\beta\left(\xi\widehat{\psi} + \widehat{\psi}_t\right) + \eta\Delta\widehat{\psi} - \sigma\widehat{\psi} - \alpha\Delta\left(\xi\widehat{\phi} + \widehat{\phi}_t\right) = 0 \text{ in } L^2(\Omega), \tag{2.11}$$

$$\left[\widehat{\phi}(T), \widehat{\phi}_t(T), \widehat{\psi}(T)\right] = \left[e^{-\xi T}\phi_0, -\xi e^{-\xi T}\phi_0 + e^{-\xi T}\phi_1, e^{-\xi T}\psi_0\right]. \tag{2.12}$$

Now multiplying the heat equation (2.11) by $\frac{\alpha}{\eta}$ and adding it to the Kirchoff plate (2.10), and subsequently taking the parameter ξ to be $\xi \equiv \frac{\alpha^2}{2\gamma\eta}$, we obtain the single equation

$$\begin{aligned}\widehat{\phi}_{tt} - \gamma\Delta\widehat{\phi}_{tt} + \mathring{\mathbf{A}}\widehat{\phi} - \gamma\mathring{\mathbf{A}}G_2\gamma_1\left(\xi^2\widehat{\phi} + 2\xi\widehat{\phi}_t + \widehat{\phi}_{tt}\right) + \alpha\mathring{\mathbf{A}}G_1\gamma_0\widehat{\psi} - \alpha\lambda\mathring{\mathbf{A}}G_2\gamma_0\widehat{\psi}\\ = c_0\widehat{\psi} + c_1\widehat{\psi}_t + c_2\widehat{\phi} + c_3\widehat{\phi}_t + c_4\Delta\widehat{\phi}\,, \qquad (2.13)\end{aligned}$$

$$\left[\widehat{\phi}(T), \widehat{\phi}_t(T), \widehat{\psi}(T)\right] = \left[e^{-\xi T}\phi_0, -\xi e^{-\xi T}\phi_0 + e^{-\xi T}\phi_1, e^{-\xi T}\psi_0\right], \qquad (2.14)$$

where the constants $c_0 = \frac{\alpha^3\beta}{2\gamma\eta^2} - \frac{\alpha\sigma}{\eta}$, $c_1 = \frac{\alpha\beta}{\eta}$, $c_2 = -\frac{\alpha^4}{4\gamma^2\eta^2}$, $c_3 = -\frac{\alpha^2}{\gamma\eta}$, and $c_4 = -\frac{\alpha^4}{4\gamma\eta^2}$. (Note that the particular choice of ξ made here eliminates the higher order term $\Delta\widehat{\phi}_t$.) System (2.13)–(2.14) may be rewritten in PDE form as the Kirchoff plate equation

$$\begin{cases} \widehat{\phi}_{tt} - \gamma\Delta\widehat{\phi}_{tt} + \Delta^2\widehat{\phi} = c_0\widehat{\psi} + c_1\widehat{\psi}_t + c_2\widehat{\phi} + c_3\widehat{\phi}_t + c_4\Delta\widehat{\phi} \quad \text{on } (0,\infty)\times\Omega, \\ \widehat{\phi} = \dfrac{\partial\widehat{\phi}}{\partial\nu} = 0 \ \text{ on } (0,\infty)\times\Gamma_0, \\ \begin{cases} \Delta\widehat{\phi} + (1-\mu)B_1\widehat{\phi} = -\alpha\widehat{\psi} \\ \dfrac{\partial\Delta\widehat{\phi}}{\partial\nu} + (1-\mu)\dfrac{\partial B_2\widehat{\phi}}{\partial\tau} - \gamma\dfrac{\partial\widehat{\phi}_{tt}}{\partial\nu} = \gamma\dfrac{\partial}{\partial\nu}\left(\xi^2\widehat{\phi} + 2\xi\widehat{\phi}_t\right) - \alpha\dfrac{\partial\widehat{\psi}}{\partial\nu} \end{cases} \text{on } (0,\infty)\times\Gamma_1, \\ \left[\widehat{\phi}(T), \widehat{\phi}_t(T), \widehat{\psi}(T)\right] = \left[e^{-\xi T}\phi_0, -\xi e^{-\xi T}\phi_0 + e^{-\xi T}\phi_1, e^{-\xi T}\psi_0\right]. \end{cases} \qquad (2.15)$$

As $\widehat{\phi} - \gamma G_2\gamma_1(\xi^2\widehat{\phi} + 2\xi\widehat{\phi}_t + \widehat{\phi}_{tt}) + \alpha G_1\gamma_0\widehat{\psi} - \alpha\lambda G_2\gamma_0\widehat{\psi} \in C([0,T]; D(\mathring{\mathbf{A}}))$ (using the last containment in (2.7)), then $[\widehat{\phi}, \widehat{\phi}_t]$ is a classical solution of (2.15).

We note at this point that one can readily derive the trace estimate Lemma 4.5 (of the appendix below) for the plate component $\Delta\widehat{\phi}|_{\Gamma_0}$ of the solution $[\widehat{\phi}, \widehat{\phi}_t, \widehat{\psi}]$ of (2.9). The proof of this is relegated to the appendix, since it is entirely analogous to that shown for the forward problem in [3] and [4]. This estimate will be critical in the proof of the following lemma, which gives an energy relation for the mechanical variable.

LEMMA 2.2. (a) *The solution* $[\widehat{\phi}, \widehat{\phi}_t, \widehat{\psi}]$ *to* (2.9) *satisfies the following relation for all* s *and* $\tau \in [0,T]$*:*

$$E_{\widehat{\phi}}(t)\Big|_{t=s}^{t=\tau} = \mathcal{F}(s,\tau), \qquad (2.16)$$

where $E_{\widehat{\phi}}(\tau)$ *is the mechanical energy function defined in* (2.3) *and* $\mathcal{F}(\cdot,\cdot)$ *is a function (defined below in* (2.34)*) that obeys the following estimate for all* s *and* $\tau \in [0,T]$ *and* $\epsilon > 0$*:*

$$\begin{aligned}\mathcal{F}(s,\tau) \le C_\epsilon\int_0^T \left\|\nabla\widehat{\phi}_t\right\|^2_{L^2(\Gamma_1)} dt + \epsilon\int_0^T \left[E_{\widehat{\phi}}(t) + \left\|A_R^{\frac{1}{2}}\widehat{\psi}(t)\right\|^2_{L^2(\Omega)}\right] dt\\ +\epsilon\left(E_{\widehat{\phi}}(s) + E_{\widehat{\phi}}(\tau)\right) + \text{l.o.t.}(\widehat{\phi}, \widehat{\phi}_t\widehat{\psi}). \qquad (2.17)\end{aligned}$$

(b) *For* $\epsilon > 0$ *small enough, the solution* $[\widehat{\phi}, \widehat{\phi}_t, \widehat{\psi}]$ *to* (2.9) *satisfies the following*

estimate for all s and $\tau \in [0,T]$:

$$E_{\widehat{\phi}}(\tau) \leq \left(\frac{1+\epsilon}{1-\epsilon}\right) E_{\widehat{\phi}}(s) + C_\epsilon \int_0^T \left\|\nabla \widehat{\phi}_t\right\|^2_{L^2(\Gamma_1)} dt$$
$$+\frac{\epsilon}{1-\epsilon}\int_0^T \left[E_{\widehat{\phi}}(t) + \left\|A_R^{\frac{1}{2}}\widehat{\psi}(t)\right\|^2_{L^2(\Omega)}\right] dt + \mathsf{l.o.t.}(\widehat{\phi}, \widehat{\phi}_t\widehat{\psi}). \tag{2.18}$$

Above, the constant C_ϵ is independent of time.

Proof. We take the duality pairing of the abstract equation (2.13) with $\widehat{\phi}_t$ and integrate in time and space so as to get

(2.19)

$$\int_s^\tau \left[\left\langle \widehat{\phi}_{tt} - \gamma\Delta\widehat{\phi}_{tt} - \gamma \mathring{\mathbf{A}} G_2\gamma_1\widehat{\phi}_{tt}, \widehat{\phi}_t\right\rangle_{H^{-1}_{\Gamma_0}(\Omega)\times H^1_{\Gamma_0}(\Omega)} + \left\langle \mathring{\mathbf{A}}\widehat{\phi}, \widehat{\phi}_t\right\rangle_{\left[D\left(\mathring{\mathbf{A}}^{\frac{1}{2}}\right)\right]'\times D\left(\mathring{\mathbf{A}}^{\frac{1}{2}}\right)}\right] dt$$
$$= \int_s^\tau \left\langle \gamma\mathring{\mathbf{A}}G_2\gamma_1\left(\xi^2\widehat{\phi} + 2\xi\widehat{\phi}_t\right) - \alpha\mathring{\mathbf{A}}G_1\gamma_0\widehat{\psi} + \alpha\lambda\ \mathring{\mathbf{A}}G_2\gamma_0\widehat{\psi}, \widehat{\phi}_t\right\rangle_{\left[D\left(\mathring{\mathbf{A}}^{\frac{1}{2}}\right)\right]'\times D\left(\mathring{\mathbf{A}}^{\frac{1}{2}}\right)} dt$$
$$+ \int_s^\tau \left(c_0\widehat{\psi} + c_1\widehat{\psi}_t + c_2\widehat{\phi} + c_3\widehat{\phi}_t + c_4\Delta\widehat{\phi}, \widehat{\phi}_t\right)_{L^2(\Omega)} dt.$$

(Note that here we are using implicitly the fact that the terminal data $[\phi_0, \phi_1, \psi_0]$ being in $D(\mathcal{A}^*_\gamma)$ implies that $\mathring{\mathbf{A}}\widehat{\phi} + \gamma\mathring{\mathbf{A}}G_2\gamma_1(\xi^2\widehat{\phi} + 2\xi\widehat{\phi}_t + \widehat{\phi}_{tt}) - \alpha\mathring{\mathbf{A}}G_1\gamma_0\widehat{\psi} + \alpha\lambda\mathring{\mathbf{A}}G_2\gamma_0\widehat{\psi}$ is an element of $C([0,T]; H^{-1}_{\Gamma_0}(\Omega))$.) Second, denoting A_D^{-1} to be the inverse of the elliptic operator defined in (1.12), we multiply the PDE (2.15) by $-\frac{c_1}{\gamma}A_D^{-1}\widehat{\psi}$, and subsequently integrate in time and space so as to get

$$-\frac{c_1}{\gamma}\int_s^\tau \left(\widehat{\phi}_{tt} - \gamma\Delta\widehat{\phi}_{tt} + \Delta^2\widehat{\phi} - \left[c_0\widehat{\psi} + c_1\widehat{\psi}_t + c_2\widehat{\phi} + c_3\widehat{\phi}_t + c_4\Delta\widehat{\phi}\right], A_D^{-1}\widehat{\psi}\right)_{L^2(\Omega)} dt = 0. \tag{2.20}$$

(A1) *Rewriting* (2.19). Using equality (1.31) and the characterizations in (1.23), we have upon the taking of adjoints that (2.19) may be rewritten as

$$E_{\widehat{\phi}}(t)\Big|_{t=s}^{t=\tau} = \int_s^\tau \left(c_0\widehat{\psi} + c_2\widehat{\phi} + c_3\widehat{\phi}_t + c_4\Delta\widehat{\phi}, \widehat{\phi}_t\right)_{L^2(\Omega)} dt + \int_s^\tau c_1\left(\widehat{\psi}_t, \widehat{\phi}_t\right)_{L^2(\Omega)} dt$$
$$-\int_s^\tau \left[\left(\gamma\xi^2\frac{\partial\widehat{\phi}}{\partial\nu} + 2\gamma\xi\frac{\partial\widehat{\phi}_t}{\partial\nu} + \alpha\lambda\gamma_0\widehat{\psi}, \widehat{\phi}_t\right)_{L^2(\Gamma_1)} + \alpha\left(\widehat{\psi}, \frac{\partial\widehat{\phi}_t}{\partial\nu}\right)_{L^2(\Gamma_1)}\right] dt\,. \tag{2.21}$$

(A2) *Rewriting* (2.20). (i) An integration by parts, the use of the heat equation (2.11), and the fact that $A_R\widehat{\psi} = -\Delta\widehat{\psi} + \Delta D\gamma_0\widehat{\psi} + \frac{\sigma}{\eta}\widehat{\psi} = A_D(\ \mathbf{I} - D\gamma_0)\widehat{\psi} + \frac{\sigma}{\eta}\widehat{\psi}$ yield

(2.22)

$$\int_s^\tau -\frac{c_1}{\gamma}\left(\widehat{\phi}_{tt}, A_D^{-1}\widehat{\psi}\right)_{L^2(\Omega)} dt = \left[-\frac{c_1}{\gamma}\left(\widehat{\phi}_t, A_D^{-1}\widehat{\psi}\right)_{L^2(\Omega)}\right]_s^\tau$$
$$+\frac{c_1}{\gamma}\int_s^\tau \left(\widehat{\phi}_t, A_D^{-1}\widehat{\psi}_t\right)_{L^2(\Omega)} dt$$

$$= \left[-\frac{c_1}{\gamma} \left(\widehat{\phi}_t, A_D^{-1} \widehat{\psi} \right)_{L^2(\Omega)} \right]_s^\tau$$
$$+ \frac{c_1}{\gamma} \int_s^\tau \left(\widehat{\phi}_t, \left(\frac{\sigma}{\beta} - \xi \right) A_D^{-1} \widehat{\psi} + \frac{\eta}{\beta} \left(\mathbf{I} - D\gamma_0 \right) \widehat{\psi} \right)_{L^2(\Omega)} dt$$
$$- \frac{c_1}{\gamma} \int_s^\tau \left(\widehat{\phi}_t, \frac{\alpha\xi}{\beta} \left(\mathbf{I} - D\gamma_0 \right) \widehat{\phi} + \frac{\alpha}{\beta} \left(\mathbf{I} - D\gamma_0 \right) \widehat{\phi}_t \right)_{L^2(\Omega)} dt.$$

(ii) An integration by parts and employment of Green's theorem yield

(2.23)

$$\int_s^\tau c_1 \left(\Delta \widehat{\phi}_{tt}, A_D^{-1} \widehat{\psi} \right)_{L^2(\Omega)} dt = - \int_s^\tau c_1 \left(\nabla \widehat{\phi}_{tt}, \nabla A_D^{-1} \widehat{\psi} \right)_{L^2(\Omega)} dt$$
$$= -c_1 \left[\left(\nabla \widehat{\phi}_t, \nabla A_D^{-1} \widehat{\psi} \right)_{L^2(\Omega)} \right]_s^\tau + c_1 \int_s^\tau \left(\nabla \widehat{\phi}_t, \nabla A_D^{-1} \widehat{\psi}_t \right)_{L^2(\Omega)} dt$$
$$= - c_1 \left[\left(\nabla \widehat{\phi}_t, \nabla A_D^{-1} \widehat{\psi} \right)_{L^2(\Omega)} \right]_s^\tau + c_1 \int_s^\tau \left(\widehat{\phi}_t, A_D A_D^{-1} \widehat{\psi}_t \right)_{L^2(\Omega)} dt$$
$$+ c_1 \int_s^\tau \left(\widehat{\phi}_t, \frac{\partial A_D^{-1} \widehat{\psi}_t}{\partial \nu} \right)_{L^2(\Gamma_1)} dt$$
$$= -c_1 \left[\left(\nabla \widehat{\phi}_t, \nabla A_D^{-1} \widehat{\psi} \right)_{L^2(\Omega)} \right]_s^\tau + c_1 \int_s^\tau \left(\widehat{\phi}_t, \widehat{\psi}_t \right)_{L^2(\Omega)} dt$$
$$+ c_1 \int_s^\tau \left(\widehat{\phi}_t, \left(\frac{\sigma}{\beta} - \xi \right) \frac{\partial A_D^{-1} \widehat{\psi}}{\partial \nu} + \frac{\eta}{\beta} \frac{\partial (\mathbf{I} - D\gamma_0) \widehat{\psi}}{\partial \nu} \right)_{L^2(\Gamma_1)} dt$$
$$- c_1 \int_s^\tau \left(\widehat{\phi}_t, \frac{\alpha\xi}{\beta} \frac{\partial (\mathbf{I} - D\gamma_0) \widehat{\phi}}{\partial \nu} + \frac{\alpha}{\beta} \frac{\partial (\mathbf{I} - D\gamma_0) \widehat{\phi}_t}{\partial \nu} \right)_{L^2(\Gamma_1)} dt.$$

(iii) Through the use of Green's theorem (1.9) and the boundary conditions in (2.15), we obtain

(2.24)

$$- \int_s^\tau \frac{c_1}{\gamma} \left(\Delta^2 \widehat{\phi}, A_D^{-1} \widehat{\psi} \right)_{L^2(\Omega)} dt$$
$$= - \frac{c_1}{\gamma} \int_s^\tau a \left(\widehat{\phi}, A_D^{-1} \widehat{\psi} \right) dt - \frac{\alpha c_1}{\gamma} \int_s^\tau \left(\widehat{\psi}, \frac{\partial A_D^{-1} \widehat{\psi}}{\partial \nu} \right)_{L^2(\Gamma_1)} dt$$
$$+ \frac{c_1}{\gamma} \int_s^\tau \left(\Delta \widehat{\phi}, \frac{\partial A_D^{-1} \widehat{\psi}}{\partial \nu} \right)_{L^2(\Gamma_0)} dt$$

(where we have used the fact that $\widehat{\phi}|_{\Gamma_0} = \frac{\partial \widehat{\phi}}{\partial \nu}|_{\Gamma_0} = 0$ implies $B_1 \widehat{\phi}|_{\Gamma_0} = 0$; see [11]).

Jointly then, equalities (2.20) and (2.22)–(2.24) give the relation

(2.25)

$$0 = -c_1 \int_s^\tau \left(\widehat{\phi}_t, \widehat{\psi}_t\right)_{L^2(\Omega)} dt - \frac{c_1}{\gamma} \int_s^\tau \left(\widehat{\phi}_t, \left(\frac{\sigma}{\beta} - \xi\right) A_D^{-1}\widehat{\psi} + \frac{\eta}{\beta}\left(\mathbf{I} - D\gamma_0\right)\widehat{\psi}\right)_{L^2(\Omega)} dt$$
$$+ \frac{c_1}{\gamma} \int_s^\tau \left(\widehat{\phi}_t, \frac{\alpha\xi}{\beta}\left(\mathbf{I} - D\gamma_0\right)\widehat{\phi} + \frac{\alpha}{\beta}\left(\mathbf{I} - D\gamma_0\right)\widehat{\phi}_t\right)_{L^2(\Omega)} dt$$
$$- c_1 \int_s^\tau \left(\widehat{\phi}_t, \left(\frac{\sigma}{\beta} - \xi\right)\frac{\partial A_D^{-1}\widehat{\psi}}{\partial\nu} + \frac{\eta}{\beta}\frac{\partial(\mathbf{I} - D\gamma_0)\widehat{\psi}}{\partial\nu}\right)_{L^2(\Gamma_1)} dt$$
$$+ c_1 \int_s^\tau \left(\widehat{\phi}_t, \frac{\alpha\xi}{\beta}\frac{\partial(\mathbf{I} - D\gamma_0)\widehat{\phi}}{\partial\nu} + \frac{\alpha}{\beta}\frac{\partial(\mathbf{I} - D\gamma_0)\widehat{\phi}_t}{\partial\nu}\right)_{L^2(\Gamma_1)} dt$$
$$+ \frac{c_1}{\gamma} \int_s^\tau a\left(\widehat{\phi}, A_D^{-1}\widehat{\psi}\right) dt + \frac{\alpha c_1}{\gamma} \int_s^\tau \left(\widehat{\psi}, \frac{\partial A_D^{-1}\widehat{\psi}}{\partial\nu}\right)_{L^2(\Gamma_1)} dt$$
$$- \frac{c_1}{\gamma} \int_s^\tau \left(\Delta\widehat{\phi}, \frac{\partial A_D^{-1}\widehat{\psi}}{\partial\nu}\right)_{L^2(\Gamma_0)} dt$$
$$- \int_s^\tau \left(c_0\widehat{\psi} + c_1\widehat{\psi}_t + c_2\widehat{\phi} + c_3\widehat{\phi}_t + c_4\Delta\widehat{\phi}, \frac{c_1}{\gamma}A_D^{-1}\widehat{\psi}\right)_{L^2(\Omega)} dt$$
$$+ c_1\left[\left(\nabla\widehat{\phi}_t, \nabla A_D^{-1}\widehat{\psi}\right)_{L^2(\Omega)} + \frac{1}{\gamma}\left(\widehat{\phi}_t, A_D^{-1}\widehat{\psi}\right)_{L^2(\Omega)}\right]_s^\tau.$$

Summing the relations (2.21) and (2.25), we obtain

(2.26)

$$E_{\widehat{\phi}}(t)\Big|_{t=s}^{t=\tau} = \int_s^\tau \left(c_0\widehat{\psi} + c_2\widehat{\phi} + c_3\widehat{\phi}_t + c_4\Delta\widehat{\phi}, \widehat{\phi}_t - \frac{c_1}{\gamma}A_D^{-1}\widehat{\psi}\right)_{L^2(\Omega)} dt$$
$$- \int_s^\tau \left[\left(\gamma\xi^2\frac{\partial\widehat{\phi}}{\partial\nu} + 2\gamma\xi\frac{\partial\widehat{\phi}_t}{\partial\nu} + \alpha\lambda\gamma_0\widehat{\psi}, \widehat{\phi}_t\right)_{L^2(\Gamma_1)} + \alpha\left(\widehat{\psi}, \frac{\partial\widehat{\phi}_t}{\partial\nu}\right)_{L^2(\Gamma_1)}\right] dt$$
$$- \frac{c_1}{\gamma} \int_s^\tau \left(\widehat{\phi}_t, \left(\frac{\sigma}{\beta} - \xi\right) A_D^{-1}\widehat{\psi} + \frac{\eta}{\beta}\left(\mathbf{I} - D\gamma_0\right)\widehat{\psi}\right)_{L^2(\Omega)} dt + c_1\left[\left(\nabla\widehat{\phi}_t, \nabla A_D^{-1}\widehat{\psi}\right)_{L^2(\Omega)}\right]_s^\tau$$
$$+ c_1 \int_s^\tau \left[\left(\frac{\widehat{\phi}_t}{\gamma}, \frac{\alpha\xi}{\beta}\left(\mathbf{I} - D\gamma_0\right)\widehat{\phi} + \frac{\alpha}{\beta}\left(\mathbf{I} - D\gamma_0\right)\widehat{\phi}_t\right)_{L^2(\Omega)}\right.$$
$$\left. - \left(\widehat{\phi}_t, \left(\frac{\sigma}{\beta} - \xi\right)\frac{\partial A_D^{-1}\widehat{\psi}}{\partial\nu} + \frac{\eta}{\beta}\frac{\partial(\mathbf{I} - D\gamma_0)\widehat{\psi}}{\partial\nu}\right)_{L^2(\Gamma_1)}\right] dt$$
$$+ c_1 \int_s^\tau \left(\widehat{\phi}_t, \frac{\alpha\xi}{\beta}\frac{\partial(\mathbf{I} - D\gamma_0)\widehat{\phi}}{\partial\nu} + \frac{\alpha}{\beta}\frac{\partial(\mathbf{I} - D\gamma_0)\widehat{\phi}_t}{\partial\nu}\right)_{L^2(\Gamma_1)} dt + \frac{c_1}{\gamma} \int_s^\tau a\left(\widehat{\phi}, A_D^{-1}\widehat{\psi}\right)_{L^2(\Omega)} dt$$
$$+ \frac{c_1}{\gamma} \int_s^\tau \left[\left(\alpha\widehat{\psi}, \frac{\partial A_D^{-1}\widehat{\psi}}{\partial\nu}\right)_{L^2(\Gamma_1)} - \left(\Delta\widehat{\phi}, \frac{\partial A_D^{-1}\widehat{\psi}}{\partial\nu}\right)_{L^2(\Gamma_0)}\right] dt + \frac{c_1}{\gamma}\left[\left(\widehat{\phi}_t, A_D^{-1}\widehat{\psi}\right)_{L^2(\Omega)}\right]_s^\tau$$

(note the cancellation of the high order term $\int_s^\tau (\widehat{\psi}_t, \widehat{\phi}_t)_{L^2(\Omega)} dt$).

We now proceed to estimate the right-hand side of this relation. In so doing, we will be using implicitly, in (B1)–(B7) below, the inequality $ab \leq \epsilon a^2 + C_\epsilon b^2$.

(B1) We have by trace theory

$$-\int_s^\tau \left[\left(\gamma\xi^2\frac{\partial\widehat{\phi}}{\partial\nu}+2\gamma\xi\frac{\partial\widehat{\phi}_t}{\partial\nu}+\alpha\lambda\gamma_0\widehat{\psi},\widehat{\phi}_t\right)_{L^2(\Gamma_1)}+\alpha\left(\widehat{\psi},\frac{\partial\widehat{\phi}_t}{\partial\nu}\right)_{L^2(\Gamma_1)}\right]dt$$

$$\leq C\int_0^T\left\|\frac{\partial\widehat{\phi}_t}{\partial\nu}\right\|^2_{L^2(\Gamma_1)}dt+\text{l.o.t.}\left(\widehat{\phi},\widehat{\phi}_t,\widehat{\psi}\right). \tag{2.27}$$

(B2) As A_D^{-1} is a bounded operator, we have

$$\int_s^\tau\left(c_0\widehat{\psi}+c_2\widehat{\phi}+c_3\widehat{\phi}_t+c_4\Delta\widehat{\phi},\widehat{\phi}_t-\frac{c_1}{\gamma}A_D^{-1}\widehat{\psi}\right)dt$$

$$\leq\frac{\epsilon}{6}\int_0^T\left\|\mathring{\mathbf{A}}^{\frac{1}{2}}\widehat{\phi}\right\|^2_{L^2(\Omega)}dt+\text{l.o.t.}\left(\widehat{\phi},\widehat{\phi}_t,\widehat{\psi}\right). \tag{2.28}$$

(B3) As $D\gamma_0\in\mathcal{L}(H^s(\Omega))$ for $s>\frac{1}{2}$ (by standard elliptic theory), and $A_D^{-1}\in\mathcal{L}(L^2(\Omega),D(A_D))$, we then have in conjunction with trace theory

(2.29)

$$-\frac{c_1}{\gamma}\int_s^\tau\left(\widehat{\phi}_t,\left(\frac{\sigma}{\beta}-\xi\right)A_D^{-1}\widehat{\psi}+\frac{\eta}{\beta}\left(\mathbf{I}-D\gamma_0\right)\widehat{\psi}\right)_{L^2(\Omega)}dt$$

$$+\frac{c_1}{\gamma}\int_s^\tau\left(\widehat{\phi}_t,\frac{\alpha\xi}{\beta}\left(\mathbf{I}-D\gamma_0\right)\widehat{\phi}+\frac{\alpha}{\beta}\left(\mathbf{I}-D\gamma_0\right)\widehat{\phi}_t\right)_{L^2(\Omega)}dt$$

$$-c_1\int_s^\tau\left(\widehat{\phi}_t,\left(\frac{\sigma}{\beta}-\xi\right)\frac{\partial A_D^{-1}\widehat{\psi}}{\partial\nu}\right)_{L^2(\Gamma_1)}dt$$

$$+\frac{c_1}{\gamma}\int_s^\tau a\left(\widehat{\phi},A_D^{-1}\widehat{\psi}\right)dt+\frac{\alpha c_1}{\gamma}\int_s^\tau\left(\widehat{\psi},\frac{\partial A_D^{-1}\widehat{\psi}}{\partial\nu}\right)_{L^2(\Gamma_1)}dt$$

$$\leq\frac{\epsilon}{6}\int_0^T\left\|\mathring{\mathbf{A}}^{\frac{1}{2}}\widehat{\phi}\right\|^2_{L^2(\Omega)}dt+\text{l.o.t.}\left(\widehat{\phi},\widehat{\phi}_t,\widehat{\psi}\right).$$

(B4) Using the fact that $D\gamma_0\in\mathcal{L}(H^s(\Omega))$ for $s>\frac{1}{2}$, and $\frac{\partial\widehat{\psi}}{\partial\nu}(t)|_\Gamma=-\lambda\widehat{\psi}(t)|_\Gamma$, we have along with trace theory that

$$c_1\int_s^\tau\left(\widehat{\phi}_t,-\frac{\eta}{\beta}\frac{\partial\widehat{\psi}}{\partial\nu}+\frac{\alpha\xi}{\beta}\frac{\partial(\mathbf{I}-D\gamma_0)\widehat{\phi}}{\partial\nu}+\frac{\alpha}{\beta}\frac{\partial\widehat{\phi}_t}{\partial\nu}\right)_{L^2(\Gamma_1)}dt$$

$$\leq C\int_0^T\left\|\frac{\partial\widehat{\phi}_t}{\partial\nu}\right\|^2_{L^2(\Gamma_1)}dt+\text{l.o.t.}\left(\widehat{\phi},\widehat{\phi}_t,\widehat{\psi}\right). \tag{2.30}$$

(B5) By [1, p. 311, Theorem 3] and trace theory we deduce that $\frac{\partial}{\partial\nu}D\gamma_0\in\mathcal{L}(H^1(\Omega),H^{-\frac{1}{2}}(\Gamma))$, and so accordingly we have

$$c_1\int_s^\tau\left(\widehat{\phi}_t,\frac{\eta}{\beta}\frac{\partial D\gamma_0\widehat{\psi}}{\partial\nu}-\frac{\alpha}{\beta}\frac{\partial D\gamma_0\widehat{\phi}_t}{\partial\nu}\right)_{L^2(\Gamma_1)}dt\leq C_\epsilon\int_0^T\left\|\widehat{\phi}_t\right\|^2_{H^{\frac{1}{2}}(\Gamma_1)}dt$$

$$+\frac{\epsilon}{6}\int_0^T\left[\left\|P_\gamma^{\frac{1}{2}}\widehat{\phi}_t\right\|^2_{L^2(\Omega)}+\left\|A_R^{\frac{1}{2}}\widehat{\psi}\right\|^2_{L^2(\Omega)}\right]dt. \tag{2.31}$$

(B6) As $A_D^{-1} \in \mathcal{L}(H^{-1}(\Omega), H_0^1(\Omega))$, by the characterizations of elliptic operators given in [7] , we then have for all $t \in [0,T]$

$$\begin{aligned}\left(\nabla\widehat{\phi}_t(t), \nabla A_D^{-1}\widehat{\psi}(t)\right)_{L^2(\Omega)} &\leq C \left\|\nabla\widehat{\phi}_t(t)\right\|_{L^2(\Omega)} \left\|\nabla A_D^{-1}\widehat{\psi}(t)\right\|_{L^2(\Omega)} \\ &\leq \frac{\epsilon}{6}\left\|P_\gamma^{\frac{1}{2}}\widehat{\phi}_t(t)\right\|_{L^2(\Omega)}^2 + \text{l.o.t.}\left(\widehat{\phi}, \widehat{\phi}_t, \widehat{\psi}\right).\end{aligned}$$

We thus have

$$\begin{aligned}&c_1\left[\left(\nabla\widehat{\phi}_t(t), \nabla A_D^{-1}\widehat{\psi}\right)_{L^2(\Omega)}\right]_{t=s}^{t=\tau} + \left[\frac{c_1}{\gamma}\left(\widehat{\phi}_t(t), A_D^{-1}\widehat{\psi}(t)\right)_{L^2(\Omega)}\right]_{t=s}^{t=\tau} \\ &\leq \frac{\epsilon}{6}\left(\left\|P_\gamma^{\frac{1}{2}}\widehat{\phi}_t(\tau)\right\|_{L^2(\Omega)}^2 + \left\|P_\gamma^{\frac{1}{2}}\widehat{\phi}_t(s)\right\|_{L^2(\Omega)}^2\right) + \text{l.o.t.}\left(\widehat{\phi}, \widehat{\phi}_t, \widehat{\psi}\right). \qquad (2.32)\end{aligned}$$

(B7) Finally, we can use the trace result Lemma 4.5 of the appendix and the fact that $A_D^{-1} \in \mathcal{L}(H^{-\frac{1}{2}+\epsilon}(\Omega), H^{\frac{3}{2}+\epsilon}(\Omega)$ (again by [7]) to have

(2.33)

$$\begin{aligned}&-\frac{c_1}{\gamma}\int_s^\tau \left(\Delta\widehat{\phi}, \frac{\partial A_D^{-1}\widehat{\psi}}{\partial\nu}\right)_{L^2(\Gamma_0)} dt \\ &\leq C\int_s^\tau \left\|\Delta\widehat{\phi}\right\|_{L^2(\Gamma_0)}\left\|\widehat{\psi}\right\|_{H^{-\frac{1}{2}+\epsilon}(\Omega)} dt \leq \frac{\epsilon}{6C_0}\int_s^\tau \left\|\Delta\widehat{\phi}\right\|_{L^2(\Gamma_0)}^2 dt + C_\epsilon\int_0^T \left\|\widehat{\psi}\right\|_{H^{-\frac{1}{2}+\epsilon}(\Omega)}^2 dt \\ &\text{(where the constant } C_0 \text{ above is the very same as that in (4.13))} \\ &\leq \frac{\epsilon}{3}\int_0^T E_{\widehat{\phi}}(t)dt + \frac{\epsilon}{3}\left[E_{\widehat{\phi}}(s) + E_{\widehat{\phi}}(\tau)\right] + \text{l.o.t.}\left(\widehat{\phi}, \widehat{\phi}_t, \widehat{\psi}\right).\end{aligned}$$

Therefore, if we define $\mathcal{F}(s,\tau)$ to be

$$\mathcal{F}(s,\tau) \equiv \text{right-hand side of (2.26)}, \qquad (2.34)$$

estimates (2.27)–(2.33), then we have

$$\begin{aligned}\mathcal{F}(s,\tau) \leq C_\epsilon\int_0^T \left\|\nabla\widehat{\phi}_t\right\|_{L^2(\Gamma_1)}^2 dt + \epsilon\int_0^T \left[E_{\widehat{\phi}}(t) + \left\|A_R^{\frac{1}{2}}\widehat{\psi}(t)\right\|_{L^2(\Omega)}^2\right] dt \\ +\epsilon\left[E_{\widehat{\phi}}(s) + E_{\widehat{\phi}}(\tau)\right] + \text{l.o.t.}\left(\widehat{\phi}, \widehat{\phi}_t, \widehat{\psi}\right), \qquad (2.35)\end{aligned}$$

where the constant C_ϵ does not depend on time T. This and equality (2.26) prove (a).

To prove (b), we combine (2.16) and (2.17) and subsequently take $\epsilon > 0$ small enough. The proof of Lemma 2.2 is concluded. □

With the radial vector field $\overline{h}$ defined in (1.3) , one has the following relation, which is essentially demonstrated in [12] (the complete proof is carried out in Proposition 4.6 of the appendix below).

PROPOSITION 2.3. *With the vector field $\overline{h}$ as defined in* (1.3)*, the solution $[\widehat{\phi}, \widehat{\phi}_t, \widehat{\psi}]$ to* (2.15)*, corresponding to terminal data $[\phi_0, \phi_1, \psi_0] \in D([\mathcal{A}_\gamma^*]^2)$, satisfies the following equality for arbitrary $\epsilon_0 \in [0,T)$:*

(2.36)

$$\int_{\epsilon_0}^{T-\epsilon_0} E_{\widehat{\phi}}(t)dt = \int_{\epsilon_0}^{T-\epsilon_0} \left(c_0\widehat{\psi} + c_2\widehat{\phi} + c_3\widehat{\phi}_t + c_4\Delta\widehat{\phi}, \overline{h}\cdot\nabla\widehat{\phi} - \frac{1}{2}\widehat{\phi} \right)_{L^2(\Omega)} dt$$

$$- c_1 \int_{\epsilon_0}^{T-\epsilon_0} \left(\widehat{\psi}, \overline{h}\cdot\nabla\widehat{\phi}_t - \frac{1}{2}\widehat{\phi}_t \right)_{L^2(\Omega)} dt + \frac{1}{2}\int_{\epsilon_0}^{T-\epsilon_0}\int_{\Gamma_1} \overline{h}\cdot\nu \left(\widehat{\phi}_t^2 + \gamma \left| \nabla\widehat{\phi}_t \right|^2 \right) d\Gamma dt$$

$$- \int_{\epsilon_0}^{T-\epsilon_0} \left\| \widehat{\phi}_t \right\|^2_{L^2(\Omega)} dt + \frac{1}{2}\int_{\epsilon_0}^{T-\epsilon_0}\int_{\Gamma_0} \overline{h}\cdot\nu \left(\Delta\widehat{\phi} \right)^2 d\Gamma dt$$

$$- \int_{\epsilon_0}^{T-\epsilon_0} \left[\alpha \left(\widehat{\psi}, \frac{\partial}{\partial\nu}\left(\overline{h}\cdot\nabla\widehat{\phi} - \frac{1}{2}\widehat{\phi} \right) \right)_{L^2(\Gamma_1)} \right.$$

$$\left. + \left(\gamma\frac{\partial}{\partial\nu}\left(\xi^2\widehat{\phi} + 2\xi\widehat{\phi}_t \right) - \alpha\frac{\partial\widehat{\psi}}{\partial\nu}, \overline{h}\cdot\nabla\widehat{\phi} - \frac{1}{2}\widehat{\phi} \right)_{L^2(\Gamma_1)} \right] dt$$

$$- \left[(\widehat{\phi}_t, \overline{h}\cdot\nabla\widehat{\phi})_{L^2(\Omega)} + \gamma\left(\nabla\widehat{\phi}_t, \nabla\left(\overline{h}\cdot\nabla\widehat{\phi} \right) \right)_{L^2(\Omega)} - \frac{1}{2}(\widehat{\phi}_t, \widehat{\phi})_{L^2(\Omega)} \right]_{t=\epsilon_0}^{t=T-\epsilon_0}$$

$$+ \left[\frac{\gamma}{2}\left(\nabla\widehat{\phi}_t, \nabla\widehat{\phi} \right)_{L^2(\Omega)} + c_1\left(\widehat{\psi}, \overline{h}\cdot\nabla\widehat{\phi} - \frac{1}{2}\widehat{\phi} \right)_{L^2(\Omega)} \right]_{t=\epsilon_0}^{t=T-\epsilon_0}$$

$$- \int_{\epsilon_0}^{T-\epsilon_0}\int_{\Gamma_1} \frac{\overline{h}\cdot\nu}{2} \left[\left(\frac{\partial^2\widehat{\phi}}{\partial x^2} \right)^2 + \left(\frac{\partial^2\widehat{\phi}}{\partial y^2} \right)^2 \right.$$

$$\left. + 2\mu\left(\frac{\partial^2\widehat{\phi}}{\partial x^2} \right)\left(\frac{\partial^2\widehat{\phi}}{\partial y^2} \right) + 2(1-\mu)\left(\frac{\partial^2\widehat{\phi}}{\partial x\partial y} \right)^2 \right] dtd\Gamma.$$

So as to derive another intermediate energy inequality, we will now estimate the right-hand side of the relation (2.36). In the course of this work, we will make critical use of the following trace estimate for (uncoupled) Kirchoff plates, which was derived in [15]. It is this regularity result that allows the controlled portion Γ_1 of the boundary to be free of geometric constraints.

TRACE THEOREM (see [15]). *Let the function $\varphi(t,x)$ satisfy the following Kirchoff equation on an open, bounded domain $\Omega \subset \mathbb{R}^n$, with smooth boundary Γ, $\Gamma = \Gamma_0 \cup \Gamma_1$, where each Γ_i is open and nonempty, with $\overline{\Gamma}_0 \cap \overline{\Gamma}_1 = \emptyset$:*

$$\text{(2.37)} \qquad \begin{cases} \varphi_{tt} - \gamma\Delta\varphi_{tt} + \Delta^2\varphi = f \quad \text{on } (0,T)\times\Omega, \\ \varphi = \dfrac{\partial\varphi}{\partial\nu} = 0 \quad \text{on } (0,T)\times\Gamma_0, \\ \begin{cases} \Delta\varphi + (1-\mu)B_1\varphi = g_1 \\ \dfrac{\partial\Delta\varphi}{\partial\nu} + (1-\mu)\dfrac{\partial B_2\varphi}{\partial\tau} - \gamma\dfrac{\partial\varphi_{tt}}{\partial\nu} = g_2 \end{cases} \quad \text{on } (0,T)\times\Gamma_1 \end{cases}$$

(here the boundary operators B_1 and B_2 are as given in (1.2)). Let $0 < \epsilon_0 < \frac{T}{2}$ and $\epsilon > 0$ be arbitrary. Then the following inequality holds true for the solution φ:

(2.38)

$$\int_{\epsilon_0}^{T-\epsilon_0} \left[\left\| \frac{\partial^2\varphi}{\partial\tau^2} \right\|^2_{L^2(\Gamma_1)} + \left\| \frac{\partial^2\varphi}{\partial\nu^2} \right\|^2_{L^2(\Gamma_1)} + \left\| \frac{\partial^2\varphi}{\partial\tau\partial\nu} \right\|^2_{L^2(\Gamma_1)} \right] dt$$

$$\leq C_{T,\epsilon_0,\gamma}\left\{\left[\int_0^T \|f\|^2_{[H^{\frac{3}{2}-\epsilon}(\Omega)]'} + \|g_1\|^2_{L^2(\Gamma_1)} + \|\varphi\|^2_{H^{\frac{3}{2}+\epsilon}(\Gamma_1)} + \||\nabla\varphi_t|\|^2_{L^2(\Gamma_1)} + \|\varphi_t\|^2_{L^2(\Gamma_1)}\right]dt + \|g_2\|^2_{H^{-1}(0,T\times\Gamma_1)}\right\}.$$

Remark 2.4. In the original statement of this theorem (see Theorem 2.1 in [15]), the term $\int_0^T \|f\|^2_{[H^{\frac{3}{2}-\epsilon}(\Omega)]'}\,dt$ in the inequality (2.38) is replaced by $\|f\|^2_{[H^q(0,T\times\Omega)]'}$, where $q < \frac{1}{2}$. However, if one replaces the $H^{-q}(0,T\times\Omega)$ spaces with $L^2(0,T;[H^q(\Omega)]')$, the values of allowed parameters extend to $q < 3/2 + \epsilon$. This is in line with elliptic theory corresponding to free boundary conditions.

By the use of this trace result in part, we have the following energy estimate.

LEMMA 2.5. *For all $\epsilon_0 \in (0, \frac{T}{2})$ and $\widetilde{\epsilon} > 0$ arbitrary, the solution $[\widehat{\phi}, \widehat{\phi}_t, \widehat{\psi}]$ to* (2.15) *satisfies*

$$\int_{\epsilon_0}^{T-\epsilon_0} E_{\widehat{\phi}}(t)dt \leq C^*\left(E_{\widehat{\phi}}(T-\epsilon_0) + E_{\widehat{\phi}}(\epsilon_0)\right)$$
$$+ C_T\int_0^T \left\|\nabla\widehat{\phi}_t\right\|^2_{L^2(\Gamma_1)} + \widetilde{\epsilon}\int_0^T \left\|A_R^{\frac{1}{2}}\widehat{\psi}\right\|^2_{L^2(\Omega)} + \text{l.o.t.}\ (\widehat{\phi}, \widehat{\phi}_t, \widehat{\psi}), \tag{2.39}$$

where the (time independent) constant $C^* \geq \sqrt{\frac{2\gamma}{1-\mu}}\max_{[x,y]\in\overline{\Omega}}\left|\overline{h}(x,y)\right|$ (where, again, μ is Poisson's ratio and $\overline{h}$ satisfies 1.3*).*

Proof. We proceed to majorize the right-hand side of (2.36).

(A.1) Handling the term $\int_{\epsilon_0}^{T-\epsilon_0}(c_0\widehat{\psi} + c_2\widehat{\phi} + c_3\widehat{\phi}_t + c_4\Delta\widehat{\phi}, \overline{h}\cdot\nabla\widehat{\phi} - \frac{1}{2}\widehat{\phi})_{L^2(\Omega)}dt$: First, by Green's theorem and the fact that $\nabla \in \mathcal{L}\left(H^s(\Omega), H^{s-1}(\Omega)\right)$ and $\nabla\left(\overline{h}\cdot\nabla\right) \in \mathcal{L}\left(H^s(\Omega), H^{s-2}(\Omega)\right)$, we obtain

$$\begin{aligned}\left(\Delta\widehat{\phi}, \overline{h}\cdot\nabla\widehat{\phi} - \frac{1}{2}\widehat{\phi}\right)_{L^2(\Omega)} &= -\left(\nabla\widehat{\phi}, \nabla\left(\overline{h}\cdot\nabla\widehat{\phi}\right) - \frac{1}{2}\widehat{\nabla\phi}\right)_{L^2(\Omega)} \\ &\quad + \left(\frac{\partial\widehat{\phi}}{\partial\nu}, \nabla\widehat{\phi} - \frac{1}{2}\widehat{\phi}\right)_{L^2(\Gamma_1)} \\ &= -\left(\nabla\widehat{\phi}, \nabla\left(\overline{h}\cdot\nabla\widehat{\phi}\right) - \frac{1}{2}\widehat{\nabla\phi}\right)_{H^\epsilon(\Omega)\times H^{-\epsilon}(\Omega)} + \left(\frac{\partial\widehat{\phi}}{\partial\nu}, \nabla\widehat{\phi} - \frac{1}{2}\widehat{\phi}\right)_{L^2(\Gamma_1)} \\ &\leq \text{l.o.t.}\left(\widehat{\phi}, \widehat{\phi}_t, \widehat{\psi}\right),\end{aligned}$$

where in the last step we have also used Cauchy–Schwarz and the trace theory. We thus have

$$\int_{\epsilon_0}^{T-\epsilon_0}\left(c_0\widehat{\psi} + c_2\widehat{\phi} + c_3\widehat{\phi}_t + c_4\Delta\widehat{\phi}, \overline{h}\cdot\nabla\widehat{\phi} - \frac{1}{2}\widehat{\phi}\right)_{L^2(\Omega)}dt \leq \text{l.o.t.}\left(\widehat{\phi}, \widehat{\phi}_t, \widehat{\psi}\right). \tag{2.40}$$

(A.2) Likewise using Sobolev trace theory, the fact that $\frac{\partial\widehat{\psi}}{\partial\nu} = -\lambda\widehat{\psi}$, and the divergence theorem, we have

(2.41)

$$
\begin{aligned}
&- \int_{\epsilon_0}^{T-\epsilon_0} \left[c_1 \left(\widehat{\psi}, \overline{h} \cdot \nabla \widehat{\phi}_t - \frac{1}{2} \widehat{\phi}_t \right)_{L^2(\Omega)} + \left\| \widehat{\phi}_t \right\|^2_{L^2(\Omega)} \right] dt \\
&+ \int_{\epsilon_0}^{T-\epsilon_0} \left[- \left(\gamma \frac{\partial}{\partial \nu} \left(\xi^2 \widehat{\phi} + 2\xi \widehat{\phi}_t \right) - \alpha \frac{\partial \widehat{\psi}}{\partial \nu}, \overline{h} \cdot \nabla \widehat{\phi} - \frac{1}{2} \widehat{\phi} \right)_{L^2(\Gamma_1)} + \frac{\alpha}{2} \left(\widehat{\psi}, \frac{\partial \widehat{\phi}}{\partial \nu} \right)_{L^2(\Gamma_1)} \right] dt \\
&\leq c_1 \int_{\epsilon_0}^{T-\epsilon_0} \left(\widehat{\psi}, h_1 \widehat{\phi}_{tx} + h_2 \widehat{\phi}_{ty} \right)_{L^2(\Omega)} dt + C \int_0^T \left\| \nabla \widehat{\phi}_t \right\|^2_{L^2(\Gamma_1)} dt + \text{l.o.t. } (\widehat{\phi}, \widehat{\phi}_t, \widehat{\psi}) \\
&= -c_1 \int_{\epsilon_0}^{T-\epsilon_0} \int_\Omega \operatorname{div} \left(\widehat{\psi h} \right) \widehat{\phi}_t d\Omega dt \\
&\quad + c_1 \int_{\epsilon_0}^{T-\epsilon_0} \int_{\Gamma_1} \overline{h} \cdot \nu \, \widehat{\psi} \widehat{\phi}_t d\Gamma dt + C \int_0^T \left\| \nabla \widehat{\phi}_t \right\|^2_{L^2(\Gamma_1)} dt + \text{l.o.t. } (\widehat{\phi}, \widehat{\phi}_t, \widehat{\psi}) \\
&\leq C \int_0^T \left\| \nabla \widehat{\phi}_t \right\|^2_{L^2(\Gamma_1)} dt + \widetilde{\epsilon} \int_0^T \left\| A_R^{\frac{1}{2}} \widehat{\psi} \right\|^2_{L^2(\Omega)} + \text{l.o.t. } (\widehat{\phi}, \widehat{\phi}_t, \widehat{\psi}).
\end{aligned}
$$

(A.3) Using (1.3), we have

$$
\frac{1}{2} \int_{\epsilon_0}^{T-\epsilon_0} \int_{\Gamma_0} \overline{h} \cdot \nu \left(\Delta \widehat{\phi} \right)^2 \leq 0. \tag{2.42}
$$

(A.4) We now estimate the terms

$$
\begin{aligned}
&- \left[(\widehat{\phi}_t, \overline{h} \cdot \nabla \widehat{\phi})_{L^2(\Omega)} + \gamma \left(\nabla \widehat{\phi}_t, \nabla \left(\overline{h} \cdot \nabla \widehat{\phi} \right) \right)_{L^2(\Omega)} - \frac{1}{2} (\widehat{\phi}_t, \widehat{\phi})_{L^2(\Omega)} \right]_{t=\epsilon_0}^{t=T-\epsilon_0} \\
&+ \left[\frac{\gamma}{2} \left(\nabla \widehat{\phi}_t, \nabla \widehat{\phi} \right)_{L^2(\Omega)} + c_1 \left(\widehat{\psi}, \overline{h} \cdot \nabla \widehat{\phi} - \frac{1}{2} \widehat{\phi} \right)_{L^2(\Omega)} \right]_{t=\epsilon_0}^{t=T-\epsilon_0}.
\end{aligned} \tag{2.43}
$$

First, as $\overline{h} \cdot \nabla \widehat{\phi}(t) \in H^{\frac{1}{2}-\epsilon}(\Omega)$ for all $t \in [0,T]$, we have

(2.44)

$$
\begin{aligned}
&\left(\widehat{\psi}(t), \overline{h} \cdot \nabla \widehat{\phi}(t) - \frac{1}{2} \widehat{\phi}(t) \right)_{L^2(\Omega)} = \left\langle \widehat{\psi}(t), \overline{h} \cdot \nabla \widehat{\phi}(t) - \frac{1}{2} \widehat{\phi}(t) \right\rangle_{H^{-\frac{1}{2}+\epsilon}(\Omega) \times H^{\frac{1}{2}-\epsilon}(\Omega)} \\
&\leq \text{l.o.t.} \left(\widehat{\phi}, \widehat{\phi}_t, \widehat{\psi} \right).
\end{aligned}
$$

Second, we have pointwise in time

$$
\begin{aligned}
&\gamma \left(\nabla \widehat{\phi}_t, \nabla \left(\overline{h} \cdot \nabla \widehat{\phi} \right) \right)_{L^2(\Omega)} + \frac{\gamma}{2} \left(\nabla \widehat{\phi}_t, \nabla \widehat{\phi} \right)_{L^2(\Omega)} \\
&= \sqrt{\gamma} \int_\Omega \sqrt{\gamma} \nabla \widehat{\phi}_t(x,y) \cdot [(x-x_0)\phi_{xx} + (y-y_0)\phi_{xy}, (x-x_0)\phi_{xy} + (y-y_0)\phi_{yy}] \, dxdy \\
&\quad + \frac{3\gamma}{2} \left(\nabla \widehat{\phi}_t, \nabla \widehat{\phi} \right)_{L^2(\Omega)}.
\end{aligned} \tag{2.45}
$$

Now, to handle the first term on the right-hand side of (2.45), we use the inequality $ab \leq \frac{\delta}{2} a^2 + \frac{1}{2\delta} b^2$ with $\delta \equiv \sqrt{2(1-\mu)}$ (where, again, Poisson's ratio $\mu \in (0, \frac{1}{2})$)

(2.46)

$$\begin{aligned}
&\int_\Omega \sqrt{\gamma}\nabla\widehat{\phi}_t(x,y)\cdot[(x-x_0)\phi_{xx}+(y-y_0)\phi_{xy},(x-x_0)\phi_{xy}+(y-y_0)\phi_{yy}]\,dxdy\\
&=\int_\Omega (x-x_0)\sqrt{\gamma}\nabla\widehat{\phi}_t\cdot[\phi_{xx},\phi_{xy}]\,dxdy+\int_\Omega (y-y_0)\sqrt{\gamma}\nabla\widehat{\phi}_t\cdot[\phi_{xy},\phi_{yy}]\,dxdy\\
&\leq \max_{[x,y]\in\overline{\Omega}}\left|\overline{h}(x,y)\right|\left\{\frac{\gamma}{\sqrt{2\,(1-\mu)}}\int_\Omega |\nabla\phi_t|^2\,d\Omega+\frac{\sqrt{(1-\mu)}}{\sqrt{2}}\int_\Omega\left[\phi_{xx}^2+\phi_{yy}^2\right]d\Omega\right.\\
&\qquad\qquad\left.+\ \sqrt{2\,(1-\mu)}\int_\Omega\phi_{xy}^2 d\Omega\right\}\\
&\leq \frac{1}{\sqrt{2\,(1-\mu)}}\max_{[x,y]\in\overline{\Omega}}\left|\overline{h}(x,y)\right|\left\{\left\|P_\gamma^{\frac{1}{2}}\widehat{\phi}_t\right\|_{L^2(\Omega)}^2+(1-\mu)\int_\Omega\left[\phi_{xx}^2+\phi_{yy}^2\right]d\Omega\right.\\
&\qquad\qquad\left.+2(1-\mu)\int_\Omega\phi_{xy}^2 d\Omega\right\}.
\end{aligned}$$

From this inequality, the definition of $a(\cdot,\cdot)$ in (1.10), and the characterization in (1.11), we obtain

(2.47)

$$\begin{aligned}
&\int_\Omega \sqrt{\gamma}\nabla\widehat{\phi}_t(x,y)\cdot[(x-x_0)\phi_{xx}+(y-y_0)\phi_{xy},(x-x_0)\phi_{xy}+(y-y_0)\phi_{yy}]\,dxdy\\
&\leq \frac{1}{\sqrt{2\,(1-\mu)}}\max_{[x,y]\in\overline{\Omega}}\left|\overline{h}(x,y)\right|\left\{\left\|P_\gamma^{\frac{1}{2}}\widehat{\phi}_t\right\|_{L^2(\Omega)}^2+\left\|\mathring{\mathbf{A}}^{\frac{1}{2}}\widehat{\phi}\right\|_{L^2(\Omega)}^2\right\}.
\end{aligned}$$

To deal with the second term on the right-hand side of (2.45), we can use the fact that $\nabla\in\mathcal{L}\left(H^s(\Omega),H^{1-s}(\Omega)\right)$ for all real s, so as to have

$$\begin{aligned}
&\left(\nabla\widehat{\phi}_t(t),\nabla\widehat{\phi}(t)\right)_{L^2(\Omega)}=\left\langle\nabla\widehat{\phi}_t(t),\nabla\widehat{\phi}(t)\right\rangle_{H^{-\epsilon}(\Omega)\times H^{\epsilon}(\Omega)}\\
&\leq C\left\|\widehat{\phi}_t(t)\right\|_{H^{1-\epsilon}(\Omega)}\left\|\widehat{\phi}(t)\right\|_{H^{1+\epsilon}(\Omega}\leq\ \mathsf{l.o.t.}\left(\widehat{\phi},\widehat{\phi}_t,\widehat{\psi}\right).
\end{aligned}\tag{2.48}$$

Combining (2.45), (2.47), and (2.48) with the definition of $E_{\widehat{\phi}}$ in (2.3), we then obtain

$$\begin{aligned}
&\gamma\left(\nabla\widehat{\phi}_t(t),\nabla\left(\overline{h}\cdot\nabla\widehat{\phi}(t)\right)\right)_{L^2(\Omega)}+\frac{\gamma}{2}\left(\nabla\widehat{\phi}_t(t),\nabla\widehat{\phi}(t)\right)_{L^2(\Omega)}\\
&\leq\sqrt{\frac{2\gamma}{1-\mu}}\max_{(x,y)\in\overline{\Omega}}\left|\overline{h}(x,y)\right|E_{\widehat{\phi}}(t)+\mathsf{l.o.t.}\ \left(\widehat{\phi},\widehat{\phi}_t,\widehat{\psi}\right).
\end{aligned}\tag{2.49}$$

Coupling (2.44) and (2.49) in turn, we arrive at the estimate

(2.50)

$$(2.43)\leq\sqrt{\frac{2\gamma}{1-\mu}}\left\{\max_{(x,y)\in\overline{\Omega}}\left|\overline{h}(x,y)\right|\right\}\left(E_{\widehat{\phi}}(T-\epsilon_0)+E_{\widehat{\phi}}(\epsilon_0)\right)+\mathsf{l.o.t.}\left(\widehat{\phi},\widehat{\phi}_t,\widehat{\psi}\right).$$

(A.5) Handling the term $-\int_{\epsilon_0}^{T-\epsilon_0}\alpha(\widehat{\psi},\frac{\partial}{\partial\nu}(\overline{h}\cdot\nabla\widehat{\phi}))$ and noting that

$$\begin{aligned}
\frac{\partial}{\partial\nu}\left(\overline{h}\cdot\nabla\widehat{\phi}\right)&=\nu_1\widehat{\phi}_x+\nu_1\,(x-x_0)\,\widehat{\phi}_{xx}\ +\nu_1\,(y-y_0)\,\widehat{\phi}_{xy}+\nu_2\,(x-x_0)\,\widehat{\phi}_{xy}\\
&\quad+\nu_2\widehat{\phi}_y+\nu_2\,(y-y_0)\,\widehat{\phi}_{yy},
\end{aligned}\tag{2.51}$$

we then have by Cauchy–Schwarz, the trace estimate (2.38) for the Kirchoff plates above, the use of the forcing data in (2.15), and the standard Sobolev trace theory that

(2.52)

$$
\begin{aligned}
&-\int_{\epsilon_0}^{T-\epsilon_0} \alpha\left(\widehat{\psi}, \frac{\partial}{\partial \nu}\left(\overline{h}\cdot\nabla\widehat{\phi}\right)\right)_{L^2(\Gamma_1)} dt \\
&\le C\int_{\epsilon_0}^{T-\epsilon_0}\left[\left\|\widehat{\psi}\right\|^2_{H^{\frac{1}{2}+\epsilon}(\Omega)} + \left\|\widehat{\phi}_{xx}\right\|^2_{L^2(\Gamma_1)} + \left\|\widehat{\phi}_{yy}\right\|^2_{L^2(\Gamma_1)} + 2\left\|\widehat{\phi}_{xy}\right\|^2_{L^2(\Gamma_1)}\right. \\
&\qquad\left. + \left\|\widehat{\phi}_x\right\|^2_{H^{\frac{1}{2}+\epsilon}(\Omega)} + \left\|\widehat{\phi}_y\right\|^2_{H^{\frac{1}{2}+\epsilon}(\Omega)}\right] dt \\
&= C\int_{\epsilon_0}^{T-\epsilon_0}\left[\left\|\widehat{\psi}\right\|^2_{H^{\frac{1}{2}+\epsilon}(\Omega)} + \left\|\frac{\partial^2\widehat{\phi}}{\partial\tau^2}\right\|^2_{L^2(\Gamma_1)} + \left\|\frac{\partial^2\widehat{\phi}}{\partial\nu^2}\right\|^2_{L^2(\Gamma_1)} + 2\left\|\frac{\partial^2\widehat{\phi}}{\partial\tau\partial\nu}\right\|^2_{L^2(\Gamma_1)}\right. \\
&\qquad\left. + \left\|\widehat{\phi}_x\right\|^2_{H^{\frac{1}{2}+\epsilon}(\Omega)} + \left\|\widehat{\phi}_y\right\|^2_{H^{\frac{1}{2}+\epsilon}(\Omega)}\right] dt \\
&\le C_T\left(\int_0^T\left[\left\|c_0\widehat{\psi} + c_1\widehat{\psi}_t + c_2\widehat{\phi} + c_3\widehat{\phi}_t + c_4\Delta\widehat{\phi}\right\|^2_{\left[H^{\frac{3}{2}-\epsilon}(\Omega)\right]'} + \left\|\widehat{\psi}\right\|^2_{H^{\frac{1}{2}+\epsilon}(\Omega)}\right.\right. \\
&\qquad\left. + \left\|\gamma\frac{\partial}{\partial\nu}\left(\xi^2\widehat{\phi} + 2\xi\widehat{\phi}_t\right)\right\|^2_{L^2(\Gamma_1)}\right] dt \\
&\qquad\left. + \int_0^T\left[\left\|\nabla\widehat{\phi}_t\right\|^2_{L^2(\Gamma_1)} + \left\|\widehat{\phi}\right\|^2_{H^{\frac{3}{2}+\epsilon}(\Omega)} + \left\|\widehat{\phi}_t\right\|^2_{H^{\frac{1}{2}+\epsilon}(\Omega)}\right] dt\right) \\
&\le C_T\int_0^T\left[\left\|\nabla\widehat{\phi}_t\right\|^2_{L^2(\Gamma_1)} + \left\|c_1\widehat{\psi}_t + c_4\Delta\widehat{\phi}\right\|^2_{\left[H^{\frac{3}{2}-\epsilon}(\Omega)\right]'} + \right] dt + \text{l.o.t.}\left(\widehat{\phi}, \widehat{\phi}_t, \widehat{\psi}\right).
\end{aligned}
$$

To handle the term $\int_0^T \|c_1\widehat{\psi}_t + c_4\Delta\widehat{\phi}\|^2_{[H^{\frac{3}{2}-\epsilon}(\Omega)]'} dt$, we use Proposition 4.4 in the appendix below and the fact that $\widehat{\psi}_t = -\xi\widehat{\psi}(t) + e^{-\xi t}\psi_t(t)$ and $\widehat{\phi}(t) = e^{-\xi t}\phi(t)$ to have

(2.53)

$$
\begin{aligned}
&\int_0^T \left\|c_1\widehat{\psi}_t + c_4\Delta\widehat{\phi}\right\|^2_{\left[H^{\frac{3}{2}-\epsilon}(\Omega)\right]'} dt \\
&= \int_0^T \left\|-\xi c_1\widehat{\psi}(t) + c_1 e^{-\xi t}\psi_t(t) + c_4 e^{-\xi t}\Delta\phi(t)\right\|^2_{\left[H^{\frac{3}{2}-\epsilon}(\Omega)\right]'} dt \\
&\le C\int_0^T\left[\left\|\widehat{\psi}\right\|^2_{L^2(\Omega)} + \|\phi\|^2_{H^{\frac{3}{2}+\epsilon}(\Omega)} + \|\psi\|^2_{H^{\frac{1}{2}+\epsilon}(\Omega)} + \|\phi_t\|^2_{H^{\frac{1}{2}+\epsilon}(\Omega)} + \left\|\frac{\partial\phi_t}{\partial\nu}\right\|^2_{L^2(\Gamma_1)}\right] dt \\
&\le C_T\int_0^T \left\|\nabla\widehat{\phi}_t\right\|^2_{L^2(\Gamma_1)} dt + \text{l.o.t.}\left(\widehat{\phi}, \widehat{\phi}_t, \widehat{\psi}\right).
\end{aligned}
$$

Collectively, estimates (2.52) and (2.53) then give

$$
\begin{aligned}
&-\int_{\epsilon_0}^{T-\epsilon_0} \alpha \left(\widehat{\psi}, \frac{\partial}{\partial \nu} \left(\overline{h} \cdot \nabla \widehat{\phi} \right) \right)_{L^2(\Gamma_1)} dt \\
&\leq C_T \int_0^T \left\| \nabla \widehat{\phi}_t \right\|^2_{L^2(\Gamma_1)} dt + \text{l.o.t.} \left(\widehat{\phi}, \widehat{\phi}_t, \widehat{\psi} \right). \qquad (2.54)
\end{aligned}
$$

(A.6) In the same way as in (A.5) we have

(2.55)

$$
\begin{aligned}
&-\int_{\epsilon_0}^{T-\epsilon_0} \int_{\Gamma_1} \frac{\overline{h} \cdot \nu}{2} \left[\left(\frac{\partial^2 \widehat{\phi}}{\partial x^2} \right)^2 + \left(\frac{\partial^2 \widehat{\phi}}{\partial y^2} \right)^2 + 2\mu \left(\frac{\partial^2 \widehat{\phi}}{\partial x^2} \right) \left(\frac{\partial^2 \widehat{\phi}}{\partial y^2} \right) \right. \\
&\qquad \left. + \ 2(1-\mu) \left(\frac{\partial^2 \widehat{\phi}}{\partial x \partial y} \right)^2 \right] dt d\Gamma \\
&\leq C_T \int_0^T \left\| \nabla \widehat{\phi}_t \right\|^2_{L^2(\Gamma_1)} + \text{l.o.t.} \left(\widehat{\phi}, \widehat{\phi}_t, \widehat{\psi} \right).
\end{aligned}
$$

(A.7) Finally,

(2.56)

$$
\frac{1}{2} \int_{\epsilon_0}^{T-\epsilon_0} \int_{\Gamma_1} \overline{h} \cdot \nu \left(\widehat{\phi}_t^2 + \gamma \left| \nabla \widehat{\phi}_t \right|^2 \right) d\Gamma dt \leq C \int_0^T \left\| \nabla \widehat{\phi}_t \right\|^2_{L^2(\Gamma_1)} dt + \text{l.o.t.} \ \left(\widehat{\phi}, \widehat{\phi}_t, \widehat{\psi} \right).
$$

Estimate (2.39) now comes about by stringing together (2.36), (2.40)–(2.42), (2.50), and (2.56), and taking $\epsilon > 0$ small enough. □

LEMMA 2.6. *For $T > T_0 \equiv 2\sqrt{\frac{2\gamma}{1-\mu}} \max_{(x,y)\in\overline{\Omega}} |\overline{h}(x,y)|$, the solution $[\widehat{\phi}, \widehat{\phi}_t, \widehat{\psi}]$ of* (2.15) *satisfies the following estimate:*

$$
\begin{aligned}
&\int_0^T E_{\widehat{\phi}}(t) dt + E_{\widehat{\phi}}(T) + \int_0^T \left\| A_R^{\frac{1}{2}} \widehat{\psi} \right\|^2_{L^2(\Omega)} dt \\
&\leq C_T \left(\|\psi_0\|^2_{L^2(\Omega)} + \int_0^T \left\| \nabla \widehat{\phi}_t \right\|^2_{L^2(\Gamma_1)} dt \right) + \text{l.o.t.} \left(\widehat{\phi}, \widehat{\phi}_t, \widehat{\psi} \right). \qquad (2.57)
\end{aligned}
$$

Proof. We have for any $\epsilon_0 \in (0, T)$,

$$
\begin{aligned}
&\int_0^T E_{\widehat{\phi}}(t) dt = \int_0^{\epsilon_0} E_{\widehat{\phi}}(t) dt + \int_{T-\epsilon_0}^T E_{\widehat{\phi}}(t) dt + \int_{\epsilon_0}^{T-\epsilon_0} E_{\widehat{\phi}}(t) dt \\
&\leq \frac{2\epsilon_0 (1+\epsilon)}{1-\epsilon} E_{\widehat{\phi}}(T) + \frac{2\epsilon_0 \epsilon}{1-\epsilon} \int_0^T \left[E_{\widehat{\phi}}(t) + \left\| A_R^{\frac{1}{2}} \widehat{\psi}(t) \right\|^2_{L^2(\Omega)} \right] dt + \int_{\epsilon_0}^{T-\epsilon_0} E_{\widehat{\phi}}(t) dt \\
&\quad + C_\epsilon \int_0^T \left\| \nabla \widehat{\phi}_t \right\|^2_{L^2(\Gamma_1)} dt + \text{l.o.t.}(\widehat{\phi}, \widehat{\phi}_t, \widehat{\psi}) \\
&\quad \text{(after applying Lemma 2.2(b) twice)} \\
&\leq C^* \left(E_{\widehat{\phi}}(T - \epsilon_0) + E_{\widehat{\phi}}(\epsilon_0) \right) + \frac{2\epsilon_0 (1+\epsilon)}{1-\epsilon} E_{\widehat{\phi}}(T) + \frac{2\epsilon_0 \epsilon}{1-\epsilon} \int_0^T E_{\widehat{\phi}}(t) dt \\
&(2.58) + \frac{2\epsilon (1+\epsilon_0)}{1-\epsilon} \int_0^T \left\| A_R^{\frac{1}{2}} \widehat{\psi} \right\|^2_{L^2(\Omega)} dt + C_{T,\epsilon} \int_0^T \left\| \nabla \widehat{\phi}_t \right\|^2_{L^2(\Gamma_1)} dt + \text{l.o.t.}(\widehat{\phi}, \widehat{\phi}_t, \widehat{\psi}),
\end{aligned}
$$

after applying Lemma 2.5 with $C^* \geq \sqrt{\frac{2\gamma}{1-\mu}} \max_{[x,y]\in\overline{\Omega}} \left|\overline{h}(x,y)\right|$, and $\widetilde{\epsilon} \equiv \frac{2\epsilon}{1-\epsilon}$ therein. Applying Lemma 2.2(b) twice more to the right-hand side of (2.58) yields now

(2.59)
$$\begin{aligned}\int_0^T E_{\widehat{\phi}}(t)dt \leq\ & 2\left(\epsilon_0 + C^*\right)\frac{1+\epsilon}{1-\epsilon}E_{\widehat{\phi}}(T) + \frac{2\epsilon\left(\epsilon_0 + C^*\right)}{1-\epsilon}\int_0^T E_{\widehat{\phi}}(t)dt \\ &+ \frac{2\epsilon\left(1+\epsilon_0 + C^*\right)}{1-\epsilon}\int_0^T \left\|A_R^{\frac{1}{2}}\widehat{\psi}\right\|^2_{L^2(\Omega)} dt + C_{T,\epsilon}\int_0^T \left\|\nabla\widehat{\phi}_t\right\|^2_{L^2(\Gamma_1)} dt \\ &+ \mathsf{l.o.t.}(\widehat{\phi},\widehat{\phi}_t,\widehat{\psi}).\end{aligned}$$

Moreover, we have by (2.16)

$$\int_0^T E_{\widehat{\phi}}(t)dt = TE_{\widehat{\phi}}(T) + \int_0^T \mathcal{F}(T,t)dt, \tag{2.60}$$

where the function $\mathcal{F}$ is as defined in (2.34). Combining (2.59) and (2.60) yields

(2.61)
$$\begin{aligned}TE_{\widehat{\phi}}(T) + \int_0^T \mathcal{F}(T,t)dt \leq\ & 2\left(\epsilon_0 + C^*\right)\frac{1+\epsilon}{1-\epsilon}E_{\widehat{\phi}}(T) + \frac{2\epsilon\left(\epsilon_0 + C^*\right)}{1-\epsilon}\int_0^T E_{\widehat{\phi}}(t)dt \\ &+ \frac{2\epsilon\left(1+\epsilon_0 + C^*\right)}{1-\epsilon}\int_0^T \left\|A_R^{\frac{1}{2}}\widehat{\psi}\right\|^2_{L^2(\Omega)} dt + C_{T,\epsilon}\int_0^T \left\|\nabla\widehat{\phi}_t\right\|^2_{L^2(\Gamma_1)} dt + \mathsf{l.o.t.}(\widehat{\phi},\widehat{\phi}_t,\widehat{\psi}).\end{aligned}$$

To use this inequality, we integrate both sides of (2.17) (with $s = T$ therein) so as to have

$$\begin{aligned}\int_0^T \mathcal{F}(T,t)dt \leq\ & \epsilon\left(T+1\right)\int_0^T E_{\widehat{\phi}}(t)dt + \epsilon T E_{\widehat{\phi}}(T) + \epsilon T\int_0^T \left\|A_R^{\frac{1}{2}}\widehat{\psi}\right\|^2_{L^2(\Omega)} dt \\ &+ C_{T,\epsilon}\int_0^T \left\|\nabla\widehat{\phi}_t\right\|^2_{L^2(\Gamma_1)} dt + \mathsf{l.o.t.}(\widehat{\phi},\widehat{\phi}_t,\widehat{\psi}).\end{aligned} \tag{2.62}$$

Combining (2.61) and (2.62), we thus obtain

(2.63)
$$\begin{aligned}TE_{\widehat{\phi}}(T) \leq\ & \left[\frac{2\left(\epsilon_0 + C^*\right)\left(1+\epsilon\right)}{1-\epsilon} + \epsilon T\right]E_{\widehat{\phi}}(T) \\ &+ \epsilon\left[\frac{2\left(\epsilon_0 + C^*\right)}{1-\epsilon} + \left(T+1\right)\right]\int_0^T E_{\widehat{\phi}}(t)dt \\ &+ \epsilon\left[\frac{2\left(1+\epsilon_0 + C^*\right)}{1-\epsilon} + T\right]\int_0^T \left\|A_R^{\frac{1}{2}}\widehat{\psi}\right\|^2_{L^2(\Omega)} dt + C_{T,\epsilon}\int_0^T \left\|\nabla\widehat{\phi}_t\right\|^2_{L^2(\Gamma_1)} dt \\ &+ \mathsf{l.o.t.}(\widehat{\phi},\widehat{\phi}_t,\widehat{\psi}).\end{aligned}$$

Taking now $T > \frac{2(\epsilon_0+C^*)(1+\epsilon)}{(1-\epsilon)^2}$, or what is the same, $T > 2C^*$ for ϵ and ϵ_0 small enough, we then have

$$\begin{aligned}E_{\widehat{\phi}}(T) \leq\ & C_{T,\epsilon}\int_0^T \left\|\nabla\widehat{\phi}_t\right\|^2_{L^2(\Gamma_1)} dt + \epsilon\widetilde{C}_T\left[\int_0^T E_{\widehat{\phi}}(t)dt + \int_0^T \left\|A_R^{\frac{1}{2}}\widehat{\psi}\right\|^2_{L^2(\Omega)} dt\right] \\ &+ \mathsf{l.o.t.}(\widehat{\phi},\widehat{\phi}_t,\widehat{\psi}),\end{aligned} \tag{2.64}$$

where throughout $\widetilde{C}_T$ will denote a constant independent of ϵ and ϵ_0 (small enough).

In turn, applying this to (2.59), we have

$$\int_0^T E_{\widehat{\phi}}(t)dt \leq C_{T,\epsilon}\int_0^T \left\|\nabla\widehat{\phi}_t\right\|^2_{L^2(\Gamma_1)}dt + \frac{2\epsilon\left(\epsilon_0+C^*\right)\left[\left(1+\epsilon\right)\widetilde{C}_T+1\right]}{1-\epsilon}\int_0^T E_{\widehat{\phi}}(t)dt$$
$$+\frac{2\epsilon\left[\left(\epsilon_0+C^*\right)\left(1+\epsilon\right)\widetilde{C}_T+\left(1+\epsilon_0+C^*\right)\right]}{1-\epsilon}\int_0^T \left\|A_R^{\frac{1}{2}}\widehat{\psi}\right\|^2_{L^2(\Omega)}dt + \mathsf{l.o.t.}(\widehat{\phi},\widehat{\phi}_t,\widehat{\psi})$$

from which follows the estimate, for $\epsilon,\epsilon_0>0$ small enough,

(2.65)
$$\int_0^T E_{\widehat{\phi}}(t)dt \leq C_{T,\epsilon}\int_0^T \left\|\nabla\widehat{\phi}_t\right\|^2_{L^2(\Gamma_1)}dt + \epsilon\widetilde{C}_T\int_0^T \left\|A_R^{\frac{1}{2}}\widehat{\psi}\right\|^2_{L^2(\Omega)}dt + \mathsf{l.o.t.}(\widehat{\phi},\widehat{\phi}_t,\widehat{\psi}).$$

Coupling together (2.64) and (2.65), we have the following preliminary inequality for the mechanical energy, again for $T>2C^*$:

$$\int_0^T E_{\widehat{\phi}}(t)dt + E_{\widehat{\phi}}(T)$$
$$\leq C_{T,\epsilon}\int_0^T \left\|\nabla\widehat{\phi}_t\right\|^2_{L^2(\Gamma_1)}dt + \epsilon\widetilde{C}_T\int_0^T \left\|A_R^{\frac{1}{2}}\widehat{\psi}\right\|^2_{L^2(\Omega)}dt + \mathsf{l.o.t.}(\widehat{\phi},\widehat{\phi}_t,\widehat{\psi}). \tag{2.66}$$

It remains to estimate the thermal component. To this end, we can multiply (2.11) by $\widehat{\psi}$, integrate in time and space, use the characterization (1.15) and (2.8) to have

$$\eta\int_0^T \left\|A_R^{\frac{1}{2}}\widehat{\psi}\right\|^2_{L^2(\Omega)}dt = \left[\frac{\beta}{2}\left\|e^{-\xi t}\psi(t)\right\|^2_{L^2(\Omega)}\right]_{t=0}^{t=T} + \xi\int_0^T\left(\beta\widehat{\psi}-\alpha\Delta\widehat{\phi},\widehat{\psi}\right)_{L^2(\Omega)}dt$$
$$+\alpha\int_0^T\left[\left(\nabla\widehat{\phi}_t,\nabla\widehat{\psi}\right)_{L^2(\Omega)} - \left(\frac{\partial\widehat{\phi}_t}{\partial\nu},\widehat{\psi}\right)_{L^2(\Gamma_1)}\right]dt. \tag{2.67}$$

Majorizing this expression results in

$$\eta\int_0^T \left\|A_R^{\frac{1}{2}}\widehat{\psi}\right\|^2_{L^2(\Omega)}dt \leq C_{\underset{\epsilon}{\sim}}\left(\|\psi_0\|^2_{L^2(\Omega)} + \int_0^T\left\|\nabla\widehat{\phi}_t\right\|^2_{L^2(\Gamma_1)}dt + \int_0^T E_{\widehat{\phi}}(t)dt\right)$$
$$+\widetilde{\epsilon}\int_0^T \left\|A_R^{\frac{1}{2}}\widehat{\psi}\right\|^2_{L^2(\Omega)}dt + \mathsf{l.o.t.}(\widehat{\phi},\widehat{\phi}_t,\widehat{\psi}),$$

and taking $\widetilde{\epsilon}>0$ small enough above, this becomes

$$\int_0^T \left\|A_R^{\frac{1}{2}}\widehat{\psi}\right\|^2_{L^2(\Omega)}dt \leq C_1\left(\|\psi_0\|^2_{L^2(\Omega)} + \int_0^T\left\|\nabla\widehat{\phi}_t\right\|^2_{L^2(\Gamma_1)}dt + \int_0^T E_{\widehat{\phi}}(t)dt\right)$$
$$+\mathsf{l.o.t.}(\widehat{\phi},\widehat{\phi}_t,\widehat{\psi}), \tag{2.68}$$

where $C_1=\frac{C_{\widetilde{\epsilon}}}{\eta-\widetilde{\epsilon}}$.

Combining (2.66) and (2.68), we have

(2.69)

$$\int_0^T E_{\widehat{\phi}}(t)dt + E_{\widehat{\phi}}(T)$$
$$\leq C_{T,\epsilon}\int_0^T \left\|\nabla\widehat{\phi}_t\right\|^2_{L^2(\Gamma_1)} dt + \epsilon\widetilde{C}_T C_1\left(\|\psi_0\|^2_{L^2(\Omega)} + \int_0^T E_{\widehat{\phi}}(t)dt\right) + \text{l.o.t.}(\widehat{\phi},\widehat{\phi}_t,\widehat{\psi}),$$

from which we obtain for $\epsilon > 0$ small enough

(2.70)
$$\int_0^T E_{\widehat{\phi}}(t)dt + E_{\widehat{\phi}}(T) \leq C_{T,\epsilon}\left(\|\psi_0\|^2_{L^2(\Omega)} + \int_0^T \left\|\nabla\widehat{\phi}_t\right\|^2_{L^2(\Gamma_1)} dt\right) + \text{l.o.t.}(\widehat{\phi},\widehat{\phi}_t,\widehat{\psi}).$$

The final estimate (2.57) finally comes about by combining (2.70) and (2.68). □

Conclusion of the proof of Theorem 2.1. Assume initially that $[\phi_0,\phi_1,\psi_0] \in D(\mathcal{A}^*_\gamma)$. Through the change of variable $\widehat{\phi}(t) = e^{-\xi t}\phi(t)$ and $\widehat{\psi}(t) = e^{-\xi t}\psi(t)$, where again $[\widehat{\phi},\widehat{\phi}_t,\widehat{\psi}]$ solves (2.15) and $\xi \equiv \frac{\alpha^2}{2\gamma\eta} > 0$, we have for $T > T_0 \equiv 2\sqrt{\frac{2\gamma}{1-\mu}}\max_{[x,y]\in\overline{\Omega}}\left|\overline{h}(x,y)\right|$

(2.71)
$$\int_0^T E_\phi(t)dt + \int_0^T \left\|A_R^{\frac{1}{2}}\psi(t)\right\|^2_{L^2(\Omega)} dt + E_\phi(T)$$
$$= \int_0^T \left[\left\|\mathring{\mathbf{A}}^{\frac{1}{2}} e^{\xi t}\widehat{\phi}(t)\right\|^2_{L^2(\Omega)} + \left\|P_\gamma^{\frac{1}{2}}\left(e^{\xi t}\widehat{\phi}_t(t) + \xi e^{\xi t}\widehat{\phi}(t)\right)\right\|^2_{L^2(\Omega)} + \left\|A_R^{\frac{1}{2}} e^{\xi t}\widehat{\psi}(t)\right\|^2_{L^2(\Omega)}\right] dt$$
$$+ \left\|\mathring{\mathbf{A}}^{\frac{1}{2}} e^{\xi T}\widehat{\phi}(T)\right\|^2_{L^2(\Omega)} + \left\|P_\gamma^{\frac{1}{2}}\left(e^{\xi T}\widehat{\phi}_t(T) + \xi e^{T}\widehat{\phi}(T)\right)\right\|^2_{L^2(\Omega)}$$
$$\leq C_T\left(\|\psi_0\|^2_{L^2(\Omega)} + \int_0^T \left\|\nabla\widehat{\phi}_t\right\|^2_{L^2(\Gamma_1)} dt\right) + \text{l.o.t.}\left(\widehat{\phi},\widehat{\phi}_t,\widehat{\psi}\right)$$
(after using estimate (2.57))
$$\leq C_T\left(\|\psi_0\|^2_{L^2(\Omega)} + \int_0^T \left\|\nabla\left(e^{-\xi t}\phi_t(t) - \xi e^{-\xi t}\phi(t)\right)\right\|^2_{L^2(\Gamma_1)} dt\right)$$
$$+ \text{l.o.t.}\left(e^{-\xi t}\phi, e^{-\xi t}\phi_t - \xi e^{-\xi t}\phi, e^{-\xi t}\psi\right)$$
$$\leq C_T\left(\|\psi_0\|^2_{L^2(\Omega)} + \int_0^T \|\nabla\phi_t(t)\|^2_{L^2(\Gamma_1)}\, dt\right) + \text{l.o.t.}\left(\phi,\phi_t,\psi\right).$$

This gives the desired inequality (2.5). □

2.3. Conclusion of the proof of Theorem 1.2. For $[\phi_0,\phi_1] \in D(\mathcal{L}^*_T\Pi^*)$, we immediately have from Theorem 2.1 the following corollary.

COROLLARY 2.7. *For $[\phi_0,\phi_1] \in D(\mathcal{L}^*_T\Pi^*)$ and $T > T_0 \equiv 2\sqrt{\frac{2\gamma}{1-\mu}}\max_{[x,y]\in\overline{\Omega}}\left|\overline{h}(x,y)\right|$, the corresponding solution $[\phi,\phi_t,\psi]$ of (1.60) satisfies the following inequality:*

(2.72)
$$\int_0^T E_\phi(t)dt + E_\phi(T) + \int_0^T \left\|A_R^{\frac{1}{2}}\psi\right\|^2_{L^2(\Omega)} dt \leq C_T\int_0^T \|\nabla\phi_t\|^2_{L^2(\Gamma_1)}\, dt + \text{l.o.t.}\left(\phi,\phi_t,\psi\right).$$

We will have the desired inequality (2.2) upon the elimination of the tainting lower order terms in (2.72). To this end, we invoke a (by now) classical compactness–uniqueness argument (see, e.g., [13] and [2]), which makes crucial use of the new Holmgren-type uniqueness result for the thermoelastic system recently derived by Isakov in [10]. It is at this point that the boundary trace $\psi|_{\Gamma_2}$, corresponding to the control u_3, comes into play.

LEMMA 2.8. *Let T^* be as defined in* (1.6). *Then for $T > T^*$ and initial data $[\phi_0, \phi_1] \in D(\mathcal{L}_T^* \Pi^*)$, there exists a C_T such that the following estimate holds true for the solution of* (1.60):

$$\begin{aligned} &\|\phi\|^2_{L^\infty(0,T;H^{\frac{3}{2}+\epsilon}(\Omega))} + \|\phi_t\|^2_{L^\infty(0,T;H^{\frac{1}{2}+\epsilon}(\Omega))} + \|\psi\|^2_{L^\infty(0,T;H^{-\frac{1}{2}+\epsilon}(\Omega))} + \int_0^T \|\psi\|^2_{H^{\frac{1}{2}+\epsilon}(\Omega)}\, dt \\ &\qquad \leq C_T \left(\int_0^T \|\nabla \phi_t\|^2_{L^2(\Gamma_1)}\, dt + \|\psi\|^2_{[H^s((0,T)\times\Gamma_2)]'} \right). \end{aligned} \tag{2.73}$$

Proof. If the proposition is false, then there exists a sequence $\{[\phi_0^{(n)}, \phi_1^{(n)}]\}_{n=1}^\infty \subseteq D(\mathcal{L}_T^* \Pi^*)$, and a corresponding solution sequence $\{[\phi^{(n)}, \phi_t^{(n)}, \psi^{(n)}]\}_{n=1}^\infty$ to (1.60), which satisfies

$$\begin{aligned} &\left\|\phi^{(n)}\right\|^2_{L^\infty\left(0,T;H^{\frac{3}{2}+\epsilon}(\Omega)\right)} + \left\|\phi_t^{(n)}\right\|^2_{L^\infty\left(0,T;H^{\frac{1}{2}+\epsilon}(\Omega)\right)} + \left\|\psi^{(n)}\right\|^2_{L^\infty\left(0,T;H^{-\frac{1}{2}+\epsilon}(\Omega)\right)} \\ &\qquad + \int_0^T \left\|\psi^{(n)}\right\|^2_{H^{\frac{1}{2}+\epsilon}(\Omega)} dt = 1 \ \ \forall\, n, \end{aligned} \tag{2.74}$$

$$\lim_{n\to\infty} \int_0^T \left\|\nabla \phi_t^{(n)}\right\|^2_{L^2(\Gamma_1)} dt + \left\|\psi^{(n)}\right\|^2_{[H^s((0,T)\times\Gamma_2)]'} = 0. \tag{2.75}$$

As $T > 2\sqrt{\frac{2\gamma}{1-\mu}} \max_{[x,y]\in\overline{\Omega}} \left|\overline{h}(x,y)\right|$, we have the existence of the inequality (2.72). This and (2.74)–(2.75) then imply the boundedness of the sequence

$$\begin{aligned} &\left\{ \int_0^T \left[\left\| \begin{bmatrix} \phi^{(n)}(t) \\ \phi_t^{(n)}(t) \end{bmatrix} \right\|^2_{D\left(\mathring{\mathbf{A}}^{\frac{1}{2}}\right)\times H^1_{\Gamma_0,\gamma}(\Omega)} + \left\| A_R^{\frac{1}{2}} \psi^{(n)}(t) \right\|^2_{L^2(\Omega)} \right] dt \right. \\ &\qquad \left. + \left\| \begin{bmatrix} \phi_0^{(n)} \\ \phi_1^{(n)} \end{bmatrix} \right\|^2_{D\left(\mathring{\mathbf{A}}^{\frac{1}{2}}\right)\times H^1_{\Gamma_0,\gamma}(\Omega)} \right\}_{n=1}^\infty. \end{aligned} \tag{2.76}$$

There thus exists a subsequence, still denoted here as $\{[\phi_0^{(n)}, \phi_1^{(n)}]\}_{n=1}^\infty$, and $[\widetilde{\phi}_0, \widetilde{\phi}_1] \in D(\mathring{\mathbf{A}}^{\frac{1}{2}}) \times H^1_{\Gamma_0,\gamma}(\Omega)$, such that

$$\phi_0^{(n)} \to \widetilde{\phi}_0 \ \text{ in } D(\mathring{\mathbf{A}}^{\frac{1}{2}}) \text{ weakly}, \tag{2.77}$$

$$\phi_1^{(n)} \to \widetilde{\phi}_1 \ \text{ in } H^1_{\Gamma_0,\gamma}(\Omega) \text{ weakly}. \tag{2.78}$$

If we further denote $[\widetilde{\phi}, \widetilde{\phi}_t, \widetilde{\psi}]$ as the solution to (1.60), corresponding to initial data $[\widetilde{\phi}_0, \widetilde{\phi}_1, 0]$, then a fortiori,

$$\left[\phi^{(n)}, \phi_t^{(n)}, \psi^{(n)}\right] \to \left[\widetilde{\phi}, \widetilde{\phi}_t, \widetilde{\psi}\right] \ \text{ in } L^\infty(0,T;\mathbf{H}_\gamma) \text{ weak star}. \tag{2.79}$$

From Proposition 4.3 of the appendix, we have that $\{\phi_{tt}^{(n)}\}_{n=1}^{\infty}$ is bounded in $L^{\infty}(0,T;[D(\mathring{\mathbf{A}}^{\frac{1}{2}}P_{\gamma}^{-1})]')$, inasmuch as $\{\|[\phi_0^{(n)},\phi_1^{(n)}]\|_{D(\mathring{\mathbf{A}}^{\frac{1}{2}})\times H^1_{\Gamma_0,\gamma}(\Omega)}\}_{n=1}^{\infty}$ is bounded in $D(\mathring{\mathbf{A}}^{\frac{1}{2}})\times H^1_{\Gamma_0,\gamma}(\Omega)$. Also, from Proposition 4.4 we have that $\psi_t^{(n)}\in L^2(0,T;[H^{\frac{3}{2}-\epsilon}(\Omega)]')$ for all n, with the estimate

$$\int_0^T \left\|\psi_t^{(n)}\right\|^2_{[H^{\frac{3}{2}-\epsilon}(\Omega)]'} dt \le C\int_0^T \left\|\nabla\phi_t^{(n)}\right\|^2_{L^2(\Gamma_1)} dt + \text{l.o.t.}\,(\phi,\phi_t,\psi)\,, \tag{2.80}$$

and this combined with (2.74)–(2.75) yields that $\{\psi_t^{(n)}\}_{n=1}^{\infty}$ is bounded in $L^2(0,T;[H^{\frac{3}{2}-\epsilon}(\Omega)]')$. This boundedness of $\{[\phi_{tt}^{(n)},\psi_t^{(n)}]\}_{n=1}^{\infty}$, and that for the sequence posted in (2.76), allows us to deduce through a compactness result of Simon's in [24] that

$$\begin{aligned}
&\phi^{(n)}\to\widetilde{\phi} \ \text{ strongly in } L^{\infty}(0,T;H^{\frac{3}{2}+\epsilon}(\Omega)),\\
&\phi_t^{(n)}\to\widetilde{\phi}_t \ \text{ strongly in } L^{\infty}(0,T;H^{\frac{1}{2}+\epsilon}(\Omega)),\\
&\psi^{(n)}\to\widetilde{\psi} \ \text{ strongly in } L^{2}(0,T;H^{\frac{1}{2}+\epsilon}(\Omega)),\\
&\psi^{(n)}\to\widetilde{\psi} \ \text{ strongly in } L^{\infty}(0,T;H^{-\frac{1}{2}+\epsilon}(\Omega)).
\end{aligned}$$

These convergences and (2.74) thus give

$$\begin{aligned}
&\left\|\widetilde{\phi}\right\|^2_{L^{\infty}\left(0,T;H^{\frac{3}{2}+\epsilon}(\Omega)\right)} + \left\|\widetilde{\phi}_t\right\|^2_{L^{\infty}\left(0,T;H^{\frac{1}{2}+\epsilon}(\Omega)\right)} + \left\|\widetilde{\psi}\right\|^2_{L^{\infty}\left(0,T;H^{-\frac{1}{2}+\epsilon}(\Omega)\right)}\\
&\qquad + \int_0^T \left\|\widetilde{\psi}\right\|^2_{H^{\frac{1}{2}+\epsilon}(\Omega)} dt = 1.
\end{aligned} \tag{2.81}$$

Moreover, the explicit representation of $\mathcal{L}_T^*\Pi^*$ in (1.59) and the convergences posted in (2.75) and (2.77)–(2.79) give that $[\widetilde{\phi}_0,\widetilde{\phi}_1]\in D(\mathcal{L}_T^*\Pi^*)$, with

$$\int_0^T \left\|\nabla\widetilde{\phi}_t\right\|^2_{H^1(\Gamma_1)} dt + \left\|\widetilde{\psi}\right\|^2_{[H^s((0,T)\times\Gamma_2)]'} = 0. \tag{2.82}$$

Now if we make the change of variable

$$z=\widetilde{\phi}_t, \qquad v=\widetilde{\psi}_t,$$

then using (2.82), $[z,v]$ solve the system

$$\begin{cases}
\begin{cases} z_{tt}-\gamma\Delta z_{tt}+\Delta^2 z+\alpha\Delta v=0\\ \beta v_t+\eta\Delta v-\sigma v-\alpha\Delta z_t=0\end{cases} & \text{on } (0,\infty)\times\Omega,\\[2ex]
z=\dfrac{\partial z}{\partial\nu}=0 \ \text{ on } (0,\infty)\times\Gamma,\\[2ex]
\begin{cases}\Delta z+(1-\mu)B_1 z+\alpha v=0\\ \dfrac{\partial\Delta z}{\partial\nu}+(1-\mu)\dfrac{\partial B_2 z}{\partial\tau}-\gamma\dfrac{\partial z_{tt}}{\partial\nu}+\alpha\dfrac{\partial v}{\partial\nu}=0\end{cases} & \text{on } (0,\infty)\times\Gamma_1,\\[2ex]
\dfrac{\partial v}{\partial\nu}+\lambda v=0 \ \text{ on } (0,\infty)\times\Gamma,\\[2ex]
v=0 \ \text{ on } (0,\infty)\times\Gamma_2.
\end{cases} \tag{2.83}$$

Now by Isakov's theorem in [10, p. 3, Corollary 1.2], we have for

$$T > 2\sqrt{\gamma} \cdot \sup_{[x,y]\in\Omega} d\left([x,y],\Gamma_2\right)$$

that the uniqueness property for the thermoelastic system is obtained, so that the solution $[z, v]$ of (2.83) is necessarily zero. Consequently $\widetilde{\phi}$ and $\widetilde{\psi}$ are each constants. From the essential boundary condition on Γ_0 in (1.60), we then have $\widetilde{\phi} = 0$ on $(0,T)\times\Omega$. In turn, the free boundary conditions on Γ_1 give that $\widetilde{\psi} = 0$ on $(0,T)\times\Omega$. Thus $[\widetilde{\phi}, \widetilde{\psi}] = [0,0]$, which contradicts the equality given in (2.81). This concludes the proof of the lemma. □

Corollary 2.7 and Lemma 2.8 in combination give inequality (2.2), the establishment of which verifies the surjectivity of the control to partial state map $\Pi\mathcal{L}_T : D(\mathcal{L}_T) \subset \mathcal{U}_s \to D(\mathring{\mathbf{A}}^{\frac{1}{2}}) \times H^1_{\Gamma_0,\gamma}(\Omega)$. This completes the proof of Theorem 1.2.

3. The proof of Theorem 1.1. Given the space $C^r(\Sigma_{2,T})$, we consider system (1.1) under the influence of boundary controls in $\mathcal{U}_{r+1}$, as defined in (1.37). The controlled PDE is then approximately controllable in $\mathcal{U}_{r+1}$ for $T > 2\sqrt{\gamma}\cdot\sup_{[x,y]\in\Omega} d\left([x,y],\Gamma_2\right)$. Indeed, if we take arbitrary $[\phi_0,\phi_1,\psi_0]$ from the null space of $\mathcal{L}_T^*$, then using the form of this operator given in (1.55), we have necessarily that $\phi_t|_{\Gamma_1} = \frac{\partial\phi_t}{\partial\nu}|_{\Gamma_1} = 0$, and $\psi|_{\Gamma_2} = 0$, where $[\phi,\phi_t,\psi]$ is the solution to (1.51). We can then use the uniqueness theorem of Isakov, in a fashion similar to that employed in Lemma 2.8, to show that $[\phi,\phi_t,\psi] = [0,0,0]$ on $(0,T)\times\Omega$ and, in particular, $[\phi_0,\phi_1,\psi_0] = [0,0,0]$.

A preliminary step (a regularity property of $\mathcal{L}_T$). With the designated control space $\mathcal{U}_{r+1}$ we then take $T > T^*$ so as to ensure both the approximate controllability of the entire system (1.1) and the exact controllability with respect to the displacement (see Theorem 1.2). In this event, we have the observability inequality (2.2), and therewith one can show in a manner identical to that done in [14, Appendix B] that the operator

$$\Pi\mathcal{L}_T\mathcal{L}_T^*\Pi^* \text{ is an isomorphism from } D(\mathcal{L}_T^*\Pi^*) \text{ into } [D(\mathcal{L}_T^*\Pi^*)]', \tag{3.1}$$

where the projection Π onto $D(\mathring{\mathbf{A}}^{\frac{1}{2}}) \times H^1_{\Gamma_0,\gamma}(\Omega)$ is as defined in (1.46). Consequently, we have

$$\Pi\mathcal{L}_T\mathcal{L}_T^*\Pi^*\left(\Pi\mathcal{L}_T\mathcal{L}_T^*\Pi^*\right)^{-1}\Pi \in \mathcal{L}\left(\mathbf{H}_\gamma, D(\mathring{\mathbf{A}}^{\frac{1}{2}}) \times H^1_{\Gamma_0,\gamma}(\Omega)\right). \tag{3.2}$$

Moreover, if we denote the maps $\mathcal{L}^{(1)}$, $\mathcal{L}^{(2)}$ by

$$\mathcal{L}^{(1)}\begin{bmatrix} u_1 \\ u_2 \end{bmatrix}(t) = \int_0^t e^{\mathcal{A}_\gamma(t-s)}\mathcal{B}\begin{bmatrix} u_1(s) \\ u_2(s) \\ 0 \end{bmatrix},$$

$$\mathcal{L}^{(2)}u(t) = \int_0^t e^{\mathcal{A}_\gamma(t-s)}\mathcal{B}\begin{bmatrix} 0 \\ 0 \\ u(s) \end{bmatrix}$$

(cf. (1.42)), then by a standard energy method one can show that

$$\mathcal{L}^{(2)} : L^2(0,T;H^{-\frac{1}{2}}(\Gamma_2)) \to C([0,T];\mathbf{H}_\gamma) \text{ continuously.} \tag{3.3}$$

To handle $\mathcal{L}^{(1)}$ on the other hand, one must appeal to a new regularity result in [17], which gives

$$\mathcal{L}^{(1)} : L^2(0,T;L^2(\Gamma_1)\times H^{-1}(\Gamma_1)) \to C([0,T];H^{\frac{3}{2}}(\Omega)\times H^{\frac{1}{2}}(\Omega)\times L^2(\Omega)) \quad \text{continuously.} \tag{3.4}$$

Combining (3.3) and (3.4) (at terminal time T) with (3.2), we thus deduce that the mapping

$$(\mathbb{I}-\Pi^*\Pi)\,\mathcal{L}_T\mathcal{L}_T^*\Pi^*\left(\Pi\mathcal{L}_T\mathcal{L}_T^*\Pi^*\right)^{-1}\Pi \in \mathcal{L}\left(\mathbf{H}_\gamma\right), \tag{3.5}$$

where $\mathbb{I}:\mathbf{H}_\gamma \to \mathbf{H}_\gamma$ denotes the identity.

Combining (3.2) and (3.5) thus gives

$$\mathcal{L}_T\mathcal{L}_T^*\Pi^*\left(\Pi\mathcal{L}_T\mathcal{L}_T^*\Pi^*\right)^{-1}\Pi \in \mathcal{L}(\mathbf{H}_\gamma). \tag{3.6}$$

Step 1. For arbitrary $\epsilon>0$ we select a $\overline{u}_1 \in D(\mathcal{L}_T) \subset \mathcal{U}_{r+1}$, so that for arbitrary terminal state $[\omega_0^T,\omega_1^T,\theta_0^T] \in \mathbf{H}_\gamma$, the corresponding solution $[\omega^{(1)}(t),\omega_t^{(1)}(t),\theta^{(1)}(t)]$ to (1.1), with $[u_1,u_2,u_3]\equiv\overline{u}_1$ and zero initial data, satisfies

$$\begin{aligned}&\left\|\begin{bmatrix}\omega^{(1)}(T)-\omega_0^T\\ \omega_t^{(1)}(T)-\omega_1^T\\ \theta^{(1)}(T)-\theta_0^T\end{bmatrix}+e^{\mathcal{A}_\gamma T}\begin{bmatrix}\omega_0\\ \omega_1\\ \theta_0\end{bmatrix}\right\|_{\mathbf{H}_\gamma}\\ &\qquad<\frac{\epsilon}{1+\left\|(\mathbb{I}-\Pi^*\Pi)\mathcal{L}_T\mathcal{L}_T^*\Pi^*\left(\Pi\mathcal{L}_T\mathcal{L}_T^*\Pi^*\right)^{-1}\Pi\right\|_{\mathcal{L}(\mathbf{H}_\gamma)}}\end{aligned} \tag{3.7}$$

(where the fact that $(\mathbb{I}\text{-}\Pi^*\Pi)\mathcal{L}_T\mathcal{L}_T^*\Pi^*\left(\Pi\mathcal{L}_T\mathcal{L}_T^*\Pi^*\right)^{-1}\Pi$ is due to (3.6)).

Step 2. We now select $\overline{u}_2 \in D(\mathcal{L}_T)$ to be the "minimal norm steering control" with respect to the (partial) terminal state $[\omega_0^T-\omega^{(1)}(T),\omega_1^T-\omega_t^{(1)}(T)]$. That is to say, $\overline{u}_2$ satisfies

$$\Pi\mathcal{L}_T\overline{u}_2+\Pi e^{\mathcal{A}_\gamma T}\begin{bmatrix}\omega_0\\ \omega_1\\ \theta_0\end{bmatrix}=\begin{bmatrix}\omega_0^T-\omega^{(1)}(T)\\ \omega_1^T-\omega_t^{(1)}(T)\end{bmatrix} \tag{3.8}$$

and minimizes the functional $\frac{1}{2}\|\overline{u}\|_{\mathcal{U}_s}^2$, over all $\overline{u}\in\mathcal{U}_s$, which satisfies

$$\Pi\mathcal{L}_T\overline{u}=\begin{bmatrix}\omega_0^T-\omega^{(1)}(T)\\ \omega_1^T-\omega_t^{(1)}(T)\end{bmatrix}-\Pi e^{\mathcal{A}_\gamma T}\begin{bmatrix}\omega_0\\ \omega_1\\ \theta_0\end{bmatrix}.$$

(By Theorem 1.2. we know there exists at least one such $\overline{u}$.) By convex optimization theory and Lax–Milgram, the minimizer $\overline{u}_2$ can be given explicitly by

$$\overline{u}_2=\mathcal{L}_T^*\Pi^*\left(\Pi\mathcal{L}_T\mathcal{L}_T^*\Pi^*\right)^{-1}\Pi\left(\begin{bmatrix}\omega_0^T-\omega^{(1)}(T)\\ \omega_1^T-\omega_t^{(1)}(T)\\ \theta_0^T-\theta^{(1)}(T)\end{bmatrix}-e^{\mathcal{A}_\gamma T}\begin{bmatrix}\omega_0\\ \omega_1\\ \theta_0\end{bmatrix}\right) \tag{3.9}$$

(see (B.20) of [14, p. 288]). With this representation, we then have from (3.7) the norm bound

$$\|(\mathbb{I}-\Pi^*\Pi)\mathcal{L}_T\overline{u}_2\|_{\mathbf{H}_\gamma}\le\frac{\left\|(\mathbb{I}-\Pi^*\Pi)\mathcal{L}_T\mathcal{L}_T^*\Pi^*\left(\Pi\mathcal{L}_T\mathcal{L}_T^*\Pi^*\right)^{-1}\Pi\right\|_{\mathcal{L}(\mathbf{H}_\gamma)}\cdot\epsilon}{1+\left\|(\mathbb{I}-\Pi^*\Pi)\mathcal{L}_T\mathcal{L}_T^*\Pi^*\left(\Pi\mathcal{L}_T\mathcal{L}_T^*\Pi^*\right)^{-1}\Pi\right\|_{\mathcal{L}(\mathbf{H}_\gamma)}}. \tag{3.10}$$

Step 3. Set the control $\overline{u}^* = \overline{u}_1 + \overline{u}_2$. Consequently, there is the equality

$$\begin{aligned}\mathcal{L}_T\overline{u}^* + e^{\mathcal{A}_\gamma T}\begin{bmatrix}\omega_0\\ \omega_1\\ \theta_0\end{bmatrix} &= \mathcal{L}_T\overline{u}_1 + \mathcal{L}_T\overline{u}_2 + e^{\mathcal{A}_\gamma T}\begin{bmatrix}\omega_0\\ \omega_1\\ \theta_0\end{bmatrix} = \begin{bmatrix}\omega_0^T\\ \omega_1^T\\ \theta^{(1)}(T)\end{bmatrix}\\ &\quad + (\mathbb{I} - \Pi^*\Pi)\left[\mathcal{L}_T\overline{u}_2 + e^{\mathcal{A}_\gamma T}\begin{bmatrix}\omega_0\\ \omega_1\\ \theta_0\end{bmatrix}\right]. \end{aligned} \tag{3.11}$$

Letting $[\omega^*, \omega_t^*, \theta^*]$ denote the solution of (1.1) corresponding to the chosen control $\overline{u}^*$, we then have from (3.11) that $[\omega^*(T), \omega_t^*(T)] = [\omega_0^T, \omega_1^T]$. Moreover, from (3.11), (3.7), and (3.10) we obtain the estimate

$$\begin{aligned}\left\|\theta^*(T) - \theta_0^T\right\|_{L^2_{\sigma+\lambda}(\Omega)} &\le \left\|\begin{bmatrix}0\\ 0\\ \theta^{(1)}(T) - \theta_0^T\end{bmatrix} + (\mathbb{I} - \Pi^*\Pi)\left[e^{\mathcal{A}_\gamma T}\begin{bmatrix}\omega_0\\ \omega_1\\ \theta_0\end{bmatrix}\right]\right\|_{\mathbf{H}_\gamma}\\ &\quad + \|(\mathbb{I} - \Pi^*\Pi)\mathcal{L}_T\overline{u}_2\|_{\mathbf{H}_\gamma}\\ &< \frac{\epsilon}{1 + \left\|(\mathbb{I} - \Pi^*\Pi)\mathcal{L}_T\mathcal{L}_T^*\Pi^*\left(\Pi\mathcal{L}_T\mathcal{L}_T^*\Pi^*\right)^{-1}\Pi\right\|_{\mathcal{L}(\mathbf{H}_\gamma)}} + \|(\mathbb{I} - \Pi^*\Pi)\mathcal{L}_T\overline{u}_2\|_{\mathbf{H}_\gamma} < \epsilon. \end{aligned} \tag{3.12}$$

Thus, the constructed control $\overline{u}^* = [u_1^*, u_2^*, u_3^*] \in \mathcal{U}_{r+1}$ satisfies the desired exact–approximate controllability property. Moreover, the Sobolev embedding theorem gives that $u_3^* \in C^r(\Sigma_{2,T})$. This concludes the proof of Theorem 1.1.

4. Appendix.

PROPOSITION 4.1. *The operator* $A_R - \frac{\sigma}{\eta} + \lambda\mathring{\mathbf{A}}G_2\gamma_0 - \mathring{\mathbf{A}}G_1\gamma_0$ *is an element of* $\mathcal{L}(L^2(\Omega), [D(\mathring{\mathbf{A}}^{\frac{1}{2}})]')$ *and* $(A_R - \frac{\sigma}{\eta} + \lambda\mathring{\mathbf{A}}G_2\gamma_0 - \mathring{\mathbf{A}}G_1\gamma_0)^* = A_D(I - D\gamma_0)$ *as elements of* $\mathcal{L}(D(\mathring{\mathbf{A}}^{\frac{1}{2}}), L^2(\Omega))$.

Proof. For every $\vartheta \in D(A_R)$ and $\varpi \in D(\mathring{\mathbf{A}}^{\frac{1}{2}})$, we have

$$\begin{aligned}&\left\langle\left(A_R - \frac{\sigma}{\eta} + \lambda\mathring{\mathbf{A}}G_2\gamma_0 - \mathring{\mathbf{A}}G_1\gamma_0\right)\vartheta, \varpi\right\rangle_{\left[D\left(\mathring{\mathbf{A}}^{\frac{1}{2}}\right)\right]'\times D|\left(\mathring{\mathbf{A}}^{\frac{1}{2}}\right)}\\ &= (-\Delta\vartheta, \varpi)_{L^2(\Omega)} + \left\langle\lambda\mathring{\mathbf{A}}G_2\gamma_0\vartheta, \varpi\right\rangle_{\left[D\left(\mathring{\mathbf{A}}^{\frac{1}{2}}\right)\right]'\times D\left(\mathring{\mathbf{A}}^{\frac{1}{2}}\right)} - \left\langle\mathring{\mathbf{A}}G_1\gamma_0\vartheta, \varpi\right\rangle_{\left[D\left(\mathring{\mathbf{A}}^{\frac{1}{2}}\right)\right]'\times D\left(\mathring{\mathbf{A}}^{\frac{1}{2}}\right)}\\ &= (\nabla\vartheta, \nabla\varpi)_{L^2(\Omega)} - \left(\frac{\partial\vartheta}{\partial\nu}, \varpi\right)_{L^2(\Gamma_1)} + \lambda\left(\gamma_0\vartheta, G_2^*\mathring{\mathbf{A}}\varpi\right)_{L^2(\Gamma_1)} - \left(\gamma_0\vartheta, G_1^*\mathring{\mathbf{A}}\varpi\right)_{L^2(\Gamma_1)}\\ &\quad \text{(after the use of Green's formula and the taking of adjoints)}\\ &= (\nabla\vartheta, \nabla\varpi)_{L^2(\Omega)} - \left(\gamma_0\vartheta, G_1^*\mathring{\mathbf{A}}\varpi\right)_{L^2(\Gamma_1)} = (\vartheta, -\Delta\varpi)_{L^2(\Omega)}\\ &\quad \text{(after one more use of Green's theorem and the characterization (1.23))}\\ &= (\vartheta, A_D(I - D\gamma_0)\varpi)_{L^2(\Omega)}. \end{aligned}$$

As $D(A_R)$ is dense in $L^2(\Omega)$, this equality proves the assertion. ∎

LEMMA 4.2. *The Hilbert space adjoint $\mathcal{A}_\gamma^*$ of $\mathcal{A}_\gamma$, as defined in* (1.36)*, is given to be*

$$\mathcal{A}_\gamma^* = \begin{pmatrix} \mathbf{I} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & P_\gamma^{-1} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{I} \end{pmatrix} \begin{pmatrix} \mathbf{0} & -\mathbf{I} & \mathbf{0} \\ \mathring{\mathbf{A}} & \mathbf{0} & -\alpha(\clubsuit) \\ \mathbf{0} & \dfrac{\alpha}{\beta} A_D(\mathbf{I} - D\gamma_0) & -\dfrac{\eta}{\beta} A_R \end{pmatrix},$$

$$\text{with } D(\mathcal{A}_\gamma^*) = \Big\{ [\phi_0, \phi_1, \psi_0] \in D\left(\mathring{\mathbf{A}}^{\frac{1}{2}}\right) \times D\left(\mathring{\mathbf{A}}^{\frac{1}{2}}\right) \times D(A_R) \\ \text{such that } \mathring{\mathbf{A}}\phi_0 + \alpha \mathring{\mathbf{A}} G_1 \gamma_0 \psi_0 \in H^{-1}_{\Gamma_0,\gamma}(\Omega) \Big\}$$

(above, $(\clubsuit)$ is the same denotation made in (1.35)*).*

Proof. We define $\mathcal{S} \subseteq \mathbf{H}_\gamma$ to be

$$\mathcal{S} \equiv \Big\{ [\phi_0, \phi_1, \psi_0] \in D\left(\mathring{\mathbf{A}}^{\frac{1}{2}}\right) \times D\left(\mathring{\mathbf{A}}^{\frac{1}{2}}\right) \times D(A_R) \\ \text{such that } \mathring{\mathbf{A}}\omega_1 + \alpha\ \mathring{\mathbf{A}} G_1 \gamma_0 \theta \in H^{-1}_{\Gamma_0,\gamma}(\Omega) \Big\}$$

and proceed to show that $D(\mathcal{A}_\gamma^*) = \mathcal{S}$. Indeed, if $[\omega_1, \omega_2, \theta] \in D(\mathcal{A}_\gamma)$ and $[\widetilde{\omega}_1, \widetilde{\omega}_2, \widetilde{\theta}] \in \mathcal{S}$, we have by using (1.36)

$$\begin{aligned}
&\left(\mathcal{A}_\gamma \begin{bmatrix} \omega_1 \\ \omega_2 \\ \theta \end{bmatrix}, \begin{bmatrix} \widetilde{\omega}_1 \\ \widetilde{\omega}_2 \\ \widetilde{\theta} \end{bmatrix} \right)_{\mathbf{H}_\gamma} \\
&= \left(\mathring{\mathbf{A}}^{\frac{1}{2}} \omega_2, \mathring{\mathbf{A}}^{\frac{1}{2}} \widetilde{\omega}_1 \right)_{L^2(\Omega)} \\
&\quad + \left(P_\gamma^{-1} \left(-\mathring{\mathbf{A}}\omega_1 + \alpha A_R \theta - \frac{\alpha\sigma}{\eta}\theta - \alpha \mathring{\mathbf{A}} G_1 \gamma_0 \theta + \alpha\lambda \mathring{\mathbf{A}} G_2 \gamma_0 \theta \right), \widetilde{\omega}_2 \right)_{H^1_{\Gamma_0,\gamma}(\Omega)} \\
&\quad - \left(\alpha A_D (I - D\gamma_0)\omega_2, \widetilde{\theta} \right)_{L^2(\Omega)} - \frac{\eta\beta}{\beta} \left(A_R \theta, \widetilde{\theta} \right)_{L^2(\Omega)} \\
&= \left\langle \omega_2, \mathring{\mathbf{A}} \widetilde{\omega}_1 \right\rangle_{\left[D\left(\mathring{\mathbf{A}}^{\frac{1}{2}}\right) \right]' \times D\left(\mathring{\mathbf{A}}^{\frac{1}{2}}\right)} - \left\langle \omega_1,\ \mathring{\mathbf{A}} \widetilde{\omega}_2 \right\rangle_{\left[D\left(\mathring{\mathbf{A}}^{\frac{1}{2}}\right) \right]' \times D\left(\mathring{\mathbf{A}}^{\frac{1}{2}}\right)} - \alpha \left(\Delta\theta, \widetilde{\omega}_2 \right)_{L^2(\Omega)} \\
&\quad - \alpha \left\langle \mathring{\mathbf{A}} G_1 \gamma_0 \theta, \widetilde{\omega}_2 \right\rangle_{\left[D\left(\mathring{\mathbf{A}}^{\frac{1}{2}}\right) \right]' \times D\left(\mathring{\mathbf{A}}^{\frac{1}{2}}\right)} + \alpha\lambda \left\langle \mathring{\mathbf{A}} G_2 \gamma_0 \theta, \widetilde{\omega}_2 \right\rangle_{\left[D\left(\mathring{\mathbf{A}}^{\frac{1}{2}}\right) \right]' \times D\left(\mathring{\mathbf{A}}^{\frac{1}{2}}\right)} \\
&\quad + \alpha \left(\Delta\omega_2, \widetilde{\theta} \right)_{L^2(\Omega)} - \beta \left(\theta, \frac{\eta}{\beta} A_R \widetilde{\theta} \right)_{L^2(\Omega)} \\
&\quad \text{(after using the equality posted in (1.32))} \\
&= \left\langle \omega_2, \mathring{\mathbf{A}} \widetilde{\omega}_1 \right\rangle_{\left[D\left(\mathring{\mathbf{A}}^{\frac{1}{2}}\right) \right]' \times D\left(\mathring{\mathbf{A}}^{\frac{1}{2}}\right)} - \left\langle \omega_1,\ \mathring{\mathbf{A}} \widetilde{\omega}_2 \right\rangle_{\left[D\left(\mathring{\mathbf{A}}^{\frac{1}{2}}\right) \right]' \times D\left(\mathring{\mathbf{A}}^{\frac{1}{2}}\right)} + \alpha \left(\nabla\theta, \nabla\widetilde{\omega}_2 \right)_{L^2(\Omega)} \\
&\quad - \alpha \left(\frac{\partial\theta}{\partial\nu}, \gamma_0 \widetilde{\omega}_2 \right)_{L^2(\Gamma_1)} - \alpha \left(\gamma_0\theta, \frac{\partial\widetilde{\omega}_2}{\partial\nu} \right)_{L^2(\Gamma_1)} - \alpha\lambda \left(\gamma_0\theta, \gamma_0\widetilde{\omega}_2 \right)_{L^2(\Gamma_1)} \\
&\quad - \alpha \left(\nabla\omega_2, \nabla\widetilde{\theta} \right)_{L^2(\Omega)} + \alpha \left(\frac{\partial\omega_2}{\partial\nu}, \gamma_0\widetilde{\theta} \right)_{L^2(\Gamma_1)} - \beta \left(\theta, \frac{\eta}{\beta} A_R \widetilde{\theta} \right)_{L^2(\Gamma_1)} \\
&\quad \text{(after using Green's theorem and (1.23))} \\
&= - \left(\mathring{\mathbf{A}}^{\frac{1}{2}} \omega_1, \mathring{\mathbf{A}}^{\frac{1}{2}} \widetilde{\omega}_2 \right)_{L^2(\Omega)} + \left\langle \omega_2, \mathring{\mathbf{A}} \widetilde{\omega}_1 \right\rangle_{\left[D\left(\mathring{\mathbf{A}}^{\frac{1}{2}}\right) \right]' \times D\left(\mathring{\mathbf{A}}^{\frac{1}{2}}\right)} + \beta \left(\theta, -\frac{\alpha}{\beta} \Delta\widetilde{\omega}_2 \right)_{L^2(\Omega)}
\end{aligned}$$

$$
\begin{aligned}
&+\alpha\left(\omega_2,\Delta\widetilde{\theta}\right)_{L^2(\Omega)}-\alpha\left(\gamma_0\omega_2,\frac{\partial\widetilde{\theta}}{\partial\nu}\right)_{L^2(\Gamma_1)}+\alpha\left(G_1^*\mathring{\mathbf{A}}\omega_2,\gamma_0\widetilde{\theta}\right)_{L^2(\Gamma_1)}\\
&-\beta\left(\theta,\frac{\eta}{\beta}A_R\widetilde{\theta}\right)_{L^2(\Gamma_1)}\\
=&-\left(\mathring{\mathbf{A}}^{\frac{1}{2}}\omega_1,\mathring{\mathbf{A}}^{\frac{1}{2}}\widetilde{\omega}_2\right)_{L^2(\Omega)}+\left\langle\omega_2,\mathring{\mathbf{A}}\widetilde{\omega}_1\right\rangle_{\left[D\left(\mathring{\mathbf{A}}^{\frac{1}{2}}\right)\right]'\times D\left(\mathring{\mathbf{A}}^{\frac{1}{2}}\right)}\\
&+\alpha\left(\omega_2,-A_R\widetilde{\theta}+\frac{\sigma}{\eta}\widetilde{\theta}\right)_{L^2(\Omega)}-\alpha\lambda\left(G_2^*\mathring{\mathbf{A}}\omega_2,\gamma_0\widetilde{\theta}\right)_{L^2(\Gamma_1)}\\
&+\alpha\left\langle\omega_2,\mathring{\mathbf{A}}G_1\gamma_0\widetilde{\theta}\right\rangle_{\left[D\left(\mathring{\mathbf{A}}^{\frac{1}{2}}\right)\right]'\times D\left(\mathring{\mathbf{A}}^{\frac{1}{2}}\right)}+\beta\left(\theta,\frac{\alpha}{\beta}A_D(I-D\gamma_0)\widetilde{\omega}_2\right)_{L^2(\Omega)}\\
&-\beta\left(\theta,\frac{\eta}{\beta}A_R\widetilde{\theta}\right)_{L^2(\Gamma_1)}\\
=&-\left(\mathring{\mathbf{A}}^{\frac{1}{2}}\omega_1,\mathring{\mathbf{A}}^{\frac{1}{2}}\widetilde{\omega}_2\right)_{L^2(\Omega)}\\
&+\left(P_\gamma^{\frac{1}{2}}\omega_2,P_\gamma^{\frac{1}{2}}P_\gamma^{-1}\left[\mathring{\mathbf{A}}\widetilde{\omega}_1+\alpha\left(-A_R\widetilde{\theta}+\frac{\sigma}{\eta}\widetilde{\theta}+\mathring{\mathbf{A}}G_1\gamma_0\widetilde{\theta}-\lambda\mathring{\mathbf{A}}G_2\gamma_0\widetilde{\theta}\right)\right]\right)_{L^2(\Omega)}\\
&+\beta\left(\theta,\frac{\alpha}{\beta}A_D(I-D\gamma_0)\widetilde{\omega}_2\right)_{L^2(\Omega)}-\beta\left(\theta,\frac{\eta}{\beta}A_R\widetilde{\theta}\right)_{L^2(\Gamma_1)}\\
&\text{(after again using (1.32), (1.23), and the fact that } \left[\widetilde{\omega}_1,\widetilde{\omega}_2,\widetilde{\theta}\right]\in\mathcal{S})\\
=&\left(\begin{bmatrix}\omega_1\\ \omega_2\\ \theta\end{bmatrix},\mathcal{T}\begin{bmatrix}\widetilde{\omega}_1\\ \widetilde{\omega}_2\\ \widetilde{\theta}\end{bmatrix}\right)_{\mathbf{H}_\gamma},
\end{aligned}
$$

where

$$
\mathcal{T}\equiv\begin{pmatrix}\mathbf{I}&\mathbf{0}&\mathbf{0}\\ \mathbf{0}&P_\gamma^{-1}&\mathbf{0}\\ \mathbf{0}&\mathbf{0}&\mathbf{I}\end{pmatrix}\begin{pmatrix}\mathbf{0}&-\mathbf{I}&\mathbf{0}\\ \mathring{\mathbf{A}}&\mathbf{0}&-\alpha(\clubsuit)\\ \mathbf{0}&\frac{\alpha}{\beta}A_D(\mathbf{I}-D\gamma_0)&-\frac{\eta}{\beta}A_R\end{pmatrix}.
$$

Thus,

$$
\mathcal{S}\subseteq D(\mathcal{A}_\gamma^*)\text{ and }\mathcal{A}_\gamma^*\big|_{\mathcal{S}}=\mathcal{T}. \tag{4.1}
$$

To show the opposite containment, one can straightforwardly compute the inverse $\mathcal{A}_\gamma^{-1}\in\mathcal{L}(\mathbf{H}_\gamma,D(\mathcal{A}_\gamma))$ as

$$
\mathcal{A}_\gamma^{-1}=\begin{pmatrix}-\frac{\alpha^2}{\eta}\mathring{\mathbf{A}}^{-1}(\clubsuit)A_R^{-1}A_D(I-D\gamma_0)&-\mathring{\mathbf{A}}^{-1}P_\gamma&-\frac{\alpha\beta}{\eta}\mathring{\mathbf{A}}^{-1}(\clubsuit)A_R^{-1}\\ \mathbf{I}&\mathbf{0}&\mathbf{0}\\ -\frac{\alpha}{\eta}A_R^{-1}A_D(I-D\gamma_0)&\mathbf{0}&-\frac{\beta}{\eta}A_R^{-1}\end{pmatrix}. \tag{4.2}
$$

In turn, one can use this quantity and Proposition 4.1 to compute the Hilbert space

adjoint $\left(\mathcal{A}_\gamma^*\right)^{-1}$ of $\mathcal{A}_\gamma^{-1}$ as

$$(4.3)\quad \left(\mathcal{A}_\gamma^*\right)^{-1} = \begin{pmatrix} -\frac{\alpha^2}{\eta}\mathring{\mathbf{A}}^{-1}(\clubsuit)A_R^{-1}A_D(I-D\gamma_0) & \mathring{\mathbf{A}}^{-1}P_\gamma & -\frac{\alpha\beta}{\eta}\mathring{\mathbf{A}}^{-1}(\clubsuit)A_R^{-1} \\ -\mathbf{I} & \mathbf{0} & \mathbf{0} \\ -\frac{\alpha}{\eta}A_R^{-1}A_D(I-D\gamma_0) & \mathbf{0} & -\frac{\beta}{\eta}A_R^{-1} \end{pmatrix}.$$

With this quantity in hand, we then have that for arbitrary $[\phi_0,\phi_1,\psi_0]\in D(\mathcal{A}_\gamma^*)$ and corresponding

$$\begin{bmatrix} \omega_0 \\ \omega_1 \\ \theta_0 \end{bmatrix} = \mathcal{A}_\gamma^* \begin{bmatrix} \phi_0 \\ \phi_1 \\ \psi_0 \end{bmatrix} \in \mathbf{H}_\gamma,$$

$$\begin{aligned} \begin{bmatrix} \phi_0 \\ \phi_1 \\ \psi_0 \end{bmatrix} &= \left(\mathcal{A}_\gamma^*\right)^{-1} \begin{bmatrix} \omega_0 \\ \omega_1 \\ \theta_0 \end{bmatrix} \\ &= \begin{bmatrix} -\frac{\alpha^2}{\eta}\mathring{\mathbf{A}}^{-1}(\clubsuit)A_R^{-1}A_D(I-D\gamma_0)\omega_0 + \mathring{\mathbf{A}}^{-1}P_\gamma\omega_1 - \frac{\alpha\beta}{\eta}\mathring{\mathbf{A}}^{-1}(\clubsuit)A_R^{-1}\theta_0 \\ -\omega_0 \\ -\frac{\alpha}{\eta}A_R^{-1}A_D(I-D\gamma_0)\omega_0 - \frac{\beta}{\eta}A_R^{-1}\theta_0 \end{bmatrix}. \end{aligned}$$

(4.4)

A fortiori then, $[\phi_0,\phi_1]\in[D(\mathring{\mathbf{A}}^{\frac12})]^2$ and $\psi_0\in D(A_R)$. Moreover, (4.4) and the definition of the operator $(\clubsuit)$ in (1.35) gives

$$(4.5)\qquad \mathring{\mathbf{A}}\phi_0 + \alpha\mathring{\mathbf{A}}G_1\gamma_0\phi_1 \in H^1_{\Gamma_0,\gamma}(\Omega).$$

Thus, $D(\mathcal{A}_\gamma^*)\subseteq\mathcal{S}$, and this combined with (4.1) concludes Lemma 4.2. □

PROPOSITION 4.3. *For arbitrary terminal data $[\phi_0,\phi_1,\psi_0]\in\mathbf{H}_\gamma$, the solution $[\phi,\phi_t,\psi]$ to (1.51) has the following additional regularity:*

$$\|\phi_{tt}\|_{L^\infty(0,T;[D(\mathring{\mathbf{A}}^{\frac12}P_\gamma^{-1})]')} \le C\,\|[\phi_0,\phi_1,\psi_0]\|_{\mathbf{H}_\gamma},$$

where $\mathring{\mathbf{A}}^{\frac12}P_\gamma^{-1}$ is taken as a closed and densely defined operator, $\mathring{\mathbf{A}}^{\frac12}P_\gamma^{-1}: D(\mathring{\mathbf{A}}^{\frac12}P_\gamma^{-1})\subset L^2(\Omega)\to L^2(\Omega)$, with $D(\mathring{\mathbf{A}}^{\frac12}P_\gamma^{-1}) = \{\varphi\in L^2(\Omega): P_\gamma^{-1}\varphi\in D(\mathring{\mathbf{A}}^{\frac12})\}$.

Proof. For terminal data $[\phi_0,\phi_1,\psi_0]\in D(\mathcal{A}_\gamma^*)$, we have for all $\varpi\in L^1(0,T;D(\mathring{\mathbf{A}}^{\frac12}P_\gamma^{-1}))$, upon using the abstract equation (1.52), the characterizations in (1.23) , the fact that $P_\gamma^{-1}\varpi\in L^1(0,T;D(\mathring{\mathbf{A}}^{\frac12})\cap D(A_N))$ (recall the definition of A_N in (1.16) and P_γ in (1.24)), and $\frac{\partial\psi}{\partial\nu}|_\Gamma = -\lambda\psi|_\Gamma$, that

(4.6)

$$\begin{aligned} &\int_0^T (\varpi,\phi_{tt})_{L^2(\Omega)}\,dt = \int_0^T \left(\varpi, P_\gamma^{-1}\left[-\mathring{\mathbf{A}}\phi - \alpha\mathring{\mathbf{A}}G_1\gamma_0\psi + \alpha\lambda\ \mathring{\mathbf{A}}G_2\gamma_0\psi - \alpha\Delta\psi\right]\right)_{L^2(\Omega)}dt \\ &= \int_0^T \left[-\left(\mathring{\mathbf{A}}^{\frac12}P_\gamma^{-1}\varpi, \mathring{\mathbf{A}}^{\frac12}\phi\right)_{L^2(\Omega)} - \alpha\left(G_1^*\mathring{\mathbf{A}}P_\gamma^{-1}\varpi, \gamma_0\psi\right)_{L^2(\Gamma_1)}\right. \end{aligned}$$

$$
\begin{aligned}
&+ \alpha\lambda \left(G_2^* \, \mathring{\mathbf{A}} P_\gamma^{-1}\varpi, \gamma_0\psi\right)_{L^2(\Gamma_1)} - \alpha\left(P_\gamma^{-1}\varpi, \Delta\psi\right)_{L^2(\Omega)}\Big] \, dt \\
&= \int_0^T \left[\left(-\mathring{\mathbf{A}}^{\frac{1}{2}} P_\gamma^{-1}\varpi, \mathring{\mathbf{A}}^{\frac{1}{2}}\phi\right)_{L^2(\Omega)} + \alpha\left(P_\gamma^{-1}\varpi, \frac{\partial\psi}{\partial\nu}\right)_{L^2(\Gamma_1)} - \alpha\left(P_\gamma^{-1}\varpi, \Delta\psi\right)_{L^2(\Omega)}\right] dt \\
&= \int_0^T \left[\left(-\mathring{\mathbf{A}}^{\frac{1}{2}} P_\gamma^{-1}\varpi, \mathring{\mathbf{A}}^{\frac{1}{2}}\phi\right)_{L^2(\Omega)} + \alpha\left(\nabla P_\gamma^{-1}\varpi, \nabla\psi\right)_{L^2(\Omega)}\right] dt \\
&= \int_0^T \left[\left(-\mathring{\mathbf{A}}^{\frac{1}{2}} P_\gamma^{-1}\varpi, \mathring{\mathbf{A}}^{\frac{1}{2}}\phi\right)_{L^2(\Omega)} - \alpha\left(\Delta P_\gamma^{-1}\varpi, \psi\right)_{L^2(\Omega)}\right] dt.
\end{aligned}
$$

Estimating the far side of this expression by using the fact that $P_\gamma^{-1} \in \mathcal{L}(L^2(\Omega), D(A_N))$, followed by the contraction of the semigroup $\{e^{\mathcal{A}_\gamma^* t}\}_{t\geq 0}$, one has the estimate

$$
\int_0^T (\varpi, \phi_{tt})_{L^2(\Omega)} \, dt \leq C \left\| \begin{bmatrix} \phi_0 \\ \phi_1 \\ \psi_0 \end{bmatrix} \right\|_{\mathbf{H}_\gamma} \|\varpi\|_{L^1(0,T;D(\mathring{\mathbf{A}}^{\frac{1}{2}} P_\gamma^{-1}))} \, dt. \tag{4.7}
$$

A density argument concludes the proof. □

PROPOSITION 4.4. *If $[\phi, \phi_t, \psi]$ denotes the solution to* (1.51)*, corresponding to terminal data $[\phi_0, \phi_1, \psi_0]$, we have the following estimates.*

1. *The map $[\phi_0, \phi_1, \psi_0] \to \Delta\phi$ is an element of $\mathcal{L}(\mathbf{H}_\gamma, L^2(0,T;[H^1(\Omega)]'))$, with the norm bound*

$$
\|\Delta\phi\|_{L^2(0,T;[H^1(\Omega)]')} \leq \mathsf{l.o.t.}\,(\phi, \phi_t, \psi)\,. \tag{4.8}
$$

2. *The map $[\phi_0, \phi_1, \psi_0] \to [\Delta\phi_t, \psi_t]$ is an element of $\mathcal{L}(D(\mathcal{L}_T^*), [L^2(0,T;[H^{\frac{3}{2}-\epsilon}(\Omega)]')]^2)$, with the norm bound*

$$
\|[\Delta\phi_t, \psi_t]\|_{\left[L^2\left(0,T;\left[H^{\frac{3}{2}-\epsilon}(\Omega)\right]'\right)\right]^2} \leq C \|\nabla\phi_t\|_{L^2(0,T;L^2(\Gamma_1))} + \mathsf{l.o.t.}\,(\phi, \phi_t, \psi)\,. \tag{4.9}
$$

Proof of (i). For all $\varpi \in L^2(0,T;H^1(\Omega))$, we easily have

$$
\begin{aligned}
&\int_0^T (\Delta\phi, \varpi)_{L^2(\Omega)} \, dt = -\int_0^T (\nabla\phi, \nabla\varpi)_{L^2(\Omega)} \, dt + \int_0^T \left(\frac{\partial\phi}{\partial\nu}, \varpi\right)_{L^2(\Gamma_1)} dt \\
&\leq C \int_0^T \left[\|\nabla\phi\|_{L^2(\Omega)} \|\nabla\varpi\|_{L^2(\Omega)} + \|\phi\|_{H^{\frac{3}{2}+\epsilon}(\Omega)} \|\varpi\|_{H^1(\Omega)}\right] dt \\
&\leq C \|\phi\|_{L^2\left(0,T;H^{\frac{3}{2}+\epsilon}(\Omega)\right)} \|\varpi\|_{L^2(0,T;H^1(\Omega))}\,,
\end{aligned} \tag{4.10}
$$

and this estimate gives the asserted result.

Proof of (ii). If $[\phi_0, \phi_1, \psi_0] \in D(\mathcal{A}_\gamma^*)$, then $[\phi, \phi_t, \psi] \in C([0,T]; D(\mathcal{A}_\gamma^*)) \cap C^1([0,T]; \mathbf{H}_\gamma)$, and so in particular $\Delta\phi_t \in L^2(0,T;L^2(\Omega))$. Taking the L^2-inner product with respect to arbitrary $\varpi \in L^2(0,T;H^{\frac{3}{2}-\epsilon}(\Omega))$, we have upon the use of Green's theorem and the definition of A_R in (1.14) that

(4.11)

$$
\begin{aligned}
&-\int_0^T (\Delta\phi_t, \varpi)_{L^2(\Omega)}\, dt \\
&= \int_0^T \left(A_R^{\frac{1}{2}}\phi_t, A_R^{\frac{1}{2}}\varpi\right)_{L^2(\Omega)} dt \\
&\quad - \int_0^T \left[\lambda\,(\phi_t, \varpi)_{L^2(\Gamma)} + \frac{\sigma}{\eta}\,(\phi_t, \varpi)_{L^2(\Omega)} + \left(\frac{\partial\phi_t}{\partial\nu}, \varpi\right)_{L^2(\Gamma_1)}\right] dt \\
&= \int_0^T \left(A_R^{\frac{1}{4}+\frac{\epsilon}{2}}\phi_t, A_R^{\frac{3}{4}-\frac{\epsilon}{2}}\varpi\right)_{L^2(\Omega)} dt \\
&\quad - \int_0^T \left[\lambda\,(\phi_t, \varpi)_{L^2(\Gamma)} + \frac{\sigma}{\eta}\,(\phi_t, \varpi)_{L^2(\Omega)} + \left(\frac{\partial\phi_t}{\partial\nu}, \varpi\right)_{L^2(\Gamma_1)}\right] dt \\
&\le \int_0^T \left(\left\|A_R^{\frac{1}{4}+\frac{\epsilon}{2}}\phi_t\right\|_{L^2(\Omega)} \left\|A_R^{\frac{3}{4}-\frac{\epsilon}{2}}\varpi\right\|_{L^2(\Omega)}\right. \\
&\qquad \left. + \left(C\,\|\nabla\phi_t\|_{L^2(0,T;L^2(\Gamma_1))} + \text{l.o.t.}\,(\phi, \phi_t, \psi)\right)\|\varpi\|_{H^{\frac{1}{2}+\epsilon}(\Omega)}\right) dt \\
&\le \left(C\,\|\nabla\phi_t\|_{L^2(0,T;L^2(\Gamma_1))} + \text{l.o.t.}\,(\phi, \phi_t, \psi)\right)\|\varpi\|_{L^2\left(0,T;H^{\frac{3}{2}-\epsilon}(\Omega)\right)}.
\end{aligned}
$$

Moreover, as $[\phi_0, \phi_1, \psi_0] \in D(\mathcal{A}_\gamma^*)$, we can take the L^2-inner product of ψ_t with arbitrary $\varpi \in L^2(0,T;H^{\frac{3}{2}-\epsilon}(\Omega))$ and use (1.53) and (4.11) to obtain

$$
\begin{aligned}
&\int_0^T (\psi_t, \varpi)_{L^2(\Omega)}\, dt = \beta^{-1}\int_0^T (\eta A_R\psi + \alpha\Delta\phi_t, \varpi)_{L^2(\Omega)}\, dt \\
&= \beta^{-1}\int_0^T \left[\left(\eta A_R^{\frac{1}{4}+\frac{\epsilon}{2}}\psi, A_R^{\frac{3}{4}-\frac{\epsilon}{2}}\varpi\right)_{L^2(\Omega)} + (\alpha\Delta\phi_t, \varpi)_{L^2(\Omega)}\right] dt \\
&\le \left(C\,\|\nabla\phi_t\|_{L^2(0,T;L^2(\Gamma_1))} + \text{l.o.t.}\,(\phi, \phi_t, \psi)\right)\|\varpi\|_{L^2\left(0,T;H^{\frac{3}{2}-\epsilon}(\Omega)\right)}. \qquad (4.12)
\end{aligned}
$$

Having obtained estimates (4.11) and (4.12) with smooth data $[\phi_0, \phi_1, \psi_0]$, a density argument (see Remark 1.8) and a recollection of the form of the adjoint $\mathcal{L}_T^*$ in (1.55) will allow us to obtain the norm bound (4.9) for all terminal data in $D(\mathcal{L}_T^*)$. □

LEMMA 4.5. *Concerning the component $\widehat{\phi}$ of the solution $[\widehat{\phi}, \widehat{\phi}_t, \widehat{\psi}]$ of (2.9), one has that $\Delta\widehat{\phi}|_{\Gamma_0} \in L^2(0,T;L^2(\Gamma_0))$ with the following estimate valid for all s and $\tau \in [0,T]$:*

$$
\int_s^\tau \left\|\Delta\widehat{\phi}\right\|^2_{L^2(\Gamma_0)} dt \le C_0\left(\int_0^T E_{\widehat{\phi}}(t)dt + E_{\widehat{\phi}}(s) + E_{\widehat{\phi}}(\tau)\right) + \text{l.o.t.}\left(\widehat{\phi}, \widehat{\phi}_t, \widehat{\psi}\right). \qquad (4.13)
$$

Proof. So as to obtain the inequality (4.13), we multiply the first equation of (2.15) by the quantity $\overline{m}\cdot\nabla\widehat{\phi}$, where $\overline{m}(x,y) \equiv [m_1(x,y), m_2(x,y)]$ is a $\left[C^2(\overline{\Omega})\right]^2$ vector field,[1] which satisfies

$$
\overline{m}|_\Gamma = \begin{cases} [\nu_1, \nu_2] & \text{on } \Gamma_0, \\ 0 & \text{on } \Gamma_1, \end{cases} \qquad (4.14)
$$

[1] Here we make use of the fact that Γ_0 and Γ_1 are separated.

and follow this by an integration from s to τ; i.e., we will work with the equation

$$\begin{aligned}&\int_s^\tau \left(\widehat{\phi}_{tt} - \gamma\Delta\widehat{\phi}_{tt} + \Delta^2\widehat{\phi}, \overline{m}\cdot\nabla\widehat{\phi}\right)_{L^2(\Omega)} dt \\ &\qquad = \int_s^\tau \left(c_0\widehat{\psi} + c_1\widehat{\psi}_t + c_2\widehat{\phi} + c_3\widehat{\phi}_t + c_4\Delta\widehat{\phi}, \overline{m}\cdot\nabla\widehat{\phi}\right)_{L^2(\Omega)} dt. \end{aligned} \tag{4.15}$$

To handle the left-hand side of (4.15), perform the following steps.

(i) First,

(4.16)

$$\begin{aligned}&\int_s^\tau \left(\widehat{\phi}_{tt}, \overline{m}\cdot\nabla\widehat{\phi}\right)_{L^2(\Omega)} dt = \left(\widehat{\phi}_t, \overline{m}\cdot\nabla\widehat{\phi}\right)_{L^2(\Omega)}\Big|_s^\tau - \int_s^\tau \left(\widehat{\phi}_t, \overline{m}\cdot\nabla\widehat{\phi}_t\right)_{L^2(\Omega)} dt \\ &= \left(\widehat{\phi}_t, \overline{m}\cdot\nabla\widehat{\phi}\right)_{L^2(\Omega)}\Big|_s^\tau - \frac{1}{2}\int_s^\tau\int_\Omega \operatorname{div}\left(\widehat{\phi}_t^2\overline{m}\right) dtd\Omega \\ &\quad + \frac{1}{2}\int_s^\tau\int_\Omega \widehat{\phi}_t^2\left[m_{1x} + m_{2y}\right] dtd\Omega \\ &= \left(\widehat{\phi}_t, \overline{m}\cdot\nabla\widehat{\phi}\right)_{L^2(\Omega)}\Big|_s^\tau + \frac{1}{2}\int_s^\tau\int_\Omega \widehat{\phi}_t^2\left[m_{1x} + m_{2y}\right] dtd\Omega, \end{aligned}$$

after making use of the divergence theorem and the fact that $\widehat{\phi}_t = 0$ on Γ_0.

(ii) Next,

(4.17)

$$\begin{aligned}&\int_s^\tau \left(-\Delta\widehat{\phi}_{tt}, \overline{m}\cdot\nabla\widehat{\phi}\right)_{L^2(\Omega)} dt = \left(\nabla\widehat{\phi}_t, \nabla\left(\overline{m}\cdot\nabla\widehat{\phi}\right)\right)_{L^2(\Omega)}\Big|_s^\tau \\ &\quad - \int_s^\tau \left(\nabla\widehat{\phi}_t, \nabla\left(\overline{m}\cdot\nabla\widehat{\phi}_t\right)\right)_{L^2(\Omega)} dt \\ &= \left(\nabla\widehat{\phi}_t, \nabla\left(\overline{m}\cdot\nabla\widehat{\phi}\right)\right)_{L^2(\Omega)}\Big|_s^\tau - \frac{1}{2}\int_s^\tau\int_\Omega \operatorname{div}\left(\left|\nabla\widehat{\phi}_t\right|^2\overline{m}\right) dtd\Omega \\ &\quad - \int_s^\tau\int_\Omega \left[\frac{\widehat{\phi}_{tx}^2 m_{1x}}{2} + \frac{\widehat{\phi}_{ty}^2 m_{2y}}{2}\right] dtd\Omega - \int_s^\tau\int_\Omega \left[\widehat{\phi}_{tx}\widehat{\phi}_{ty}m_{2x} + \widehat{\phi}_{tx}\widehat{\phi}_{ty}m_{1y}\right] dtd\Omega \\ &\quad + \int_s^\tau\int_\Omega \left[\frac{\widehat{\phi}_{tx}^2 m_{2y}}{2} + \frac{\widehat{\phi}_{ty}^2 m_{1x}}{2}\right] dtd\Omega \\ &= \left(\nabla\widehat{\phi}_t, \nabla\left(\overline{m}\cdot\nabla\widehat{\phi}\right)\right)_{L^2(\Omega)}\Big|_s^\tau \\ &\quad + \int_s^\tau\int_\Omega \left[\frac{\widehat{\phi}_{tx}^2 m_{2y}}{2} + \frac{\widehat{\phi}_{ty}^2 m_{1x}}{2} - \frac{\widehat{\phi}_{tx}^2 m_{1x}}{2} - \frac{\widehat{\phi}_{ty}^2 m_{2y}}{2}\right] dtd\Omega \\ &\quad - \int_s^\tau\int_\Omega \left[\widehat{\phi}_{tx}\widehat{\phi}_{ty}m_{2x} + \widehat{\phi}_{tx}\widehat{\phi}_{ty}m_{1y}\right] dtd\Omega, \end{aligned}$$

after again using the divergence theorem and the fact that $\int_\Omega \operatorname{div}(|\nabla\widehat{\phi}_t|^2\overline{m})d\Omega = \int_{\Gamma_0} |\nabla\widehat{\phi}_t|^2 d\Gamma_0 = 0$ (as $\widehat{\phi}_t(t) \in H^2_{\Gamma_0}(\Omega)$).

(iii) To handle the biharmonic term, we use Green's theorem (1.9), the given boundary conditions of (2.15), (4.14), and the fact that $\widehat{\phi} \in H^2_{\Gamma_0}(\Omega)$ to obtain

(4.18)
$$\int_s^\tau \left(\Delta^2\widehat{\phi}, \overline{m}\cdot\nabla\widehat{\phi}\right)_{L^2(\Omega)} dt = \int_s^\tau a\left(\widehat{\phi}, \overline{m}\cdot\nabla\widehat{\phi}\right) dt$$
$$+\alpha\int_s^\tau\int_{\Gamma_1}\widehat{\psi}\cdot\frac{\partial\overline{m}\cdot\nabla\widehat{\phi}}{\partial\nu}\,d\Gamma_1 dt \;-\int_s^\tau\int_{\Gamma_0}\left(\Delta\widehat{\phi}+(1-\mu)B_1\widehat{\phi}\right)\frac{\partial^2\widehat{\phi}}{\partial\nu^2}d\Gamma_0 dt.$$

We note at this point that we can rewrite the first term on the right-hand side of (4.18) as

(4.19)
$$\begin{aligned}&\int_s^\tau a\left(\widehat{\phi}, \overline{m}\cdot\nabla\widehat{\phi}\right) dt\\ &= \frac{1}{2}\int_s^\tau\int_\Omega \overline{m}\cdot\nabla\left[\widehat{\phi}_{xx}^2+\widehat{\phi}_{yy}^2+2\mu\widehat{\phi}_{xx}\widehat{\phi}_{yy}+2(1-\mu)\widehat{\phi}_{xy}^2\right]dtd\Omega\\ &\quad+\mathcal{O}\left(\int_0^T\left\|\mathring{\mathbf{A}}^{\frac{1}{2}}\widehat{\phi}\right\|^2_{L^2(\Omega)}dt\right),\end{aligned}$$

where $\mathcal{O}(\int_0^T \|\mathring{\mathbf{A}}^{\frac{1}{2}}\widehat{\phi}\|^2_{L^2(\Omega)}dt)$ denotes a series of terms that can be majorized by the $L^2(0,T;D(\mathring{\mathbf{A}}^{\frac{1}{2}}))$-norm of $\widehat{\phi}$. We consequently have by the divergence theorem that

(4.20)
$$\begin{aligned}&\int_s^\tau a\left(\widehat{\phi}, \overline{m}\cdot\nabla\widehat{\phi}\right) dt\\ &= \frac{1}{2}\int_s^\tau\int_\Omega \overline{m}\cdot\nabla\left[\widehat{\phi}_{xx}^2+\widehat{\phi}_{yy}^2+2\mu\widehat{\phi}_{xx}\widehat{\phi}_{yy}+2(1-\mu)\widehat{\phi}_{xy}^2\right]dtd\Omega\\ &\quad+\mathcal{O}\left(\int_0^T\left\|\mathring{\mathbf{A}}^{\frac{1}{2}}\widehat{\phi}\right\|^2_{L^2(\Omega)}dt\right)\\ &= \frac{1}{2}\int_s^\tau\int_\Omega \operatorname{div}\left\{\overline{m}\left[\widehat{\phi}_{xx}^2+\widehat{\phi}_{yy}^2+2\mu\widehat{\phi}_{xx}\widehat{\phi}_{yy}+2(1-\mu)\widehat{\phi}_{xy}^2\right]\right\}dtd\Omega\\ &\quad+\mathcal{O}\left(\int_0^T\left\|\mathring{\mathbf{A}}^{\frac{1}{2}}\widehat{\phi}\right\|^2_{L^2(\Omega)}dt\right)\\ &= \frac{1}{2}\int_s^\tau\int_{\Gamma_0}\left[\widehat{\phi}_{xx}^2+\widehat{\phi}_{yy}^2+2\mu\widehat{\phi}_{xx}\widehat{\phi}_{yy}+2(1-\mu)\widehat{\phi}_{xy}^2\right]dtd\Gamma_0\\ &\quad+\mathcal{O}\left(\int_0^T\left\|\mathring{\mathbf{A}}^{\frac{1}{2}}\widehat{\phi}\right\|^2_{L^2(\Omega)}dt\right)\\ &= \frac{1}{2}\int_s^\tau\int_{\Gamma_0}\left(\Delta\widehat{\phi}\right)^2dt+\mathcal{O}\left(\int_0^T\left\|\mathring{\mathbf{A}}^{\frac{1}{2}}\widehat{\phi}\right\|^2_{L^2(\Omega)}dt\right),\end{aligned}$$

where in the last step above, we have used the fact (as reasoned in [11, Chapter 4]) that $\widehat{\phi}|_{\Gamma_0} = \frac{\partial\widehat{\phi}}{\partial\nu}|_{\Gamma_0} = 0$ implies that $\widehat{\phi}_{xx}^2+\widehat{\phi}_{yy}^2+2\mu\widehat{\phi}_{xx}\widehat{\phi}_{yy}+2(1-\mu)\widehat{\phi}_{xy}^2 = (\Delta\widehat{\phi})^2$ on Γ_0.

To handle the last term on the right-hand side of (4.18), we note that $B_1\widehat{\phi} = 0$ on Γ_0, which implies that

$$\Delta\widehat{\phi} = \Delta\widehat{\phi} + (1-\mu)B_1\widehat{\phi} = \frac{\partial^2\widehat{\phi}}{\partial\nu^2} \text{ on } \Gamma_0 \ . \tag{4.21}$$

We consequently have upon the insertion of (4.20) into (4.18), followed by the consideration of (4.21) that

$$\begin{aligned}\int_s^\tau \left(\Delta^2\widehat{\phi}, \overline{m}\cdot\nabla\widehat{\phi}\right)_{L^2(\Omega)} dt &= -\frac{1}{2}\int_s^\tau \left\|\Delta\widehat{\phi}\right\|^2_{L^2(\Gamma_0)} dt \\ &+ \alpha\int_s^\tau\int_{\Gamma_1} \widehat{\psi}\cdot\frac{\partial\overline{m}\cdot\nabla\widehat{\phi}}{\partial\nu}\, d\Gamma_1 dt \ + \ \mathcal{O}\left(\int_0^T \left\|\mathring{\mathbf{A}}^{\frac{1}{2}}\widehat{\phi}\right\|^2_{L^2(\Omega)} dt\right).\end{aligned} \tag{4.22}$$

(iv) To handle the right-hand side of (4.15), an integration by parts yields

$$\begin{aligned}&\int_s^\tau \left(c_0\widehat{\psi} + c_1\widehat{\psi}_t + c_2\widehat{\phi} + c_3\widehat{\phi}_t + c_4\Delta\widehat{\phi}, \overline{m}\cdot\nabla\widehat{\phi}\right)_{L^2(\Omega)} dt \\ &= c_1\left[\left(\widehat{\psi}, \overline{m}\cdot\nabla\widehat{\phi}\right)_{L^2(\Omega)}\right]_s^\tau - c_1\int_s^\tau \left(\widehat{\psi}, \overline{m}\cdot\nabla\widehat{\phi}_t\right)_{L^2(\Omega)} dt \\ &\quad + \int_s^\tau \left(c_0\widehat{\psi} + c_2\widehat{\phi} + c_3\widehat{\phi}_t + c_4\Delta\widehat{\phi}, \overline{m}\cdot\nabla\widehat{\phi}\right)_{L^2(\Omega)} dt.\end{aligned}$$

As $\overline{m}\cdot\nabla\widehat{\phi} \in C([0,T]; H^1_{\Gamma_0,\gamma}(\Omega))$, we have for all $t \in [0,T]$,

$$\left(\widehat{\psi}(t), \overline{m}\cdot\nabla\widehat{\phi}(t)\right)_{L^2(\Omega)} = \left\langle\widehat{\psi}(t), \overline{m}\cdot\nabla\widehat{\phi}(t)\right\rangle_{H^{-1}_{\Gamma_0,\gamma}(\Omega)\times H^1_{\Gamma_0,\gamma}(\Omega)}.$$

Accordingly, we have

$$\begin{aligned}&\int_s^\tau \left(c_0\widehat{\psi} + c_1\widehat{\psi}_t + c_2\widehat{\phi} + c_3\widehat{\phi}_t + c_4\Delta\widehat{\phi}, \overline{m}\cdot\nabla\widehat{\phi}\right)_{L^2(\Omega)} dt \\ &\quad = c_1\left[\left\langle\widehat{\psi}, \overline{m}\cdot\nabla\widehat{\phi}\right\rangle_{H^{-1}_{\Gamma_0,\gamma}(\Omega)\times H^1_{\Gamma_0,\gamma}(\Omega)}\right]_s^\tau \\ -c_1&\int_s^\tau \left(\widehat{\psi}, \overline{m}\cdot\nabla\widehat{\phi}_t\right)_{L^2(\Omega)} dt + \int_s^\tau \left(c_0\widehat{\psi} + c_2\widehat{\phi} + c_3\widehat{\phi}_t + c_4\Delta\widehat{\phi}, \overline{m}\cdot\nabla\widehat{\phi}\right)_{L^2(\Omega)} dt.\end{aligned} \tag{4.23}$$

To finish the proof, we rewrite (4.15) by collecting the relations given above in (4.16), (4.17), (4.22), and (4.23) to attain

$$\begin{aligned}&\frac{1}{2}\int_s^\tau \left\|\Delta\widehat{\phi}\right\|^2_{L^2(\Gamma_0)} dt = \alpha\int_s^\tau\int_{\Gamma_1} \widehat{\psi}\cdot\frac{\partial\overline{m}\cdot\nabla\widehat{\phi}}{\partial\nu}\, d\Gamma_1 dt \\ &+ \ \mathcal{O}\left(\int_0^T \left\|\mathring{\mathbf{A}}^{\frac{1}{2}}\widehat{\phi}\right\|^2_{L^2(\Omega)} dt\right) + \frac{1}{2}\int_s^\tau\int_\Omega \widehat{\phi}_t^2\left[m_{1x} + m_{2y}\right] dt d\Omega\end{aligned} \tag{4.24}$$

$$
\begin{aligned}
&- \int_s^\tau \left(c_0\widehat{\psi} + c_2\widehat{\phi} + c_3\widehat{\phi}_t + c_4\Delta\widehat{\phi}, \overline{m}\cdot\nabla\widehat{\phi}\right)_{L^2(\Omega)} dt \\
&+ \gamma \int_s^\tau \int_\Omega \left[\frac{\widehat{\phi}_{tx}^2 m_{2y}}{2} + \frac{\widehat{\phi}_{ty}^2 m_{1x}}{2} - \frac{\widehat{\phi}_{tx}^2 m_{1x}}{2} - \frac{\widehat{\phi}_{ty}^2 m_{2y}}{2}\right] dt d\Omega \\
&- \gamma \int_s^\tau \int_\Omega \left[\widehat{\phi}_{tx}\widehat{\phi}_{ty} m_{2x} + \widehat{\phi}_{tx}\widehat{\phi}_{ty} m_{1y}\right] dt d\Omega + c_1 \int_s^\tau \left(\widehat{\psi}, \overline{m}\cdot\nabla\widehat{\phi}_t\right)_{L^2(\Omega)} dt \\
&+ \left[\left(\widehat{\phi}_t, \overline{m}\cdot\nabla\widehat{\phi}\right)_{L^2(\Omega)} + \gamma\left(\nabla\widehat{\phi}_t, \nabla\left(\overline{m}\cdot\nabla\widehat{\phi}\right)\right)_{L^2(\Omega)} - c_1\left\langle\widehat{\psi}, \overline{m}\cdot\nabla\widehat{\phi}\right\rangle_{H^{-1}_{\Gamma_0,\gamma}(\Omega)\times H^1_{\Gamma_0,\gamma}(\Omega)}\right]_s^\tau.
\end{aligned}
$$

The desired inequality (4.13) now comes about by majorizing the right-hand side of this expression (note that in this majorization we are using implicitly the fact that $\frac{\partial\overline{m}\cdot\nabla\widehat{\phi}}{\partial\nu}|_{\Gamma_1}$ is a "lower order term," as $\overline{m}|_{\Gamma_1} = 0$). □

PROPOSITION 4.6. *With the vector field $\overline{h}$ as defined in* (1.3)*, the solution $[\widehat{\phi}, \widehat{\phi}_t, \widehat{\psi}]$ to* (2.15)*, corresponding to terminal data $[\phi_0, \phi_1, \psi_0] \in D([\mathcal{A}_\gamma^*]^2)$, satisfies equality* (2.36) *for arbitrary $\epsilon_0 \in [0, T)$.*

Proof. We multiply (2.15) by $\overline{h}\cdot\nabla\widehat{\phi} - \frac{1}{2}\widehat{\phi}$ and subsequently integrate in time and space; i.e., we will consider the equation

$$
\int_{\epsilon_0}^{T-\epsilon_0} \left(\widehat{\phi}_{tt} - \gamma\Delta\widehat{\phi}_{tt} + \Delta^2\widehat{\phi} - \left[c_0\widehat{\psi} + c_1\widehat{\psi}_t + c_2\widehat{\phi} + c_3\widehat{\phi}_t\right], \overline{h}\cdot\nabla\widehat{\phi} - \frac{1}{2}\widehat{\phi}\right)_{L^2(\Omega)} dt = 0. \tag{4.25}
$$

First, using directly the computations performed in [12] for the quantity $\int_{\epsilon_0}^{T-\epsilon_0}(\widehat{\phi}_{tt} - \gamma\Delta\widehat{\phi}_{tt} + \Delta^2\widehat{\phi}, \overline{h}\cdot\nabla\widehat{\phi} - \frac{1}{2}\widehat{\phi})_{L^2(\Omega)}dt$, in the case that $\overline{h}$ is a radial vector field (see the relations (3.12) and (3.16) of [12]), we have

(4.26)

$$
\begin{aligned}
&\int_{\epsilon_0}^{T-\epsilon_0} \left(\widehat{\phi}_{tt} - \gamma\Delta\widehat{\phi}_{tt} + \Delta^2\phi, \overline{h}\cdot\nabla\widehat{\phi} - \frac{1}{2}\widehat{\phi}\right)_{L^2(\Omega)} dt \\
&= \left[(\widehat{\phi}_t, \overline{h}\cdot\nabla\widehat{\phi})_{L^2(\Omega)} + \gamma\left(\nabla\widehat{\phi}_t, \nabla\left(\overline{h}\cdot\nabla\widehat{\phi}\right)\right)_{L^2(\Omega)} - \frac{1}{2}(\widehat{\phi}_t, \widehat{\phi})_{L^2(\Omega)}\right]_{t=\epsilon_0}^{t=T-\epsilon_0} \\
&\quad - \left[\frac{\gamma}{2}\left(\nabla\widehat{\phi}_t, \nabla\widehat{\phi}\right)_{L^2(\Omega)}\right]_{t=\epsilon_0}^{t=T-\epsilon_0} - \frac{1}{2}\int_{\epsilon_0}^{T-\epsilon_0}\int_{\Gamma_1} \overline{h}\cdot\nu\left(\widehat{\phi}_t^2 + \gamma\left|\nabla\widehat{\phi}_t\right|^2\right) d\Gamma dt \\
&\quad + \int_{\epsilon_0}^{T-\epsilon_0} \left[\frac{1}{2}\left\|P_\gamma^{\frac{1}{2}}\widehat{\phi}_t\right\|_{L^2(\Omega)}^2 + \left\|\widehat{\phi}_t\right\|_{L^2(\Omega)}^2 - \gamma\left(\overline{h}\cdot\nabla\widehat{\phi} - \frac{1}{2}\widehat{\phi}, \frac{\partial\widehat{\phi}_{tt}}{\partial\nu}\right)_{L^2(\Gamma_1)}\right] dt \\
&\quad + \int_{\epsilon_0}^{T-\epsilon_0} \left[\frac{1}{2}\left\|\mathring{\mathbf{A}}^{\frac{1}{2}}\widehat{\phi}\right\|^2 + \frac{1}{2}\int_{\Gamma_0} \overline{h}\cdot\nu\left(\Delta\widehat{\phi}\right)^2 d\Gamma - \left(\Delta\widehat{\phi}, \frac{\partial}{\partial\nu}\left(\overline{h}\cdot\nabla\widehat{\phi}\right)\right)_{L^2(\Gamma_0)}\right] dt \\
&\quad + \int_{\epsilon_0}^{T-\epsilon_0} \left[\left(\frac{\partial\Delta\widehat{\phi}}{\partial\nu} + (1-\mu)\frac{\partial B_2\widehat{\phi}}{\partial\tau}, \overline{h}\cdot\nabla\widehat{\phi} - \frac{1}{2}\widehat{\phi}\right)_{L^2(\Gamma_1)}\right. \\
&\qquad \left. - \left(\Delta\widehat{\phi} + (1-\mu)B_1\widehat{\phi}, \frac{\partial}{\partial\nu}\left(\overline{h}\cdot\nabla\widehat{\phi} - \frac{1}{2}\widehat{\phi}\right)\right)_{L^2(\Gamma_1)}\right] dt
\end{aligned}
$$

$$+\int_{\epsilon_0}^{T-\epsilon_0}\int_{\Gamma_1}\frac{\overline{h}\cdot\nu}{2}\left[\left(\frac{\partial^2\widehat{\phi}}{\partial x^2}\right)^2+\left(\frac{\partial^2\widehat{\phi}}{\partial y^2}\right)^2+2\mu\left(\frac{\partial^2\widehat{\phi}}{\partial x^2}\right)\left(\frac{\partial^2\widehat{\phi}}{\partial y^2}\right)+2(1-\mu)\left(\frac{\partial^2\widehat{\phi}}{\partial x\partial y}\right)^2\right]dtd\Gamma.$$

Using the boundary conditions in (2.15) and the fact that $\frac{\partial(\overline{h}\cdot\nabla\phi)}{\partial\nu}|_{\Gamma_0}=(\overline{h}\cdot\nu)\Delta\phi|_{\Gamma_0}$, this equation becomes

(4.27)

$$\begin{aligned}
&\int_{\epsilon_0}^{T-\epsilon_0}\left(\widehat{\phi}_{tt}-\gamma\Delta\widehat{\phi}_{tt}+\Delta^2\phi,\overline{h}\cdot\nabla\widehat{\phi}-\frac{1}{2}\widehat{\phi}\right)_{L^2(\Omega)}dt=\frac{1}{2}\int_{\epsilon_0}^{T-\epsilon_0}E_{\widehat{\phi}}(t)dt\\
&+\left[(\widehat{\phi}_t,\overline{h}\cdot\nabla\widehat{\phi})_{L^2(\Omega)}+\gamma\left(\nabla\widehat{\phi}_t,\nabla\left(\overline{h}\cdot\nabla\widehat{\phi}\right)\right)_{L^2(\Omega)}-\frac{1}{2}(\widehat{\phi}_t,\widehat{\phi})_{L^2(\Omega)}\right]_{t=\epsilon_0}^{t=T-\epsilon_0}\\
&-\left[\frac{\gamma}{2}\left(\nabla\widehat{\phi}_t,\nabla\widehat{\phi}\right)_{L^2(\Omega)}\right]_{t=\epsilon_0}^{t=T-\epsilon_0}-\frac{1}{2}\int_{\epsilon_0}^{T-\epsilon_0}\int_{\Gamma_1}\overline{h}\cdot\nu\left(\widehat{\phi}_t^2+\gamma\left|\nabla\widehat{\phi}_t\right|^2\right)d\Gamma dt\\
&+\int_{\epsilon_0}^{T-\epsilon_0}\left[\left\|\widehat{\phi}_t\right\|_{L^2(\Omega)}^2-\frac{1}{2}\int_{\Gamma_0}\overline{h}\cdot\nu\left(\Delta\widehat{\phi}\right)^2d\Gamma\right]dt\\
&+\int_{\epsilon_0}^{T-\epsilon_0}\left[\alpha\left(\widehat{\psi},\frac{\partial}{\partial\nu}\left(\overline{h}\cdot\nabla\widehat{\phi}-\frac{1}{2}\widehat{\phi}\right)\right)_{L^2(\Gamma_1)}\right.\\
&\qquad\left.+\left(\gamma\frac{\partial}{\partial\nu}\left(\xi^2\widehat{\phi}+2\xi\widehat{\phi}_t\right)-\alpha\frac{\partial\widehat{\psi}}{\partial\nu},\overline{h}\cdot\nabla\widehat{\phi}-\frac{1}{2}\widehat{\phi}\right)_{L^2(\Gamma_1)}\right]dt\\
&+\int_{\epsilon_0}^{T-\epsilon_0}\int_{\Gamma_1}\frac{\overline{h}\cdot\nu}{2}\left[\left(\frac{\partial^2\widehat{\phi}}{\partial x^2}\right)^2+\left(\frac{\partial^2\widehat{\phi}}{\partial y^2}\right)^2+2\mu\left(\frac{\partial^2\widehat{\phi}}{\partial x^2}\right)\left(\frac{\partial^2\widehat{\phi}}{\partial y^2}\right)\right.\\
&\qquad\left.+2(1-\mu)\left(\frac{\partial^2\widehat{\phi}}{\partial x\partial y}\right)^2\right]dtd\Gamma.
\end{aligned}$$

Second, we multiply $[c_0\widehat{\psi}+c_1\widehat{\psi}_t+c_2\widehat{\phi}+c_3\widehat{\phi}_t+c_4\Delta\widehat{\phi}]$ by $\overline{h}\cdot\nabla\widehat{\phi}-\frac{1}{2}\widehat{\phi}$ and integrate by parts to obtain

(4.28)

$$\begin{aligned}
&\int_{\epsilon_0}^{T-\epsilon_0}\left(c_0\widehat{\psi}+c_1\widehat{\psi}_t+c_2\widehat{\phi}+c_3\widehat{\phi}_t+c_4\Delta\widehat{\phi},\overline{h}\cdot\nabla\widehat{\phi}-\frac{1}{2}\widehat{\phi}\right)_{L^2(\Omega)}dt\\
&=c_1\left[\left(\widehat{\psi},\overline{h}\cdot\nabla\widehat{\phi}-\frac{1}{2}\widehat{\phi}\right)_{L^2(\Omega)}\right]_{t=\epsilon_0}^{t=T-\epsilon_0}\\
&\quad-c_1\int_{\epsilon_0}^{T-\epsilon_0}\left(\widehat{\psi},\overline{h}\cdot\nabla\widehat{\phi}_t-\frac{1}{2}\widehat{\phi}_t\right)_{L^2(\Omega)}dt
\end{aligned}$$

$$+\int_{\epsilon_0}^{T-\epsilon_0}\left(c_0\widehat{\psi}+c_2\widehat{\phi}+c_3\widehat{\phi}_t+c_4\Delta\widehat{\phi},\overline{h}\cdot\nabla\widehat{\phi}-\frac{1}{2}\widehat{\phi}\right)_{L^2(\Omega)}dt.$$

To now obtain (2.36), we combine the expressions (4.25) and (4.27)–(4.28) and follow this by a rearrangement of terms. □

REFERENCES

[1] J. P. Aubin, *Analyse fonctionelle appliquée,* Tome 2, Presses Universitaire de France, Paris, 1979.

[2] G. Avalos, *The exponential stability of a coupled hyperbolic/parabolic system arising in structural acoustics*, Abstr. Appl. Anal., 1 (1996), pp. 203–217.

[3] G. Avalos and I. Lasiecka, *Exponential stability of a thermoelastic system with free boundary conditions without mechanical dissipation*, SIAM J. Math. Anal., 29 (1998), pp. 155–182.

[4] G. Avalos and I. Lasiecka, *Exponential stability of a thermoelastic system without mechanical dissipation*, Rend. Instit. Mat. Univ. Trieste, 28 (1997), pp. 1–28.

[5] G. Avalos, *Exact controllability of a thermoelastic system with control in the thermal component only*, Differential Integral Equations, to appear.

[6] L. de Teresa and E. Zuazua, *Controllability for the linear system of thermoelastic plates*, Adv. Differential Equations, 1 (1996), pp. 369–402.

[7] P. Grisvard, *Caracterization de quelques espaces d'interpolation*, Arch. Rational Mech. Anal., 25 (1967), pp. 40–63.

[8] S. Hansen and B. Zhang, *Boundary control of a linear thermoelastic beam*, J. Math. Anal. Appl., 210 (1997), pp. 182–205.

[9] V. Hutson and J. S. Pym, *Applications of Functional Analysis and Operator Theory*, Academic Press, New York, 1980.

[10] V. Isakov, *On the uniqueness of the continuation for a thermoelasticity system*, Differential Equations, to appear.

[11] J. E. Lagnese, *Boundary Stabilization of Thin Plates*, SIAM Stud. Appl. Math. 10, SIAM, Philadelphia, 1989.

[12] J. Lagnese, *The reachability problem for thermoelastic plates*, Arch. Rational Mech. Anal., 112 (1990), pp. 223–267.

[13] I. Lasiecka and R. Triggiani, *Uniform stabilization of the wave equation with Dirichlet or Neumann feedback control without geometrical conditions*, Appl. Math. Optim., 25 (1992), pp. 189–224.

[14] I. Lasiecka and R. Triggiani, *Exact controllability of the wave equation with Neumann boundary control*, Appl. Math. Optim., 19 (1989), pp. 243–290.

[15] I. Lasiecka and R. Triggiani, *Sharp trace estimates of solutions to Kirchoff and Euler–Bernoulli equations*, Appl. Math. Optim., 28 (1993), pp. 277–306.

[16] I. Lasiecka and R. Triggiani, *Structural decomposition of thermoelastic semigroups with rotational forces*, Semigroup Forum, 5 (1999), pp. 585–599.

[17] I. Lasiecka and R. Triggiani, *Sharp regularity theory for elastic and thermo-elastic Kirchoff equations with free boundary conditions*, Rocky Mountain J. Math., to appear.

[18] I. Lasiecka and R. Triggiani, *Analyticity of thermo-elastic semigroups with free boundary conditions*, Ann. Scuola Norm. Sup. Pisa Cl. Sci. (3-4), 27 (1998), pp. 457–482.

[19] G. Lebeau and E. Zuazua, *Null–controllability of a system of linear thermoelasticity*, Arch. Rational Mech. Anal., 141 (1998), pp. 297–329.

[20] G. Leugering *A decomposition method for integro–partial differential equations and applications*, J. Math. Pures Appl., 71 (1992), pp. 561–587.

[21] J. L. Lions and E. Magenes, *Non-Homogeneous Boundary Value Problems and Applications*, vol. 1., Springer-Verlag, New York, 1972.

[22] W. Liu, *Partial exact controllability and exponential stability in higher–dimensional linear thermoelasticity*, ESAIM Control Optim. Calc. Var., 3 (1998), pp. 23–48.

[23] W. Liu, *Erratum on partial exact controllability and exponential stability in higher-dimensional linear thermoelasticity*, ESAIM Control Optim. Calc. Var., 3 (1998), pp. 323–328.

[24] J. Simon, *Compact sets in the space $L^p(0,T;B)$*, Ann. Mat. Pura Appl. (4), 146 (1987), pp. 65–96.

SIAM J. CONTROL OPTIM.
Vol. 38, No. 2, pp. 384–399

MINIMIZATION OF FUNCTIONALS OF THE GRADIENT BY BAIRE'S THEOREM*

SANDRO ZAGATTI†

Abstract. We give sufficient conditions for the existence of solutions of the minimum problem

$$\mathcal{P}_{u_0}: \qquad \text{Minimize} \quad \int_\Omega g(Du(x))dx, \quad u \in u_0 + W_0^{1,p}(\Omega,\mathbb{R}),$$

based on the structure of the epigraph of the lower convex envelope of g, which is assumed be lower semicontinuous and to grow at infinity faster than the power p with p larger than the dimension of the space. No convexity conditions are required on g, and no assumptions are made on the boundary datum $u_0 \in W_0^{1,p}(\Omega,\mathbb{R})$.

Key words. calculus of variations, minimum problem, Baire category method

AMS subject classifications. 49J10, 49J45

PII. S0363012998335206

1. Introduction.

Consider the problem

$$\mathcal{P}_{u_0}: \qquad \text{Minimize} \ \ \mathcal{I}(u) = \int_\Omega g(Du(x))dx$$

on $u \in u_0 + W_0^{1,1}(\Omega,\mathbb{R})$, where Ω is an open and bounded subset of $\mathbb{R}^n$ and $g : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ is lower semicontinuous and satisfies $g(y) \geq \phi(|y|)$ for some continuous function ϕ with superlinear growth at infinity. As the theorem quoted below shows clearly, for $n > 1$, if we do not assume the convexity of g, in general $\mathcal{P}_{u_0}$ fails to have solutions since $\mathcal{I}$ is not weak lower semicontinuous on $W^{1,1}(\Omega,\mathbb{R})$. However, if the boundary datum u_0 is assumed to be affine, i.e., $u_0 = A_\xi := \langle \xi, \cdot \rangle$ for some $\xi \in \mathbb{R}^n$, the existence of minimum points for $\mathcal{I}$ on $A_\xi + W_0^{1,p}(\Omega,\mathbb{R})$ can be completely characterized by the following result, due to Cellina [C1], [C2].

THEOREM. *Problem $\mathcal{P}_{A_\xi}$ admits at least one solution if and only if either $g(\xi) = g^{**}(\xi)$ or the point $(\xi, g^{**}(\xi))$ belongs to the relative interior of an n-dimensional proper face of* epi(g^{**})*, where g^{**} is the lower convex envelope of g and* epi(g^{**}) *denotes its epigraph.*

By the lower semicontinuity of g, if $(y, g^{**}(y))$ is an extreme point of epi(g^{**}), then necessarily $g^{**}(y) = g(y)$; hence Cellina's result implies in particular that if g^{**} satisfies hypothesis (H) (the dimension of any proper face of the epigraph of g^{**} is either 0 or n), then problem $\mathcal{P}_{A_\xi}$ admits at least one solution for any affine boundary datum A_ξ.

In this paper we prove that under condition (H), if the n-dimensional faces of epi(g^{**}) are finitely many and pairwise disjoint and if $\phi(|y|) = a|y|^p - b$ for some positive constant a, b and for $p > n$, problem $\mathcal{P}_{u_0}$ admits at least one solution for any $u_0 \in W^{1,p}(\Omega,\mathbb{R})$.

*Received by the editors March 6, 1998; accepted for publication (in revised form) April 19, 1999; published electronically January 11, 2000.

http://www.siam.org/journals/sicon/38-2/33520.html

†Scuola Internazionale Superiore di Studi Avanzati (SISSA), via Beirut 2-4, I-34014 Trieste, Italy (zagatti@sissa.it).

To prove this result we make use of Baire's theorem, considering the nonempty set $\mathcal{S}$ of solutions of the minimum problem for the relaxed functional

$$\overline{\mathcal{P}}_{u_0}: \qquad \text{Minimize } \overline{\mathcal{I}}(u) = \int_\Omega g^{**}(Du(x))dx, \quad u \in u_0 + W_0^{1,p}(\Omega, \mathbb{R}),$$

endowed with a distance with respect to which it is a complete metric space, and introducing suitable open and dense subsets $\mathcal{S}_k \subseteq \mathcal{S}$ such that, for any element u of $\bigcap_k \mathcal{S}_k$, $g^{**}(Du) = g(Du)$ almost everywhere in Ω. Consequently there exists (a dense subset of) $v \in \mathcal{S}$ such that $\mathcal{I}(v) = \overline{\mathcal{I}}(v)$, and, since the infimum of $\mathcal{I}$ equals the minimum of $\overline{\mathcal{I}}$, v solves $\mathcal{P}_{u_0}$.

To do this we use techniques introduced in the framework of differential inclusions (see, for example, [B], [BF], [DP1] and the references quoted there) and more recently implemented in the study of existence of generalized solutions for Hamilton–Jacobi equations [DM1], [DM2], [DP2], [Z].

2. Preliminaries and notation. In this paper an element x of $\mathbb{R}^n$ is written $x = (x_1, \dots, x_n)$ and we denote by $\langle \cdot, \cdot \rangle$ and by $|\cdot|$, respectively, the scalar product and the Euclidean norm in $\mathbb{R}^n$; for $x \in \mathbb{R}^n$ and $r > 0$, $B(x, r)$ is the open ball centered in x of radius r. If x and y are elements of $\mathbb{R}^n$ by $[x, y]$ we mean the closed line segment of endpoints x and y, i.e., $[x, y] = \{z \in \mathbb{R}^n : z = (1 - t)x + ty, t \in [0, 1]\}$. Given $A \subseteq \mathbb{R}^n$, ∂A, $\text{int}(A)$, $\overline{A}$, $\text{diam}(A)$, $\text{Ls}(A)$, $\text{co}(A)$, and χ_A, $m_n(A)$, respectively, denote the boundary, the interior, the closure, the diameter, the linear span, the convex hull, the characteristic function, and the n-dimensional Lebesgue measure of A; by $\dim(A)$ we mean the dimension of $\text{Ls}(A)$. The relative interior of A (r.i.(A)) and the relative boundary of A (r.b.(A)) are the interior and the boundary of A relative to $\text{Ls}(A)$. Given two sets A and B, $A \triangle B$ is the symmetric difference $(A - B) \cup (B - A)$.

We need the notion of *face* of a convex set. Given a convex subset C of $\mathbb{R}^n$, a face of C is a convex subset C' of C such that every closed line segment in C whose relative interior intersects C' has both endpoints in C'. A face of C different from C itself is said to be a *proper face*. The zero-dimensional faces of C are actually singletons and are called *extreme points* of C. The set of extreme points of C is denoted by $\text{extr}(C)$. The following facts are well known (see [R, pp. 162–169]).

PROPOSITION 2.1. *Let $C \subseteq \mathbb{R}^n$ be a convex set.*

(i) *If C'' is a face of C' and C' is a face of C, then C'' is a face of C.*

(ii) *A proper face C' of C is entirely contained in the relative boundary of C so that, in particular, $\dim(C') < \dim(C)$.*

(iii) *The collection of the relative interiors of the faces of C is a partition of C itself.*

A family $\mathcal{V}$ of closed subsets of $\mathbb{R}^n$ is said to cover a set $A \subseteq \mathbb{R}^n$ in the Vitali sense (or to be a Vitali covering of E) if for any $x \in A$ there exist a positive number $r(x) > 0$, a sequence of n-dimensional balls $(B(x, \rho_k))_{k \in \mathbb{N}}$, with $\rho_k \to 0$ as $k \to \infty$, and a sequence $(V_k)_{k \in \mathbb{N}}$ in $\mathcal{V}$ such that $V_k \subseteq B(x, \rho_k)$ for any k and

$$\frac{m_n(V_k)}{m_n(B(x, \rho_k))} > r(x) \qquad \forall k \in \mathbb{N}.$$

According to the Vitali covering theorem (see [F, pp. 205–207]) given a Vitali covering $\mathcal{V}$ and a positive ϵ, there exists a finite subfamily $\{V_k, k = 1, \dots, m\}$ of $\mathcal{V}$ such that $V_k \bigcap V_j = \emptyset$ for $j \neq k$ and $m_n\left(A - \bigcup_{k=1}^m V_k\right) \leq \epsilon$.

We recall (see for example [WZ, p. 107]) that, given a measurable subset A of $\mathbb{R}^n$, almost every point x in A is a point of density, i.e.,

$$\lim_{\rho\to 0}\frac{m_n(A\cap B(x,\rho))}{m_n(B(x,\rho))}=1.$$

For a map $u:\Omega(\subseteq\mathbb{R}^n)\to\mathbb{R}$, Du denotes the gradient of u: $Du=(\frac{\partial u}{\partial x_1},\dots,\frac{\partial u}{\partial x_n})$. We shall use the spaces $L^\infty(\Omega,\mathbb{R})$, $L^1(\Omega,\mathbb{R})$, $W^{1,p}(\Omega,\mathbb{R})$, $W_0^{1,p}(\Omega,\mathbb{R})$, $W^{1,\infty}(\Omega,\mathbb{R})$, and $W_0^{1,\infty}(\Omega,\mathbb{R})$ endowed with their usual topologies.

We recall (see [KS, p. 50]) the following well-known proposition.

PROPOSITION 2.2. *Let $\Omega\subseteq\mathbb{R}^n$ and $u\in W^{1,q}(\Omega,\mathbb{R})$ $(1\le q\le+\infty)$; set*

$$v(x):=\max\{0,u(x)\}.$$

Then v belongs to $W^{1,q}(\Omega,\mathbb{R})$, and

$$Dv(x)=\begin{cases}Du(x) & \text{a.e. in } \{x\in\Omega:u(x)>0\},\\ 0 & \text{a.e. in } \{x\in\Omega:u(x)\le 0\}.\end{cases}$$

Let $g:\mathbb{R}^n\to\mathbb{R}\cup\{+\infty\}$; we set $\mathrm{dom}(g)=\{y\in\mathbb{R}^n:g(y)\in\mathbb{R}\}$ and denote by g^{**} the lower convex envelope of g (see, for example, [ET]), remarking that the epigraph of g^{**}, $\mathrm{epi}(g^{**}):=\{(y,t)\in\mathbb{R}^n\times\mathbb{R}:t\ge g^{**}(y)\}$ is a closed convex subset of $\mathbb{R}^n\times\mathbb{R}$.

DEFINITION 2.3. *Let $g:\mathbb{R}^n\to\mathbb{R}\cup\{+\infty\}$ be lower semicontinuous. We say that g satisfies condition* (A) *if*

(A.i) *there exist two positive constant a,b such that*

$$g(y)\ge a|y|^p-b\qquad\forall y\in\mathbb{R}^n,$$

where $p>n$.

(A.ii) *There exist $F^1,\dots,F^N$ proper nonvertical faces of* $\mathrm{epi}(g^{**})$ *such that* $\dim(F^i)=n$ *for any $i=1,\dots,N$, $F^i\bigcap F^j=\emptyset$ when $i\ne j$ and any proper face $F\ne F^i$ for any $i\in\{1,\dots,N$ has dimension equal to zero, i.e., is an extreme point.*

Remark 2.4. For any $i=1,\dots,N$, $F^i\subseteq\partial(\mathrm{epi}(g^{**}))$ and any extreme point of $\mathrm{epi}(g^{**})$ is of the form $(y,g^{**}(y))$.

DEFINITION 2.5. *For any n-dimensional face F^i $(i=1,\dots,N)$ of* $\mathrm{epi}(g^{**})$ *we set*

$$F_0^i:=\{y\in\mathbb{R}^n:(y,g^{**}(y))\in F^i\}.$$

Example. Let $n=3$ and $g(y):=(|y|^2-1)^2$. The map g satisfies (A.i) and (A.ii). Indeed we may choose $a=\frac{1}{2}$, $b=3$, and $p=4$; moreover, any point $(y,g^{**}(y))$ such that $|y|\ge 1$ is an extreme point of $\mathrm{epi}(g^{**})$ and $F^1:=\{(y,0),|y|\le 1\}$, which is a n-dimensional set, is the unique proper face of $\mathrm{epi}(g^{**})$ having nonzero dimension. In this case we have $F_0^1=\overline{B}(0,1)$.

We collect now some properties of $\mathrm{epi}(g^{**})$ that will be useful in the following.

PROPOSITION 2.6. *Let g satisfy* (A). *Then*

(i) *F^i is a compact convex subset of $\mathbb{R}^n\times\mathbb{R}$ for any $i=1,\dots,N$.*

(ii) *F_0^i is a compact convex subset of $\mathbb{R}^n$ with nonempty interior for any $i=1,\dots,N$.*

(iii) $\mathrm{r.b.}(F^i)=\mathrm{extr}(F^i)\subseteq\mathrm{extr}(\mathrm{epi}(g^{**}))$ *for any $i=1,\dots,N$.*

(iv) $\mathrm{extr}(F_0^i)=\partial(F_0^i)$ *and, if $i\ne j$, $F_0^i\bigcap F_0^j=\emptyset$.*

(v) *The set* $\{y \in \mathbb{R}^n : g(y) = g^{**}(y)\}$ *is contained in the set* $(\mathbb{R}^n - \bigcup_{i=1}^N \operatorname{int}(F_0^i))$.

(vi) *The map* g^{**} *is affine on each* F_0^i*; i.e., there exist* $\alpha_i \in \mathbb{R}^n$, $\beta_i \in \mathbb{R}$, *such that* $g^{**}(y) = \langle \alpha_i, y\rangle + \beta_i$ *for any* $y \in F_0^i$ *and for any* $i = 1, \dots, N$.

(vii) *Let* $y_1, y_2 \in \mathbb{R}^n$, $y_1 \neq y_2$ *be such that either* $y_1, y_2 \in \mathbb{R}^n - \bigcup_{i=1}^N F_0^i$ *or* $y_1 \in F_0^j$ *for some* j *and* $y_2 \in \mathbb{R}^n - F_0^j$. *Then, for any* $\lambda \in]0,1[$,

$$g^{**}((1-\lambda)y_1 + \lambda y_2) < (1-\lambda)g^{**}(y_1) + \lambda g^{**}(y_2). \tag{2.1}$$

Proof. Statement (i) is an obvious consequence of the definition of face and of the growth condition expressed in point (A.i) of Definition 2.3. Moreover it implies obviously (ii).

To prove (iii) take $z \in$ r.b.(F^i). By point (i) and (iii) of Proposition 2.1 z belongs to the relative interior of a proper face F of F^i and, by point (ii) of the same proposition, $\dim(F) < n$; but, again by point (i) of Proposition 2.1, F is a face of $\operatorname{epi}(g^{**})$ and this implies, by assumption (A.ii) of Definition 2.3, that $\dim(F) = 0$.

Point (iv) is a consequence of the fact that

$$\operatorname{extr}(F_0^i) = \{y \in \mathbb{R}^n : (y, g^{**}(y)) \in \operatorname{extr}(F^i)\}$$

of previous point (iii) and of the fact that $F^i \bigcap F^j = \emptyset$.

Point (v) follows from (iv), from Remark 2.4, and from the lower semicontinuity of g. To prove (vi) see [ET, Chap. 2] or [C1].

To prove point (vii) assume by contradiction that for some $\lambda \in]0,1[$ equality holds in equation (2.1) instead of strict inequality. This means that the point $P_\lambda = ((1-\lambda)y_1 + \lambda y_2), g^{**}((1-\lambda)y_1 + \lambda y_2)$, which belongs to the relative interior of the line segment $L = [(y_1, g^{**}(y_1)), (y_2, g^{**}(y_2))] \subseteq \operatorname{epi}(g^{**})$, belongs to $\partial(\operatorname{epi}(g^{**}))$; then P_λ belongs to some proper face H of $\operatorname{epi}(g^{**})$ and, by the definition of face, both endpoints of the line segment L lie in H. But H must have dimension larger than one, since it contains L, and this, recalling that by Definition 2.3 the only faces of $\operatorname{epi}(g^{**})$ having dimension larger than zero are the n-dimensional faces F_0^i, implies that $(y_1, g^{**}(y_1))$ and $(y_2, g^{**}(y_2))$ belongs to F^j for some $j \in \{1, \dots, N\}$, which is a contradiction in both cases. ∎

DEFINITION 2.7. *Let* $E \subseteq \mathbb{R}^n$ *be measurable and* $K \subseteq \mathbb{R}^n$ *compact and convex. We define the map* $h(\cdot, K) : \mathbb{R}^n \to \mathbb{R} \cup \{-\infty\}$ *by*

$$h(\xi, K) = \begin{cases} \sup\left\{ \left(\int_0^1 |\phi(y) - \xi|^2 dy\right)^{\frac{1}{2}}, \phi : [0,1] \to K, \int_0^1 \phi(y)dy = \xi \right\}, & \xi \in K, \\ -\infty, & \xi \in \mathbb{R}^n - K, \end{cases}$$

and the likelihood functional

$$\mathcal{L}(V, E) := \int_E h(V(x), K)dx$$

for $V \in L^\infty(E, \mathbb{R}^n)$.

PROPOSITION 2.8.

(i) *The map* $\xi \to h(\xi, K)$ *is upper semicontinuous and strictly concave.*

(ii) $h(\xi, K) \geq 0$ *for any* $\xi \in K$ *and* $h(\xi, K) = 0$ *if and only if* $\xi \in \operatorname{extr}(K)$.

(iii) $|h(\xi, K)| \leq \operatorname{diam}(K)$.

(iv) *The functional* $\mathcal{L}(\cdot, E)$ *is upper semicontinuous with respect to weak* topology of* $L^\infty(E, \mathbb{R}^n)$*; i.e., for any sequence* $\{V_k\}_{k\in\mathbb{N}}$ *such that* $V_k \overset{*}{\rightharpoonup} V$ *in* $L^\infty(E, \mathbb{R}^n)$, *we have*

$$\mathcal{L}(V, E) \geq \limsup_{k\to\infty} \mathcal{L}(V_k, E).$$

For the proof of (i), (ii), and (iii) see [B]. To prove (iv) see Theorem 1.2, p. 49, of [D].

3. Statement of the result. Throughout this paper we assume that Ω is an open bounded subset of $\mathbb{R}^n$, $g : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ satisfies condition (A), and u_0 denotes an element of $W^{1,p}(\Omega, \mathbb{R})$.

We define the set of admissible functions

$$\mathcal{W} := u_0 + W_0^{1,p}(\Omega, \mathbb{R})$$

and the functionals

$$\mathcal{I}(u) := \int_\Omega g(Du(x))dx, \qquad \overline{\mathcal{I}}(u) := \int_\Omega g^{**}(Du(x))dx,$$

and we consider the problems

$$\mathcal{P}: \quad \text{Minimize } \mathcal{I}(u), \qquad u \in \mathcal{W},$$

and

$$\overline{\mathcal{P}}: \quad \text{Minimize } \overline{\mathcal{I}}(u), \qquad u \in \mathcal{W}.$$

Clearly for some boundary data u_0 we could have $\mathcal{I}(u) = \overline{\mathcal{I}}(u) = +\infty$ for any $u \in \mathcal{W}$. To avoid triviality, we shall assume, however, that the functionals are finite for some $u \in \mathcal{W}$.

The following is a well-known fact (see, for example, [ET, Chap. IX] or [D, Chap. 3]).

THEOREM 3.1. *$\overline{\mathcal{I}}$ is coercive and weak lower semicontinuous on $W^{1,p}(\Omega, \mathbb{R})$. Consequently $\overline{\mathcal{P}}$ admits at least one solution. Moreover,*

$$m := \min\{\overline{\mathcal{I}}(u), u \in \mathcal{W}\} \le \inf\{\mathcal{I}(u), u \in \mathcal{W}\}.$$

The main result of this paper, which will be proved in the next section, is the following theorem.

THEOREM 3.2. *$\mathcal{P}$ admits at least one solution.*

4. Proof of Theorem 3.2.

DEFINITION 4.1. *We set $\mathcal{S}(\overline{\mathcal{P}}) := \{u \in \mathcal{W} : u \text{ is a solution of } \overline{\mathcal{P}}\}$.*

DEFINITION 4.2. *Let $u \in \mathcal{S}(\overline{\mathcal{P}})$ and $i \in \{1, \dots, N\}$; we set*

$$E_u^i := \{x \in \Omega : Du(x) \in F_0^i\}.$$

Remark 4.3. The sets E_u^i are measurable.

PROPOSITION 4.4. *Let $u, v \in \mathcal{S}(\overline{\mathcal{P}})$, $i, j \in \{1, \dots, N\}$, $i \neq j$. Then*

$$m_n(E_u^i \triangle E_v^i) = 0, \tag{4.1}$$

$$m_n(E_u^i \cap E_u^j) = 0. \tag{4.2}$$

Proof. We prove that $m_n(E_u^i - E_v^i) = 0$; then (4.1) follows interchanging u with v. Assume, by contradiction, $m_n(E_u^i - E_v^i) > 0$ and set $G = E_u^i - E_v^i$. For almost

every $x \in G$, $Du(x) \in F_0^i$, and $Dv(x) \in \mathbb{R}^n - (F_0^i)$; hence, recalling point (vii) of Proposition 2.6, for any $\lambda \in]0,1[$, we have

$$g^{**}((1-\lambda)Du(x) + \lambda Dv(x)) < (1-\lambda)g^{**}(Du(x)) + \lambda g^{**}(Dv(x))$$

almost everywhere in G. Define the map $w := (1-\lambda)u + \lambda v$. Clearly w belongs to $\mathcal{W}$ and, by the convexity of g^{**} and the above inequality, we have

$$\begin{aligned}
\overline{\mathcal{I}}(w) &= \int_G g^{**}(Dw(x))dx + \int_{\Omega - G} g^{**}(Dw(x))dx \\
&< (1-\lambda)\int_G g^{**}(Du(x))dx + \lambda \int_G g^{**}(Dv(x))dx + \int_{\Omega-G} g^{**}(Dw(x))dx \\
&\leq (1-\lambda)\int_G g^{**}(Du(x))dx + \lambda \int_G g^{**}(Dv(x))dx \\
&\quad + (1-\lambda)\int_{\Omega-G} g^{**}(Du(x))dx + \lambda \int_{\Omega-G} g^{**}(Dv(x))dx \\
&= (1-\lambda)\overline{\mathcal{I}}(u) + \lambda\overline{\mathcal{I}}(v) = m,
\end{aligned}$$

which is a contradiction.

To prove the second assertion assume $m_n(E_u^i \bigcap E_u^j) > 0$. This implies that given any representative of u there exists some $x \in E_u^i \bigcap E_u^j$ such that $Du(x) \in F_0^i \bigcap F_0^j$, and this contradicts point (iv) of Proposition 2.6. □

DEFINITION 4.5. *Take $u \in \mathcal{S}(\overline{\mathcal{P}})$ and set*

$$E^i := E_u^i, \qquad E := \bigcup_{i=1}^N E^i.$$

Remark 4.6.

(i) In view of (4.1) the sets E^i are defined modulo null sets in the sense that, given any $v \in \mathcal{S}(\overline{\mathcal{P}})$, $Dv(x) \in F_0^i$ for almost every $x \in E^i$, and then, choosing a suitable representative of v, we may assume

$$E_v^i = E^i \quad \forall v \in \mathcal{S}(\overline{\mathcal{P}}) \quad \forall i \in \{1, \dots, N\}. \tag{4.3}$$

(ii) By (4.2) we may also assume

$$E^i \bigcap E^j = \emptyset, \tag{4.4}$$

observing that the sets E^i may be empty.

PROPOSITION 4.7. *Let $u, v \in \mathcal{S}(\overline{\mathcal{P}})$. Then $Du(x) = Dv(x)$ for almost every $x \in \Omega - E$.*

Proof. First, we remark that by the definition of E, $Du(x), Dv(x) \in \mathbb{R}^n - \left(\bigcup_{i=1}^s F_0^i\right)$ for almost every $x \in \Omega - E$.

Assume that there exists $G \subseteq \Omega - E$ with positive measure such that $Du(x) \neq Dv(x)$ for $x \in G$; recalling point (iii) of Proposition 2.6 we have, for any $\lambda \in]0,1[$,

$$g^{**}((1-\lambda)Du(x) + \lambda Dv(x)) < (1-\lambda)g^{**}(Du(x)) + \lambda g^{**}(Dv(x)), \quad x \in G.$$

Then we define $w := (1-\lambda)u + \lambda v \in \mathcal{W}$ and, by the same computations of the proof of Proposition 4.4, get

$$\overline{\mathcal{I}}(w) = \int_G g^{**}(Dw(x))dx + \int_{\Omega-G} g^{**}(Dw(x))dx < m,$$

which is a contradiction. □

DEFINITION 4.8.

(i) *For any $i \in \{1, \dots, N\}$ we define the functional*

$$\mathcal{L}(u, E^i) := \int_{E^i} h(Du(x), F_0^i)dx, \qquad u \in \mathcal{S}(\overline{\mathcal{P}}),$$

where h is given by Definition 4.1 *and the first argument of the integrand is intended as the restriction to E^i of the $L^1(\Omega, \mathbb{R}^n)$-map Du. By Remark* 4.6 *and by the boundedness of F_0^i such a restriction belongs to $L^\infty(E^i, \mathbb{R}^n)$.*

(ii) *For any $\alpha > 0$ we define the sets*

$$\mathcal{S}_\alpha := \{u \in \mathcal{S}(\overline{\mathcal{P}}) : \mathcal{L}(u, E^i) < \alpha \text{ for every } i \in \{1, \dots, N\}\}.$$

Remark 4.9. By Propositions 2.8 and 4.4, for any $u \in \mathcal{S}(\overline{\mathcal{P}})$ and for any $i \in \{1, \dots, N\}$, we have that $\mathcal{L}(u, E^i) \geq 0$ and that $\mathcal{L}(u, E^i) = 0$ if and only if $Du(x) \in \mathrm{extr}(F_0^i) = \partial F_0^i$ for almost every $x \in E^i$.

DEFINITION 4.10. *Letting $u, v \in \mathcal{S}(\overline{\mathcal{P}})$ we set $d(u, v) := \|u - v\|_{L^1(\Omega)}$.*

PROPOSITION 4.11. *The pair $\big(\mathcal{S}(\overline{\mathcal{P}}), d\big)$ is a complete metric space.*

Proof. We prove that $\mathcal{S}(\overline{\mathcal{P}})$ is a closed subset of $L^1(\Omega)$. For this purpose take a sequence $(u_k)_{k\in\mathbb{N}}$ in $\mathcal{S}(\overline{\mathcal{P}})$ such that $u_k \to u$ in $L^1(\Omega)$; we have to show that u is a solution of $\overline{\mathcal{P}}$.

Recalling the growth condition expressed in (A.ii) of Definition 2.3, we have

$$\int_\Omega |Du(x)|^p dx \leq M \quad \text{for any } k \in \mathbb{N}$$

for some positive constant M; hence, by the superlinear growth of ϕ, there exist a subsequence $\big(u_{k_j}\big)_{j\in\mathbb{N}}$ and $v \in W^{1,p}(\Omega, \mathbb{R})$ such that

$$u_{k_j} \rightharpoonup v \quad \text{in } W^{1,p}(\Omega, \mathbb{R}). \tag{4.5}$$

By weak lower semicontinuity of $\overline{\mathcal{I}}$ (Theorem 3.1), v belongs to $\mathcal{S}(\overline{\mathcal{P}})$. By Proposition 4.7, $Du_{k_j} = Dv$ almost everywhere in $\Omega - E$ and, for almost every $x \in E$, $Du_{k_j}(x), Dv(x) \in \big(\bigcup_{i=1}^s F_0^i\big) \subseteq B(0, R)$ for some $R > 0$; hence the sequence $\big(u_{k_j} - v\big)_{j\in\mathbb{N}}$ belongs to $W_0^{1,\infty}(\Omega, \mathbb{R})$ and $\|u_{k_j} - v\|_{W_0^{1,\infty}(\Omega,\mathbb{R})} \leq R$. Consequently we may extract a subsequence that we still call $(u_{k_j} - v)$, weakly* converging in $W_0^{1,\infty}(\Omega, \mathbb{R})$; by (4.5) its weak* limit must be zero and then, by the Rellich–Kondrachov compactness theorem we have $(u_{k_j} - v) \to 0$ in $L^\infty(\Omega, \mathbb{R})$ as $j \to \infty$. This implies $u(x) = v(x)$ for almost every $x \in \Omega$ and then u is a solution of $\overline{\mathcal{P}}$. □

PROPOSITION 4.12. *For every $\alpha > 0$, $\mathcal{S}_\alpha$ is relatively open in $\mathcal{S}(\overline{\mathcal{P}})$.*

Proof. Fix $\alpha > 0$; we prove that $\mathcal{S}(\overline{\mathcal{P}}) - \mathcal{S}_\alpha$ is closed. Let $(u_k)_{k\in\mathbb{N}}$ be a sequence in $\mathcal{S}(\overline{\mathcal{P}}) - \mathcal{S}_\alpha$ converging in $L^1(\Omega, \mathbb{R})$ to some u. Extracting if necessary a subsequence, we may assume that for some $j \in \{1, \dots, N\}$, $\mathcal{L}(u_k, E^j) \geq \alpha$ for every $k \in \mathbb{N}$. By the same argument used in the proof of Proposition 4.11 we infer the existence of a subsequence that we still call $(u_k)_{k\in\mathbb{N}}$ such that $u_k \rightharpoonup u$ in $W^{1,p}(\Omega, \mathbb{R})$. Hence u belongs to $\mathcal{S}(\overline{\mathcal{P}})$ and $\|Du_k - Du\|_{L^\infty(\Omega,\mathbb{R})} \leq R$, where $R > 0$ is the radius of a ball containing $\cup_{i=1}^N F_0^i$. Extracting if necessary a subsequence we may assume that $(Du_k - Du) \overset{*}{\rightharpoonup} 0$ in $L^\infty(\Omega, \mathbb{R})$ so that, in particular, remarking that $|Du_k|$ and $|Du|$ are bounded by R on each set E^i ($i \in \{1, \dots, N\}$), $(Du_k)\big|_{E^j} \overset{*}{\rightharpoonup} (Du)\big|_{E^j}$ in $L^\infty(E^j, \mathbb{R})$. Hence, by point (iv) of Proposition 2.8,

$$\mathcal{L}(u, E^j) \geq \limsup_{k\to\infty} \mathcal{L}(u_k, E^j) \geq \alpha.$$

Hence u belongs to $\mathcal{S}(\overline{\mathcal{P}}) - \mathcal{S}_\alpha$ and the statement is proved. □

In the following we shall consider $\mathcal{S}(\overline{\mathcal{P}})$ endowed with the distance $d(\cdot,\cdot)$.

The following proposition is an adaptation of our case of the approximation lemma proved in [DP2].

PROPOSITION 4.13. *Let $u \in \mathcal{S}(\overline{\mathcal{P}})$, $j \in \{1,\dots,N\}$. Assume that the set*

$$H^j = \{x \in \Omega : Du(x) \in \operatorname{int}(F_0^j)\}$$

has positive measure.

Then there exists $E_0^j \subseteq H^j$ with $m_n(H^j - E_0^j) = 0$ such that, given $\epsilon > 0$, the following conditions hold:

(i) *For any $y \in E_0^j$ there exist $s_y^0 > 0$, a family $\{A_{(y,s)}, s \in (0, s_y^0)\}$ of compact subsets of E^j, and a family $\{u_{(y,s)}, s \in (0, s_y^0)\}$ of elements of $\mathcal{S}(\overline{\mathcal{P}})$, satisfying the following properties:*

$$u_{(y,s)} - u \in W_0^{1,\infty}(\Omega, \mathbb{R}) \qquad \forall s \in (0, s_y^0), \tag{4.6}$$

$$u_{(y,s)} = u \text{ a.e. in } \Omega - A_{(y,s)} \qquad \forall s \in (0, s_y^0), \tag{4.7}$$

$$\|u_{(y,s)} - u\|_{L^\infty(\Omega)} \le \epsilon \qquad \forall s \in (0, s_y^0), \tag{4.8}$$

$$Du_{(y,s)} = Du \quad \text{a.e. in } \Omega - A_{(y,s)} \qquad \forall s \in (0, s_y^0), \tag{4.9}$$

$$Du_{(y,s)} \in \operatorname{extr}(K) \quad \text{a.e. in } A_{(y,s)} \qquad \forall s \in (0, s_y^0), \tag{4.10}$$

$$\int_\Omega Du_{(y,s)}(x)dx = \int_\Omega Du(x)dx \qquad \forall s \in (0, s_y^0). \tag{4.11}$$

(ii) *The family $\{A_{(y,s)}, s \in (0, s_y^0), y \in A_0\}$ is a Vitali covering of E_0^j.*

Proof. First, we suppress the dependence on j and set

$$E_0 := \{y \in H : u \text{ is differentiable at } y\}. \tag{4.12}$$

By Theorem 1, p. 235, of [EG], $m_n(H - E_0) = 0$. We fix $y \in E_0$ and proceed to the construction of $A_{(y,s)}$ and of $u_{(y,s)}$; for the sake of simplicity we omit the dependence on y writing s^0, A_s, and u_s instead of s_y^0, $A_{(y,s)}$, and $u_{(y,s)}$, respectively.

Let $\rho > 0$ be such that $B(y,\rho) \subseteq \Omega$. Since $Du(y) \in \operatorname{int}(F_0)$, by Carathéodory's theorem (see [R, p. 155]) there exists a set

$$\{v_j, j = 0, 1, \dots, n\} \subseteq \operatorname{extr}(F_0) \tag{4.13}$$

such that $Du(y) \in \operatorname{int}(\operatorname{co}(\{v_j, j = 0, 1, \dots, n\}))$.

Define the set

$$P_1 := \{z \in \mathbb{R}^n : \langle v_j - Du(y), z\rangle \le 1, j = 0, \dots, n\}$$

and, for any $r > 0$, set $P_r = rP_1 := \{w \in \mathbb{R}^n : w = rz, z \in P_1\}$. We notice that

$$\begin{aligned} P_r &= \{z \in \mathbb{R}^n : \langle v_j - Du(y), z\rangle \le r, j = 0, \dots, n\}, \\ \partial P_r &= \{z \in \mathbb{R}^n : \langle v_j - Du(y), z\rangle = r \text{ for some } j = 0, \dots, n\}. \end{aligned} \tag{4.14}$$

It is easy to see that, for any $r > 0$, P_r is a compact neighborhood of the origin, and, in particular, there exist two positive numbers d, D such that

$$0 < d < 1 < D \qquad \text{and} \qquad B(0,d) \subseteq P_1 \subseteq B(0,D). \tag{4.15}$$

Now, for $s > 0$, we define the map $w_s : \mathbb{R}^n \to \mathbb{R}$ by setting

$$w_s(x) := \max_{0\le j\le n} \{\langle v_j - Du(y), x - y\rangle\} - s. \tag{4.16}$$

It is easy to check that w_s is a piecewise affine continuous map belonging to $W^{1,\infty}_{\mathrm{loc}}(\mathbb{R}^n, \mathbb{R})$; moreover, recalling (4.13) and (4.14) we have that

$$\begin{aligned} &w_s(x) \le 0 \quad \text{if and only if} \quad x \in y + P_s, \\ &w_s(x) = 0 \quad \text{if and only if} \quad x \in y + \partial P_s. \end{aligned} \tag{4.17}$$

In addition

$$w_s(x) \ge s \qquad \forall x \in y + \partial P_{2s}, \tag{4.18}$$

while (see [C1])

$$Dw_s(x) \in \{v_j - Du(y), j = 0, 1, \dots, n\} \quad \text{for a.e. } x \in \mathbb{R}^n. \tag{4.19}$$

Claim 4.14. There exists $s_1 > 0$ such that, for every $s \in (0, s_1)$ and $\forall x \in y + \partial P_{2s}$,

$$u(x) < u(y) + \langle Du(y), x - y\rangle + w_s(x). \tag{4.20}$$

Proof of Claim 4.14. Since u is differentiable at y (recall (4.12)) we may choose $s_1 > 0$ in such a way that $\forall x \in B(y, 2s_1 D)$,

$$\left| \frac{u(x) - u(y) - \langle Du(y), x - y\rangle}{|x - y|} \right| \le \frac{1}{4D}.$$

Hence, for any $s \in (0, s_1)$, remarking that $y + P_{2s} \subseteq B(y, 2sD) \subseteq B(y, 2s_1 D)$,

$$u(x) - u(y) - \langle Du(y), x - y\rangle \le \frac{1}{4D}(2sD) = \frac{s}{2} \qquad \forall x \in y + P_{2s}.$$

Since, by (4.18), $w_s(x) \ge s$ for any $x \in y + \partial P_{2s}$, the claim is proved.

Now we fix a positive s^0 such that

$$s^0 < \min\left\{ s_1, \frac{\rho}{2D}, \frac{\epsilon}{4\mathrm{diam}(F_0)D} \right\}$$

and define the map $\tilde{u}_s : y + P_{2s} \to \mathbb{R}$, by

$$\tilde{u}_s(x) := \min\left\{u(x), u(y) + \langle Du(y), x - y\rangle + w_s(x)\right\}, \quad x \in y + P_{2s}. \tag{4.21}$$

We set also

$$A_s := \left\{x \in y + P_{2s} : \tilde{u}_s(x) = u(y) + \langle Du(y), x - y\rangle + w_s(x)\right\}. \tag{4.22}$$

First, we notice that, by Proposition 2.1, $\tilde{u}_s$ belongs to $W^{1,p}(\Omega, \mathbb{R})$. We recall that $y + P_{2s} \subseteq B(y, 2sD)$, that $s < s^0$, and that $B(y, 2s^0 D) \subseteq B(y, \rho) \subseteq \Omega$; hence $y + \mathrm{int}(P_{2s}) \subseteq \Omega$ and, by Proposition 2.1, $\tilde{u}_s$ belongs to $W^{1,p}((y + P_{2s}), \mathbb{R})$. Then it is continuous and, consequently, the set A_s turns out to be a compact subset of $y + P_{2s}$. Moreover

$$\partial A_s \subseteq \{x \in y + P_{2s} : \tilde{u}_s(x) = u(x)\}$$

and

$$A_s \subseteq y + \mathrm{int}(P_{2s}). \tag{4.23}$$

Indeed, assuming by contradiction the existence of some $\overline{x} \in A_s \cap (y + \partial P_{2s})$ we would have, by (4.20) and (4.23),

$$u(y) + \langle Du(y), \overline{x} - y\rangle + w_s(\overline{x}) = \tilde{u}(\overline{x}) < u(y) + \langle Du(y), \overline{x} - y\rangle + w_s(\overline{x}).$$

We have so proved that u and $\tilde{u}_s$ agree on the open set $\mathrm{int}(y + P_{2s}) - A_s \subseteq \Omega$; hence, by (4.21), (4.22), and (4.24),

$$\tilde{u}_s - u \in W_0^{1,p}(y + \mathrm{int}(P_{2s}), \mathbb{R}). \tag{4.24}$$

Moreover, by Proposition 2.1, we have

$$D\tilde{u}_s(x) = \begin{cases} Du(x), & x \in (y + P_{2s}) - A_s, \\ \\ Du(y) + Dw_s(x), & x \in A_s. \end{cases} \tag{4.25}$$

Now we define the map u_s by setting

$$u_s(x) := \begin{cases} \tilde{u}_s(x), & x \in (y + P_{2s}), \\ \\ u(x), & x \in \Omega - (y + P_{2s}). \end{cases} \tag{4.26}$$

Claim 4.15. The families $\{A_s, s \in (0, s^0)\}$ and $\{u_s, s \in (0, s^0)\}$ satisfy (4.6)–(4.11).

Proof of Claim 4.15. By (4.23), (4.24), and (4.26), prolonging the map $\tilde{u}_s - u$ by zero outside $(y + P_{2s})$ we may write $u_s = u + (u_s - u)\chi_{(y+P_{2s})}$. Consequently, u_s belongs to $W^{1,p}(\Omega, \mathbb{R})$ and

$$u_s - u \in W_0^{1,p}(\Omega, \mathbb{R}). \tag{4.27}$$

Hence u_s belongs to $\mathcal{S}(\overline{\mathcal{P}})$ and (4.7) follows easily from (4.21), (4.23), and (4.26).

Moreover, by Proposition 2.1, we have

$$Du_s(x) := \begin{cases} D\tilde{u}_s(x), & x \in (y + P_{2s}), \\ \\ Du(x), & x \in \Omega - (y + P_{2s}). \end{cases} \tag{4.28}$$

Hence (4.9) holds trivially, while (4.19), (4.25), and (4.28) imply (4.10).

By (4.27) and (4.9), applying the divergence theorem, we have

$$\int_\Omega (Du_s(x) - Du(x))\, dx = \int_{A_s} (Du_s(x) - Du(x))\, dx = 0,$$

i.e., (4.11). This last fact implies that u_s belongs to $\mathcal{S}(\overline{\mathcal{P}})$; indeed recalling (4.25), (4.28), and that g^{**} is affine on F_0, we have, by Jensen's inequality,

$$\begin{aligned} \frac{1}{m_n(A_s)} \int_{A_s} g^{**}(Du_s(x))dx &= g^{**}\left(\frac{1}{m_n(A_s)} \int_{A_s} Du_s(x)dx\right) \\ &= g^{**}\left(\frac{1}{m_n(A_s)} \int_{A_s} Du(x)dx\right) \le \left(\frac{1}{m_n(A_s)} \int_{A_s} g^{**}(Du(x))dx\right). \end{aligned}$$

Hence

$$\begin{aligned}\overline{\mathcal{I}}(u_s) &= \int_{\Omega - A_s} g^{**}(Du_s(x))dx + \int_{A_s} g^{**}(Du_s(x))dx \\ &\leq \int_{\Omega - A_s} g^{**}(Du(x))dx + \int_{A_s} g^{**}(Du(x))dx = m.\end{aligned}$$

Since u_s belongs to $\mathcal{S}(\overline{\mathcal{P}})$ and $Du_s(x) \in F_0$ for almost every $x \in A_s$ it follows that $A_s \subseteq E$. Hence we may infer that $u_s - u$ belongs to $W_0^{1,\infty}(\Omega, \mathbb{R})$ and that $|Du_s(x) - Du(x)| \leq 2\text{diam}(F_0)$ for almost every $x \in \Omega$. Then, recalling that $Du_s(x) - Du(x) \neq 0$ only for $x \in y + P_{2s} \subseteq B(y, 2sD)$, we have, for every $s \in (0, s^0)$,

$$|u_s(x) - u(x)| \leq 4\text{diam}(F_0)sD \leq \epsilon.$$

Hence (4.6) and (4.8) are proved.

Claim 4.16. Point (ii) of Proposition 4.13 holds true.

Proof of Claim 4.16. First note that, given $\eta > 0$, we have that

$$w_s(x) \geq \eta s - s \qquad \forall x \in y + P_{\eta s}.$$

Choose η_0 sufficiently small so that

$$\left| \frac{u(x) - u(y) - \langle Du(y), x - y \rangle}{|x - y|} \right| \leq 1$$

for any $x \in B(y, \eta_0 sD)$. Then fix $\eta < \min\{\eta_0, \frac{1}{2+2D}\}$; we have, for any $x \in y + P_{\eta s} \subseteq B(y, \eta sD) \subseteq B(y, \eta_0 sD)$,

$$u(x) - u(y) - \langle Du(y), x - y \rangle - w_s(x) \geq -\eta sD - \eta s + s \geq \frac{s}{2}.$$

This implies that any $x \in y + P_{\eta s}$ belongs to A_s; hence,

$$B(y, \eta sd) \subseteq y + P_{\eta s} \subseteq A_s \subseteq y + P_{2s} \subseteq B(y, 2sD) \qquad \forall s \in (0, s^0). \tag{4.29}$$

Now we recall that the construction depends on $y \in E_0$ and set

$$A_{(y,s)} := A_s \qquad \text{and} \qquad u_{(y,s)} := u_s.$$

In general (4.29) takes the form

$$B(y, \eta_y s d_y) \subseteq y + P_{\eta_y s} \subseteq A_{(y,s)} y + P_{2s} \subseteq B(y, 2sD_y) \qquad \forall s \in (0, s_y^0), \tag{4.30}$$

where η_y, d_y, D_y, and s_y^0 are suitable positive numbers satisfying the same requirements of η, d, D, and s^0. Now we observe that for any $y \in E_0$, $A_{(y,s)}$ is a compact subset of Ω and that, by (4.30),

$$\frac{m_n(A_{(y,s)})}{m_n(B(y, 2sD))} \geq \frac{m_n(B(y, \eta_y s d_y))}{m_n(B(y, 2sD))} \geq \left(\frac{\eta_y d_y}{2D_y} \right)^n \qquad \forall s \in (0, s_y^0).$$

This inequality, together with (4.30), shows that $\{A_{(y,s)}, s \in (0, s_y^0), y \in E_0\}$ is a Vitali covering of E_0. □

PROPOSITION 4.17. *For any $\alpha > 0$ the set $\mathcal{S}_\alpha$ is dense in $\mathcal{S}(\overline{\mathcal{P}})$.*

Proof. Let $\epsilon > 0$ and $u \in \mathcal{S}(\overline{\mathcal{P}})$ be given. We want to construct $v \in \mathcal{S}_\alpha$ such that $\|u - v\|_{L^1(\Omega,\mathbb{R})} \leq \epsilon$. Suppose first that, for any $i \in \{1, \dots, N\}$,

$$m_n\left(\{x \in E^i : Du(x) \in \mathrm{int}(F_0^i)\}\right) = 0. \tag{4.31}$$

Recalling the definition of E^i and point (iv) of Proposition 2.6, we imply that $Du \in \mathrm{extr}(F_0^i)$ almost everywhere in E^i and then, by point (ii) of Proposition 2.8, $\mathcal{L}(u, E^i) = 0$ for any $i \in \{1, \dots, N\}$. In this case we set $v = u$ and there is nothing to prove.

Suppose now that (4.31) does not hold. Renumbering the indices, we may assume that for some $P \leq N$,

$$m_n\left(\{x \in E^i : Du(x) \in \mathrm{int}F_0^i\}\right) > 0 \quad \text{for any } i \in \{1, \dots, P\} \tag{4.32}$$

and that

$$m_n\left(\{x \in \Omega : Du(x) \in \mathrm{int}F_0^i\}\right) = 0 \quad \text{for any } i \in \{P+1, \dots, N\}.$$

This last inequality implies, as above, that $\mathcal{L}(u, E^i) = 0$ for any $i \in \{P+1, \dots, N\}$. We consider then the other indices. Set

$$H^i := \{x \in E^i : Du(x) \in \mathrm{int}F_0^i\}, \qquad i \in \{1, \dots, P\}.$$

For any $i \in \{1, \dots, P\}$ we apply Proposition 4.13, with $\frac{\epsilon}{m_n(\Omega)}$ in place of ϵ, obtaining sets $E_0^i \subseteq H^i$ such that

$$m_n(H^i - E_0^i) = 0 \tag{4.33}$$

and families

$$\left\{A^i_{(y,s)}, s \in (0, s_y^0), y \in E_0^i, \right\}, \qquad i \in \{1, \dots, P\},$$

$$\left\{u^i_{(y,s)}, s \in (0, s_y^0), y \in E_0^i, \right\}, \qquad i \in \{1, \dots, P\},$$

satisfying (4.6)–(4.11), so that, in particular,

$$\|u^i_{(y,s)} - u\|_{L^\infty}(\Omega) \leq \frac{\epsilon}{m_n(\Omega)}. \tag{4.34}$$

By the Vitali covering theorem there exists a finite subfamily

$$\left\{A^i_{(y_j,s_j)}, s \in (0, s^0_{y_j}), y_j \in E_0^i, j \in \{1, \dots, \ell_i\}, \right\}, \qquad i \in \{1, \dots, P\},$$

such that

$$A^i_{(y_j,s_j)} \bigcap A^i_{(y_k,s_k)} = \emptyset \qquad \forall j, k \in \{1, \dots, \ell_i\}, \quad j \neq k, \tag{4.35}$$

and

$$m_n\left(E_0^i - G^i\right) \leq \frac{\alpha}{2\mathrm{diam}(F_0^i)}, \tag{4.36}$$

where we have set

$$G^i := \bigcup_{j=1}^{\ell_i} A^i_{(y_j,s_j)}. \tag{4.37}$$

Claim 4.18. Let $i, m \in \{1, \dots, P\}$, $i \neq m$. Then

$$A^i_{(y_j,s_j)} \bigcap A^m_{(y_k,s_k)} = \emptyset \qquad \forall j \in \{1, \dots, \ell_i\} \quad \forall k \in \{1, \dots, \ell_m\}. \tag{4.38}$$

Proof of Claim 4.18. Keeping the index i fixed, we define

$$v^i(x) := \begin{cases} u(x) & \text{for } x \in \Omega - G^i, \\ u^i_{(y_j,s_j)}(x) & \text{for } x \in A^i_{(y_j,s_j)} \text{ for some index } j \in \{1, \dots, \ell_i\}. \end{cases}$$

We have to show that this definition is well posed, i.e., that v^i belongs to $W^{1,p}(\Omega, \mathbb{R})$. By (4.35), recalling the compactness of the sets $A^i_{(y,s)}$, we may find open disjoint sets Λ^i_j, $j \in \{1, \dots, \ell_i\}$ such that

$$A^i_{(y_j,s_j)} \subseteq \Lambda^i_j \subseteq \overline{\Lambda^i_j} \subseteq \Omega.$$

Recalling (4.6) and (4.7), we have

$$W^{1,p}(\Omega, \mathbb{R}) \ni u^i_{(y_j,s_j)} = u + (u^i_{(y_j,s_j)} - u) = u + (u^i_{(y_j,s_j)} - u)\chi_{\Lambda^i_j};$$

hence $(u^i_{(y_j,s_j)} - u)\chi_{\Lambda^i_j} \in W^{1,p}(\Omega, \mathbb{R})$ and, consequently,

$$v^i = u + \sum_{j=1}^{\ell_i} (u^i_{(y_j,s_j)} - u)\chi_{\Lambda^i_j}$$

belongs to $W^{1,p}(\Omega, \mathbb{R})$. Moreover, by (4.6) and (4.35), $v^i - u \in W^{1,\infty}_0(\Omega, \mathbb{R})$ and then v^i belongs to $\mathcal{W}$.

Recalling (4.28), we have

$$Dv^i(x) := \begin{cases} Du(x) & \text{for } x \in \Omega - G^i, \\ Du^i_{(y_j,s_j)}(x) & \text{for } x \in A^i_{(y_j,s_j)} \text{ for some index } j \in \{1, \dots, \ell_i\}, \end{cases}$$

and, by (4.11),

$$\int_{A^i_{(y_j,s_j)}} Dv^i(x)dx = \int_{A^i_{(y_j,s_j)}} Du(x)dx.$$

This implies that

$$\int_{G^i} Dv^i(x)dx = \sum_{j=1}^{\ell_i} \int_{A^i_{(y_j,s_j)}} Dv^i(x)dx = \sum_{j=1}^{\ell_i} \int_{A^i_{(y_j,s_j)}} Du(x)dx = \int_{G^i} Du(x)dx. \tag{4.39}$$

By (4.10) we have that

$$Dv^i(x) \in \operatorname{extr}(F^i_0) \quad \text{for a.e. } x \in G^i; \tag{4.40}$$

recalling that g^{**} is affine on F_0^i (point (vi) of Proposition 2.6), we have, by Jensen's inequality by (4.39) and (4.40),

$$\begin{aligned}\frac{1}{m_n(G^i)}\int_{G^i} g^{**}(Dv^i(x))dx &= g^{**}\left(\frac{1}{m_n(G^i)}\int_{G^i} Dv^i(x)dx\right)\\ &= g^{**}\left(\frac{1}{m_n(G^i)}\int_{G^i} Du(x)dx\right)\\ &\le \frac{1}{m_n(G^i)}\int_{G^i} g^{**}(Du(x))dx;\end{aligned}$$

i.e.,

$$\int_{G^i} g^{**}(Dv^i(x))dx \le \int_{G^i} g^{**}(Du(x))dx.$$

This last inequality, the definition of v^i, and the convexity of g^{**} imply that

$$\overline{\mathcal{I}}(v^i) = \int_{G^i} g^{**}(Dv^i(x))dx + \int_{\Omega - G^i} g^{**}(Du(x))dx \le \int_{\Omega} g^{**}(Du(x))dx = m.$$

Hence v^i belongs to $\mathcal{S}(\overline{\mathcal{P}})$ and, by (4.40) and point (ii) of Remark 4.6,

$$G^i \subseteq E_v^i \subseteq E^i; \tag{4.41}$$

i.e., by (4.37),

$$A^i_{(y_j,s_j)} \subseteq E^i \quad \forall j \in \{1,\dots,\ell_i\}.$$

Recalling (4.4) we conclude that $A^i_{(y_j,s_j)} \cap E^m = \emptyset$ for any $j \in \{1,\dots,\ell_i\}$ and, conversely, that $A^m_{(y_k,s_k)} \cap E^i = \emptyset$ for any $k \in \{1,\dots,\ell_m\}$. Hence, again by (4.4), (4.38) holds true, and the claim is proved.

Now we define

$$v(x) := \begin{cases} u(x) & \text{for } x \in \Omega - \left(\bigcup_{i=1}^P G^i\right),\\ u^i_{(y_j,s_j)}(x) & \text{for } x \in A^i_{(y_j,s_j)}\\ & \text{for some indices } j \in \{1,\dots,\ell_i\} \text{ and } i \in \{1,\dots,P\}.\end{cases}$$

By virtue of Claim 4.18 we may apply to v the same argument used for map v^i in the proof of Claim 4.19, obtaining that v belongs to $\mathcal{S}(\overline{\mathcal{P}})$.

Claim 4.19. The map v belongs to $\mathcal{S}_\alpha$ and $\|u - v\|_{L^1(\Omega,\mathbb{R})} \le \epsilon$.

Proof of Claim 4.19. By the definition of v and by point (4.6) of Proposition 4.12 we have that $u - v$ belongs to $W_0^{1,\infty}(\Omega,\mathbb{R})$; moreover, by (4.8) and (4.34),

$$\|u^i_{(y_j,s_j)} - v\|_{L^\infty(\Omega,\mathbb{R})} \le \frac{\epsilon}{m_n(\Omega)} \quad \forall j \in \{1,\dots,\ell_i\} \quad \forall i \in \{1,\dots,P\}.$$

Hence $\|u - v\|_{L^1(\Omega,\mathbb{R})} \le \epsilon$.

We have to prove now that

$$\mathcal{L}(v, E^i) < \alpha \quad \forall i \in \{1,\dots,N\}.$$

By (4.33), recalling Remark 4.9, we have that

$$\mathcal{L}(v, E^i) = 0 \quad \forall i \in \{P+1, \dots, N\}.$$

Take then $i \in \{1, \dots, P\}$. By the definition of v, by (4.10) and (4.32), we have that

$$Dv(x) \in \operatorname{extr}(F_0^i) \quad \text{for a.e. } x \in G^i \cup (E^i - H^i);$$

hence, again by Proposition 2.8,

$$\int_{G^i} h(Dv(x), F_0^i)dx = \int_{E^i - H^i} h(Dv(x), F_0^i)dx = 0. \tag{4.42}$$

Then, by (4.33), (4.36), (4.41), and (4.42),

$$\begin{aligned}
\mathcal{L}(v, E^i) &= \int_{E^i} h(Dv(x), F_0^i)dx \\
&\le \int_{E^i - H^i} h(Dv(x), F_0^i)dx + \int_{H^i - E_0^i} h(Dv(x), F_0^i)dx \\
&\quad + \int_{E_0^i - G^i} h(Dv(x), F_0^i)dx + \int_{G^i} h(Dv(x), F_0^i)dx \\
&= \int_{E_0^i - G^i} h(Dv(x), F_0^i)dx \\
&\le \operatorname{diam}(F_0^i) m_n(E_0^i - G^i) \le \frac{\alpha}{2} < \alpha. \qquad \square
\end{aligned}$$

Proof of Theorem 3.2. By Baire's theorem and by Propositions 4.11, 4.12, and 4.13, we have that

$$\mathcal{S}_0 := \bigcap_{k \in \mathbb{N}} \mathcal{S}_{\frac{1}{k}}$$

is a dense subset of $\mathcal{S}(\overline{\mathcal{P}})$. Take an element $\overline{u} \in \mathcal{S}_0$. Clearly $\mathcal{L}(\overline{u}, E^i) = 0$ for any $i \in \{1, \dots, N\}$; hence, recalling Remark 4.9,

$$D\overline{u}(x) \in \operatorname{extr}(F_0^i) \qquad \text{for a.e. } x \in E^i. \tag{4.43}$$

Moreover, by Definition 4.5 and Proposition 4.7,

$$D\overline{u}(x) \in \mathbb{R}^n - \bigcup_{i=1}^N F_0^i \qquad \text{for a.e. } x \in \Omega - E. \tag{4.44}$$

Putting together (4.43) and (4.44) we have that

$$D\overline{u}(x) \in \mathbb{R}^n - \bigcup_{i=1}^N \operatorname{int}(F_0^i) \qquad \text{for a.e. } x \in \Omega$$

and then, by point (iv) of Proposition 2.6,

$$g(D\overline{u}(x)) = g^{**}(D\overline{u}(x)) \qquad \text{for a.e. } x \in \Omega.$$

Since $\overline{u}$ is a solution of $\overline{\mathcal{P}}$, this implies that

$$\mathcal{I}(\overline{u}) = \overline{\mathcal{I}}(\overline{u}),$$

and, recalling Theorem 3.1, we conclude that $\overline{u}$ is a solution of $\mathcal{P}$. $\square$

REFERENCES

[B] A. Bressan, *The most likely path of a differential inclusion*, J. Differ. Equations, 88 (1990), pp. 155–174.

[BF] A. Bressan and F. Flores, *On total differential inclusions*, Rend. Sem. Mat. Univ. Padova, 92 (1994), pp. 9–16.

[C1] A. Cellina, *On minima of a functional of the gradient: Necessary conditions*, Nonlinear Anal., 20 (1993), pp. 337–341.

[C2] A. Cellina, *On minima of a functional of the gradient: Sufficient conditions*, Nonlinear Anal., 20 (1993), pp. 343–347.

[D] B. Dacorogna, *Direct Methods in the Calculus of Variations*, Springer-Verlag, Berlin, 1989.

[DM1] B. Dacorogna and P. Marcellini, *General existence theorems for Hamilton-Jacobi equations in the scalar and vectorial cases*, Acta Math., 178 (1997), pp. 1–37.

[DM2] B. Dacorogna and P. Marcellini, *Cauchy-Dirichlet problem for first order nonlinear systems*, J. Funct. Anal., 152 (1998), pp. 404–446.

[DP1] F. S. De Blasi and G. Pianigiani, *Non convex valued differential inclusions in Banach spaces*, Funkcial. Ekvac., 25 (1982), pp. 153–162.

[DP2] F. S. De Blasi and G. Pianigiani, *On the Dirichlet problem for first order differential equations. A Baire category approach*, Nonlinear Differential Equations Appl., 6 (1999), pp. 13–34.

[EG] L. C. Evans and R. F. Gariepy, *Measure Theory and Fine Properties of Functions*, CRC Press, Boca Raton, FL, 1992.

[ET] I. Ekeland and R. Temam, *Convex Analysis and Variational Problems*, North–Holland, Amsterdam, 1976.

[F] J. Foran, *Fundamentals of Real Analysis*, Dekker, New York, 1991.

[KS] D. Kinderleher and G. Stampacchia, *An Introduction to Variational Inequalities and Their Applications*, Academic Press, New York, 1980.

[R] R. T. Rockafellar, *Convex Analysis*, Princeton University Press, Princeton, NJ, 1972.

[WZ] R. L. Wheeden and A. Zygmund, *Measure and Integral. An Introduction to Real Analysis*, Dekker, New York, 1977.

[Z] S. Zagatti, *On the Dirichlet problem for vectorial Hamilton–Jacobi equations*, SIAM J. Math. Anal., 29 (1998), pp. 1481–1491.

SIAM J. CONTROL OPTIM.
Vol. 38, No. 2, pp. 400–430

SECOND ORDER HAMILTON–JACOBI EQUATIONS IN HILBERT SPACES AND STOCHASTIC BOUNDARY CONTROL*

FAUSTO GOZZI†, ELISABETH ROUY‡, AND ANDRZEJ ŚWIĘCH§

Abstract. The paper is concerned with fully nonlinear second order Hamilton–Jacobi–Bellman–Isaacs equations of elliptic type in separable Hilbert spaces which have unbounded first and second order terms. The viscosity solution approach is adapted to the equations under consideration and the existence and uniqueness of viscosity solutions are proved. A stochastic optimal control problem driven by a parabolic stochastic PDE with control of Dirichlet type on the boundary is considered. It is proved that the value function of this problem is the unique viscosity solution of the associated Hamilton–Jacobi–Bellman equation.

Key words. stochastic boundary control, Hamilton–Jacobi equations, viscosity solutions

AMS subject classifications. 49L25, 93E20, 35R15

PII. S0363012997324909

1. Introduction.

In this paper we study second order infinite dimensional Hamilton–Jacobi–Bellman–Isaacs equations

$$\lambda v(x) + \big\langle Ax, Dv(x)\big\rangle + H\big(x, Dv(x), D^2v(x)\big) = 0, \qquad x \in X, \tag{1.1}$$

and their relationship with stochastic optimal control problems. Above, X is a real, separable Hilbert space, D denotes the Fréchet derivative, λ is a positive number, and $-A : D(A) \subset X \to X$ is a closed linear operator that generates an analytic C_0–semigroup e^{-tA} on X. Moreover, we assume that A is positive and self-adjoint and has compact resolvent $R(\mu, A)$, and that $H : X_1 \to \mathbb{R}$, where $X_1 \subset X \times X \times \Sigma(X)$ (Σ denotes the space of all bounded, self-adjoint linear operators from X to itself). X_1 will be specified later. We call such equations *unbounded*.

Equations of this type arise in stochastic optimal control problems driven by parabolic stochastic PDE, for instance, when the control is given at the boundary with Dirichlet or Neumann type conditions (see section 5.4) and the stochastic term is given by the so-called "white noise." This has been one of the main motivations of our study. Our approach is very flexible since the model control problem we study in section 5 includes both distributed and boundary controls and applies also to cases with purely distributed controls (infinite or finite dimensional). In this paper we define a suitable notion of solution of (1.1), prove existence and uniqueness of solutions, and show that if (1.1) comes from a stochastic optimal control problem the value function is its unique solution. These results are general and apply also to equations with no

*Received by the editors July 23, 1997; accepted for publication (in revised form) August 21, 1998; published electronically January 11, 2000. Most of this research was completed when the authors visited Scuola Normale Superiore, Pisa, Italy, and when the first author visited Georgia Institute of Technology, Atlanta, GA.

http://www.siam.org/journals/sicon/38-2/32490.html

†Dipartimento di Matematica, Università di Pisa, Via F. Buonarroti, 56127 Pisa, Italy (gozzi@sab.sns.it). The reseach of this author was supported in part by the Italian National Project MURST 40% "Problemi nonlineari...."

‡Institut Elie Cartan, INRIA Lorraine, CNRS UMR 9973, B.P. 239, 54506 Vandoeuvre-lès-Nancy Cedex, France. Present address: Laboratorie de Mathématiques et Physique Théorique, UPRES-A 6083, Université de Tours, Parc de Grandmont, 37200 Tours, France.

§School of Mathematics, Georgia Institute of Technology, Atlanta, GA 30332.

"sup" structure, in particular to Isaacs' equations. Therefore, instead of studying a boundary control problem in section 5 we could study a differential game problem.

To have an idea of what we have in mind let us look at the following stochastic optimal control problem (P_0) that we will study in detail in section 5.4. Let $\Omega \subset \mathbb{R}^n$ be an open, connected, and bounded set with smooth boundary. Consider the stochastic controlled PDE

$$\begin{cases} \dfrac{\partial x}{\partial t}(t,\xi) = \Delta_\xi x(t,\xi) + f_1\big(x(t,\xi),\alpha_1(t,\xi)\big) \\ \qquad\qquad + f_2\big(x(t,\xi),\alpha_1(t,\xi)\big)\dot{W}_Q(t,\xi) & \text{in } (0,\infty)\times\Omega, \\ x(0,\xi) = x_0(\xi) & \text{on } \Omega, \\ x(t,\xi) = \alpha_2(t,\xi) & \text{on } (0,\infty)\times\partial\Omega, \end{cases} \tag{1.2}$$

where W_Q is a Wiener process with values in $L^2(\Omega)$ and with covariance operator Q, $x_0 \in L^2(\Omega)$, and the controls $\alpha_1 : (0,\infty) \to L^2(\Omega)$, $\alpha_2 : (0,\infty) \to L^2(\partial\Omega)$ are measurable and adapted to the Wiener process W_Q, and $f_1, f_2 : \mathbb{R}^2 \to \mathbb{R}$. In this equation we have two control functions. The first, indicated by α_1, is the "distributed" control, while the other, indicated by α_2, is the "boundary" control. The Cauchy problem (1.2) may be rewritten in an abstract form as explained in section 5.

Consider now the problem of minimizing the cost functional

$$J(x_0;\alpha_1,\alpha_2) = \mathbb{E}\int_0^{+\infty} e^{-\lambda t} L\big(x(t;x_0,\alpha_1,\alpha_2),\alpha_1(t),\alpha_2(t)\big)\,dt,$$

where $\lambda > 0$, $L : L^2(\Omega)\times L^2(\Omega)\times L^2(\partial\Omega) \to \mathbb{R}$ is a bounded, uniformly continuous function and $x(\cdot\,;x_0,\alpha)$ is the solution of the Cauchy problem (1.2). The value function of this problem is defined as

$$v(x) = \inf J(x;\alpha_1,\alpha_2),$$

where the infimum is taken over all admissible controls (α_1,α_2) considered above. We expect that the function v is a solution (in a suitable sense) of (1.1), where the Hamiltonian H depends on the data of the problem (see section 5.3). We point out here that the presence of the boundary control term in the state equation (1.2) causes $H(x,Dv(x),D^2v(x))$ to be defined only if $Dv(x) \in D(A^\beta)$, where A is the Laplace operator with Dirichlet boundary conditions and $\beta \in (\frac{3}{4},1)$. Clearly, in the case of different boundary conditions, different values of β have to be considered; for example, the case of Neumann boundary conditions (which also fits in our assumptions) gives $\beta \in (\frac{1}{4},1)$. This "bad behavior" appears as a result of "transforming" the boundary control into a distributed one, as it is explained, e.g., in Bensoussan et al. [3] and Cannarsa and Tessitore [10].

To deal with the difficulties posed by this problem we introduce a change of variables which is similar to the one used by Cannarsa and Tessitore in [10] to study a first order Hamilton–Jacobi equation associated with a boundary control problem with Dirichlet boundary conditions. Given a solution $x(\cdot)$ of the state equation (1.2), we set $y(\cdot) = A^{-\frac{\beta}{2}}x(\cdot)$ for a chosen $\beta \in (0,1)$, and we study a new control problem (P_1) with cost $J(A^{\frac{\beta}{2}}y,\alpha)$ and state $y(\cdot)$. We then obtain a different Hamilton–Jacobi–Bellman equation,

$$\lambda u(y) + \langle Ay, Du(y)\rangle + H\left(A^{\frac{\beta}{2}}y, A^{-\frac{\beta}{2}}Du(y), A^{-\frac{\beta}{2}}D^2u(y)A^{-\frac{\beta}{2}}\right) = 0, \tag{1.3}$$

which can also be obtained directly from (1.1) by making the change of variable $y = A^{-\frac{\beta}{2}}x$ and setting $u(y) = v(A^{\frac{\beta}{2}}y)$. This equation contains fewer unbounded terms and is easier to handle in spite of the additional difficulty created by the presence of the unbounded term $A^{\frac{\beta}{2}}y$. For (1.3) we are able to prove existence and uniqueness of viscosity solutions. Then, since (1.1) and (1.3) are related by the change of variable $u(y) = v(A^{\frac{\beta}{2}}y)$, we define a viscosity solution of (1.1) as a function v such that $u(\cdot) \overset{def}{=} v(A^{\frac{\beta}{2}}\cdot)$ is a viscosity solution of (1.3). The function v is uniquely determined once u has been characterized on $D(A^{\frac{\beta}{2}})$. This is the idea behind the definition of viscosity solution we employ and we will make it rigorous in section 3. The definition is meaningful since in the case when (1.1) comes from a stochastic control problem, v and u can be respectively characterized as the value functions of problems (P_0) and (P_1).

The techniques of this paper could also be employed to study evolution equations of parabolic type and stochastic boundary control problems with finite horizon (see Cannarsa and Tessitore [9] for an analogue in the first order case). Moreover, some of the assumptions made throughout the paper can be relaxed but we do not attempt to do so here.

Finally we observe that the problems of optimality conditions and synthesis of optimal controls for stochastic control problems in infinite dimensions when the value function is not regular are open. Even in the case of finite dimensional stochastic control problems very few results are available (see [37]). These questions will be studied in the future. We briefly discuss what is known in section 5.5.

There is an increasing interest in and a growing literature on Hamilton–Jacobi equations in infinite dimensions. These equations were first studied by Barbu and Da Prato (see, e.g., [2]), setting the problem in classes of convex functions and using semigroup and perturbation methods (see also Da Prato [15] and Havarneanu [28]). Much progress has been made recently due to the introduction of the notion of viscosity solutions. We refer the reader to Crandall and Lions [14], Ishii [31], Soner [42], and Tataru [44, 45, 46] for the first order equations. As regards the second order, "bounded" equations have been investigated by P.-L. Lions in [39, Parts I and III], and "unbounded" in [39, Part II], Ishii [32], Kocan and Święch [34], and Święch [43]. Except for [32] the unboundedness in the studied equations was always coming from the term $\langle Ax, Du\rangle$. This paper is concerned with equations that exhibit "bad behavior" in the Hamiltonian H also in Du and D^2u.

To compare our work with the existing literature let us look at the following model equation:

$$\lambda v(x) - \frac{1}{2}\operatorname{Tr} QD^2v(x) + \big\langle Ax, Dv(x)\big\rangle + H\big(x, Dv(x)\big) = 0, \quad x \in X, \tag{1.4}$$

where $Q : X \to X$ is a self-adjoint nonnegative linear operator. Leaving aside the unboundedness of the term $\langle Ax, Dv(x)\rangle$ that has been investigated in various cases, let us concentrate our attention on the other terms. If Q is a nuclear operator, the term $\operatorname{Tr} QC$ is well defined for all $C \in \Sigma(X)$, and (1.4) is well studied. If Q is not nuclear, $\operatorname{Tr} QC$ does not make sense for many operators, notably for $C = I$. However, equations like this arise in stochastic optimal control of infinite dimensional systems driven by "nondegenerate processes," in particular by the so-called "white noise" (see Albeverio and Röckner [1], and Jona Lasinio and Mitter [33]). Concerning the "bad behavior" with respect to Dv, in the case of Bellman equations (1.4) associated with stochastic boundary control problems, H is well defined only on $X \times D(A^\beta)$

(where $D(A^\beta)$ denotes the domain of the β-fractional power of A for a suitable $\beta \in (0,1)$), and hence the term $H(x, Dv(x))$ is not well defined even if v is Fréchet differentiable. To our knowledge equations of this kind have been studied only in the first order case (i.e., when $Q = 0$) by Cannarsa, Gozzi and Soner [8] and Cannarsa and Tessitore [10].

In the case when Q is not nuclear and H is continuous in $X \times X$, Hamilton–Jacobi equations (1.4) and their parabolic analogues have been recently studied by Cannarsa and Da Prato [4, 5, 7], Gozzi [25, 26], and Gozzi and Rouy [27] from the point of view of strong solutions (which are in particular differentiable in the space variable x). The theory is based on stochastic representation of solutions and uses techniques related to properties of transition semigroups in infinite dimensions recently developed by Da Prato and Zabczyk [17, 18]. Another approach to second order Hamilton–Jacobi equations in infinite dimensions is presented by Chow and Menaldi in [11].

The plan of the paper is the following. In section 2 we give some preliminaries. In section 3 and section 4 we present the definition of a viscosity solution and we prove a general uniqueness and existence result for the transformed equation (1.3) and for the original one (1.1), respectively. Section 5 is devoted to the control problem (P_0). In sections 5.1 and 5.2 we introduce the problem and we prove various estimates for the solutions of the transformed state equation. In section 5.3 we prove that the value function of the control problem (P_0) is the unique viscosity solution of the associated Hamilton–Jacobi–Bellman equation, while in section 5.4 we present examples of stochastic boundary control problems. Finally, in section 5.5 we briefly discuss the relationship between the notions of viscosity and strong solutions and the problem of synthesis of optimal controls.

We refer the reader to the survey paper of Crandall, Ishii, and Lions [12] for the introduction to the notion of viscosity solutions and a complete treatment of finite dimensional equations and to the book of Fleming and Soner [24] for the connection with stochastic optimal control.

2. Notation and preliminaries. Throughout this paper X will denote a real separable Hilbert space endowed with the inner product $\langle \cdot, \cdot \rangle$ and the norm $|\cdot|$. We denote by $\mathcal{L}(X)$ the Banach space of the continuous linear operators $T : X \to X$ with the operator norm $\|\cdot\|$, and we set

$$\Sigma(X) = \big\{T \in \mathcal{L}(X),\ T \text{ self-adjoint}\big\}.$$

Moreover, we denote by $\mathcal{L}_2(X)$ the set of all Hilbert–Schmidt operators $T : X \to X$. It is well known that $\mathcal{L}_2(X)$ is a Hilbert space with the norm

$$\|T\|^2_{\mathcal{L}_2(X)} = \sum_{k=1}^{\infty} |Te_k|^2,$$

where $\{e_k\}_{k\in\mathbb{N}}$ is any orthonormal basis of X.

For any Hilbert spaces X and Y, we denote by $B(X,Y)$, $UC(X,Y)$, and $BUC(X,Y)$ the Banach space of all functions $\varphi : X \to Y$ which are, respectively, bounded, uniformly continuous, uniformly continuous and bounded on X with the usual norm

$$\|\varphi\| = \sup_{x\in X} |\varphi(x)|_Y.$$

For $k \in \mathbb{N}$, we denote by $BUC^k(X,Y)$ the set of all functions $\varphi : X \to Y$ which are uniformly continuous and bounded on X together with all their Fréchet derivatives up to the order k. If $Y = \mathbb{R}$, then we write $BUC^k(X)$ instead of $BUC^k(X,\mathbb{R})$.

We say that a function $\rho : [0,+\infty) \to [0,+\infty)$ is a modulus if ρ is continuous, nondecreasing, subadditive, and $\rho(0) = 0$. Subadditivity in particular implies that for all $\varepsilon > 0$, there exists $C_\varepsilon > 0$ such that

$$\rho(r) \le \varepsilon + C_\varepsilon r \quad \text{for every } r \ge 0.$$

Moreover, a function $\rho : [0,+\infty) \times [0,+\infty) \to [0,+\infty)$ is a local modulus if ρ is continuous, nondecreasing in both variables, subadditive in the first variable, and $\rho(0,r) = 0$ for every $r \ge 0$.

For any $\varphi \in UC(X)$, we denote by ρ_φ a continuity modulus of φ, i.e., a modulus such that $|\varphi(x) - \varphi(y)| \le \rho_\varphi(|x-y|)$ for every $x, y \in X$. We recall that, if $\varphi \in UC(X,Y)$, then its modulus of continuity always exists and so there exist positive constants C_0, C_1 such that

$$|\varphi(x)|_Y \le C_0 + C_1|x| \quad \text{for every } x \in X.$$

We now briefly recall some properties of the stochastic convolution. Let $(\Omega, \mathcal{F}, \mathcal{F}_t, \mathbf{P})$ be a complete probability space with a normal filtration $\{\mathcal{F}_t : t \ge 0\}$, and let W be a cylindrical Wiener process with respect to $\mathcal{F}_t$ (see [17, Chapter 4], for the definition and properties of a cylindrical Wiener process W).

Given $T \in (0,+\infty]$, let us denote by $M^0(0,T;X)$ the space of all X-valued processes measurable on $(0,T)$ and adapted to the filtration $\mathcal{F}_t$ and by $M^2(0,T;X)$ the space of the X-valued processes x such that $x \in M^0(0,T;X)$ and

$$\mathbb{E}\left(\int_0^T |x(s)|^2\, ds\right) < +\infty.$$

The result below can be found in [17, Theorem 7.6] and [18, Chapter 5].

PROPOSITION 2.1. *Let $A : D(A) \subset X \to X$ be the generator of a strongly continuous semigroup e^{tA} on X. Assume that $B : [0,T] \times X \times \Omega \to \mathcal{L}(X)$ is a strongly continuous predictable process such that $e^{tA}B(s,x,\omega)$ is a Hilbert–Schmidt operator for all $s,t \in (0,T]$, $x \in X$, $\omega \in \Omega$ and*

$$\mathbb{E}\int_0^T \left\|e^{tA}B(s,x)\right\|^2_{\mathcal{L}_2(X)} dt < +\infty$$

uniformly in $(s,x) \in [0,T] \times X$. Then, setting

$$\bar{W}(t,x) = \int_0^t e^{(t-s)A}B(s,x)\, dW(s),$$

we have that, for every $x \in X$,

(i) *the process $\bar{W}(\cdot,x)$ is mean square continuous,*
(ii) *the process $\bar{W}(\cdot,x)$ has $\mathbf{P}$-almost surely (a.s.) square integrable trajectories and $\bar{W}(\cdot,x) \in M^2(0,T;X)$,*
(iii) *if for some $\delta > 0$*

$$\mathbb{E}\int_0^T t^{-\delta}\left\|e^{tA}B(s,x)\right\|^2_{\mathcal{L}_2} dt < +\infty \tag{2.1}$$

uniformly in $(s,x) \in [0,T] \times X$, *then* $\bar{W}(\cdot,x)$ *has continuous trajectories and*

$$\mathbb{E}\left(\sup_{0\le t\le T} |\bar{W}(t,x)|^2\right) < +\infty.$$

3. Definition of viscosity solution and a general existence and uniqueness result for the transformed equation. In this section we study the "transformed" equation (1.3). Note that, for the purpose of studying (1.1), we need only to solve the transformed problem on $D(A^{\beta/2})$, but since Hamilton–Jacobi equations of this kind may be of independent interest, we investigate a general equation

$$\lambda u(y) + \langle Ay, Du(y)\rangle + G\big(y, Du(y), D^2u(y)\big) = 0, \quad y \in X. \tag{3.1}$$

We assume the following.

HYPOTHESIS 3.1. $A : D(A) \subset X \to X$ *is a self-adjoint operator, there exists* $a > 0$ *such that* $\langle Ax, x\rangle \ge a|x|^2$ *for all* $x \in D(A)$, *and* A^{-1} *is compact.*

Remark 3.2. Hypothesis 3.1 implies in particular that $-A$ is the infinitesimal generator of an analytic semigroup with compact resolvent satisfying $\|e^{-tA}\| \le e^{-at}$ for all $t \ge 0$ and that there is an orthonormal basis of X made of eigenvectors of A such that the corresponding sequence of eigenvalues diverges to $+\infty$ as $n \to \infty$. It also follows that for every $\theta \in [0,1]$ there exists a constant $M_\theta > 0$ such that

$$\big|A^\theta e^{-tA}x\big| \le M_\theta \frac{e^{-ta}}{t^\theta}|x| \quad \text{for every } t > 0,\ x \in X. \tag{3.2}$$

Moreover, for $\gamma \in (0,1]$ and $\alpha \in (0,\gamma)$, a well known interpolation inequality (see, e.g., [41, pp. 73–74]) states that for every $\sigma > 0$ there exists $C_\sigma > 0$ such that

$$|A^\alpha x| \le \sigma |A^\gamma x| + C_\sigma |x| \quad \text{for every } x \in D(A^\gamma). \tag{3.3}$$

Let $X_1 \subset X_2 \subset \cdots$ be finite dimensional subspaces of X generated by eigenvectors of A such that $\overline{\bigcup_{N=1}^\infty X_N} = X$. Given $N \in \mathbb{N}$, denote by P_N the orthogonal projection onto X_N, let $Q_N = I - P_N$, and let $X_N^\perp = Q_N X$. We then have an orthogonal decomposition $X = X_N \times X_N^\perp$ and we will denote by x_N an element of X_N and by $x_N^\perp$ an element of $X_N^\perp$. For $x \in X$ we will write $x = (P_N x, Q_N x)$. For $\gamma > 0$ we denote by $X_{-\gamma}$ the completion of X in the norm $|x|_{-\gamma} = |A^{-\frac{\gamma}{2}}x|$. We make the following assumptions about G.

HYPOTHESIS 3.3.

(A0) *There exists* $\beta \in (0,1)$ *such that the function* $G : D(A^{\frac{\beta}{2}}) \times D(A^{\frac{\beta}{2}}) \times \Sigma(X) \to \mathbb{R}$ *is continuous (in the topology of* $D(A^{\frac{\beta}{2}}) \times D(A^{\frac{\beta}{2}}) \times \Sigma(X)$*).*

(A1)
$$G(x,p,S_1) \le G(x,p,S_2) \quad \text{if } S_1 \ge S_2$$
for all $x, p \in D(A^{\frac{\beta}{2}})$.

(A2) *There exists a modulus* ρ *such that*
$$\begin{aligned}&\big|G(x,p,S_1) - G(x,q,S_2)\big| \\ &\qquad \le \rho\Big(\big(1 + \big|A^{\frac{\beta}{2}}x\big|\big)|A^{\frac{\beta}{2}}(p-q)| + \big(1 + |A^{\frac{\beta}{2}}x|^2\big)\|S_1 - S_2\|\Big)\end{aligned}$$
for all $x, p, q \in D(A^{\frac{\beta}{2}})$ *and* $S_1, S_2 \in \Sigma(X)$.

(A3) *There exist* $0 < \eta < 1 - \beta$ *and a modulus* ω *such that, for all* $N \geq 1$,

$$G\Big(x, \frac{A^{-\eta}(x-y)}{\varepsilon}, Z\Big) - G\Big(y, \frac{A^{-\eta}(x-y)}{\varepsilon}, Y\Big) \geq -\omega\Big(|A^{\frac{\beta}{2}}(x-y)|\Big(1 + \frac{|A^{\frac{\beta}{2}}(x-y)|}{\varepsilon}\Big)\Big)$$

for all $x, y \in D(A^{\frac{\beta}{2}})$ *and* $Z, Y \in \Sigma(X_N)$ *such that*

$$\begin{pmatrix} Z & 0 \\ 0 & -Y \end{pmatrix} \leq \frac{2}{\varepsilon} \begin{pmatrix} P_N A^{-\eta} P_N & -P_N A^{-\eta} P_N \\ -P_N A^{-\eta} P_N & P_N A^{-\eta} P_N \end{pmatrix}. \tag{3.4}$$

(A4) *For every* $R < +\infty, |\lambda| \leq R$, $p, x \in D(A^{\frac{\beta}{2}})$

$$\sup\big\{|G(x,p,S+\lambda Q_N) - G(x,p,S)| : \|S\| \leq R, S = P_N S P_N\big\} \to 0 \quad \text{as } N \to \infty. \tag{3.5}$$

Remark 3.4. By the properties of moduli, condition (A2) guarantees that there exists a constant C such that, for every x, p, S,

$$\big|G(x,p,S)\big| \leq C\Big(1 + \big(1 + |A^{\frac{\beta}{2}}x|\big)\big|A^{\frac{\beta}{2}}p\big| + \big(1 + |A^{\frac{\beta}{2}}x|^2\big)\|S\|\Big) + \big|G(x,0,0)\big|. \tag{3.6}$$

Some of the conditions above can be weakened; however, we want to keep the technicalities down. We refer the reader to [43] for techniques leading to possible generalizations.

The definition of viscosity solution is motivated by [10].

DEFINITION 3.5. *We say that a function* φ *belongs to the space* $\tilde{C}^2_-(X)$ *(resp.,* $\tilde{C}^2_+(X)$*) if*

(i) $\varphi \in C^2(X)$ *and is weakly sequentially lower (resp., upper) semicontinuous on* X.
(ii) $D\varphi \in UC(X,X) \cap UC(D(A^{\frac{1}{2}-\varepsilon}), D(A^{\frac{1}{2}}))$ *for some* $\varepsilon = \varepsilon(\varphi) > 0$.
(iii) $D^2\varphi \in BUC(X, \Sigma(X))$.

DEFINITION 3.6. *Given* $\delta > 0$ *and* $w, \varphi : X \to \mathbb{R}$ *we say that a point* $y_0 \in X$ *belongs to the set* $M^+_\delta(w,\varphi)$ *if* y_0 *is a point of a local maximum for the function*

$$w - \varphi - \frac{\delta}{2}|\cdot|^2.$$

Similarly, we say that y_0 *belongs to* $M^-_\delta(w,\varphi)$ *if* y_0 *is a point of a local minimum for the function*

$$w - \varphi + \frac{\delta}{2}|\cdot|^2.$$

DEFINITION 3.7. *We say that a function* $w : X \to \mathbb{R}$ *is a viscosity subsolution of* (3.1) *if* w *is weakly sequentially upper semicontinuous on* X*, and for every* $\varphi \in \tilde{C}^2_-(X)$ *and* $\delta > 0$,

(i) $M^+_\delta(w,\varphi) \subset D(A^{\frac{1}{2}})$,
(ii) *for all* $y \in M^+_\delta(w,\varphi)$,

$$\lambda w(y) + \langle A^{\frac{1}{2}} y, A^{\frac{1}{2}} D\varphi(y)\rangle + \delta |A^{\frac{1}{2}} y|^2 + G\left(y, D\varphi(y) + \delta y, D^2\varphi(y) + \delta I\right) \le 0.$$

We say that w is a viscosity supersolution of (3.1) *if w is weakly sequentially lower semicontinuous on X, and for every $\varphi \in \tilde{C}^2_+(X)$ and $\delta > 0$,*

(i) $M^-_\delta(w,\varphi) \subset D(A^{\frac{1}{2}})$,

(ii) *for all $y \in M^-_\delta(w,\varphi)$,*

$$\lambda w(y) + \langle A^{\frac{1}{2}} y, A^{\frac{1}{2}} D\varphi(y)\rangle - \delta \left|A^{\frac{1}{2}} y\right|^2 + G\left(y, D\varphi(y) - \delta y, D^2\varphi(y) - \delta I\right) \ge 0.$$

We say that w is a viscosity solution of (3.1) *if it is both a viscosity subsolution and a supersolution.*

THEOREM 3.8. *Let Hypotheses* 3.1 *and* 3.3 *be satisfied. Then we have the following.*

Comparison: Let $u, -v \le M$ for some constant M. If u is a viscosity subsolution of (3.1) *and v is a viscosity supersolution of* (3.1)*, then $u \le v$ on X. Moreover, if u is a viscosity solution then*

$$\left|u(x) - u(y)\right| \le m\left(|A^{-\frac{\eta}{2}}(x-y)|\right) \tag{3.7}$$

for all $x, y \in X$ and some modulus m, where η is the constant in (A3).

Existence: If

$$\sup_{x \in D\left(A^{\frac{\beta}{2}}\right)} \left|G(x,0,0)\right| = K < \infty, \tag{3.8}$$

then there exists a unique viscosity solution $u \in BUC(X_{-\eta})$ of (3.1).

Proof. Comparison. Let $\varepsilon, \delta > 0$. Put

$$\Phi(x,y) = u(x) - v(y) - \frac{\left|A^{-\frac{\eta}{2}}(x-y)\right|^2}{2\varepsilon} - \frac{\delta}{2}|x|^2 - \frac{\delta}{2}|y|^2.$$

Since $u - v$ is bounded from above and weakly upper-semicontinuous in $X \times X$, Φ must attain its maximum at some point $(\bar{x}, \bar{y})$ (which can be assumed to be strict by subtracting, for instance, $\mu(|A^{-1}(x-\bar{x})|^2 + |A^{-1}(y-\bar{y})|^2)$ and then letting $\mu \to 0$). Moreover, standard considerations (see, for instance, [32]) yield that

$$\lim_{\varepsilon \to 0} \limsup_{\delta \to 0} \left(\frac{|A^{-\frac{\eta}{2}}(\bar{x} - \bar{y})|^2}{\varepsilon} \right) = 0 \tag{3.9}$$

and

$$\lim_{\delta \to 0} \left(\delta |\overline{x}|^2 + \delta |\overline{y}|^2\right) = 0 \quad \text{for every fixed } \varepsilon > 0. \tag{3.10}$$

We now fix $N \in \mathbb{N}$. Then obviously

$$\langle A^{-\eta}(x-y), x-y\rangle = \langle P_N A^{-\eta} P_N (x-y), x-y\rangle + \left|A^{-\frac{\eta}{2}} Q_N (x-y)\right|^2,$$

and we have

$$\begin{aligned}\left|A^{-\frac{\eta}{2}} Q_N (x-y)\right|^2 \le{}& 2\langle Q_N A^{-\eta} Q_N(\bar{x} - \bar{y}), x - y\rangle - \langle Q_N A^{-\eta} Q_N(\bar{x} - \bar{y}), \bar{x} - \bar{y}\rangle \\ &+ 2\left|A^{-\frac{\eta}{2}} Q_N (x - \bar{x})\right|^2 + 2\left|A^{-\frac{\eta}{2}} Q_N (y - \bar{y})\right|^2\end{aligned}$$

with equality if $x = \bar{x}, y = \bar{y}$. Therefore, if we define

$$u_1(x) = u(x) - \frac{\langle x, Q_N A^{-\eta} Q_N(\bar{x} - \bar{y})\rangle}{\varepsilon} + \frac{\langle Q_N A^{-\eta} Q_N(\bar{x} - \bar{y}), \bar{x} - \bar{y}\rangle}{2\varepsilon} - \frac{\left|A^{-\frac{\eta}{2}} Q_N(x - \bar{x})\right|^2}{\varepsilon} - \frac{\delta}{2}|x|^2$$

and

$$v_1(y) = v(y) - \frac{\langle y, Q_N A^{-\eta} Q_N(\bar{x} - \bar{y})\rangle}{\varepsilon} + \frac{\left|A^{-\frac{\eta}{2}} Q_N(y - \bar{y})\right|^2}{\varepsilon} + \frac{\delta}{2}|y|^2,$$

it follows that the function

$$\widetilde{\Phi}(x, y) \stackrel{def}{=} u_1(x) - v_1(y) - \frac{\langle P_N A^{-\eta} P_N(x - y), x - y\rangle}{2\varepsilon} \tag{3.11}$$

always satisfies $\widetilde{\Phi} \leq \Phi$ and attains a strict global maximum at $\bar{x}, \bar{y}$, where $\widetilde{\Phi}(\bar{x}, \bar{y}) = \Phi(\bar{x}, \bar{y})$. We now define, for $x_N,\ y_N \in X_N$, the functions

$$\tilde{u}_1(x_N) = \sup_{x_N^\perp \in X_N^\perp} u_1\big(x_N, x_N^\perp\big), \qquad \tilde{v}_1(y_N) = \inf_{y_N^\perp \in X_N^\perp} v_1\big(y_N, y_N^\perp\big).$$

Since u_1 and $-v_1$ are weakly upper semicontinuous on X, $\tilde{u}_1$ and $-\tilde{v}_1$ are upper semicontinuous on X_N (see [13]). Moreover, by definition of u_1 and $-v_1$ and by the form of $\widetilde{\Phi}$, it follows that

$$\tilde{u}_1(P_N\bar{x}) = u_1(\bar{x}), \qquad \tilde{v}_1(P_N\bar{y}) = v_1(\bar{y}). \tag{3.12}$$

Defining now the map $\Phi_N : X_N \times X_N \to \mathbb{R}$ as

$$\begin{aligned}\Phi_N(x_N, y_N) &= \tilde{u}_1(x_N) - \tilde{v}_1(y_N) - \frac{\langle P_N A^{-\eta} P_N(x_N - y_N), x_N - y_N\rangle}{2\varepsilon} \\ &= \sup_{x_N^\perp,\, y_N^\perp \in X_N^\perp} \widetilde{\Phi}\left(\big(x_N, x_N^\perp\big), \big(y_N, y_N^\perp\big)\right),\end{aligned}$$

it is not difficult to check that Φ_N attains a strict global maximum over $X_N \times X_N$ at $(\bar{x}_N, \bar{y}_N) = (P_N\bar{x}, P_N\bar{y})$. By a finite dimensional result (see [12]) for every $n \in \mathbb{N}$ there exist points $x_N^n,\ y_N^n \in X_N$ such that

$$x_N^n \to \bar{x}_N, y_N^n \to \bar{y}_N, \qquad \tilde{u}_1(x_N^n) \to \tilde{u}_1(\bar{x}_N), \qquad \tilde{v}_1(y_N^n) \to \tilde{v}_1(\bar{y}_N) \quad \text{as } n \to \infty \tag{3.13}$$

and there exist functions $\varphi_n, \psi_n \in C^2(X_N)$ such that $\tilde{u}_1 - \varphi_n$, and $-\tilde{v}_1 + \psi_n$ have unique, strict, global maxima at x_N^n, and y_N^n respectively, and

$$\begin{aligned} D\varphi_n(x_N^n) &\to \frac{1}{\varepsilon} P_N A^{-\eta} P_N(\bar{x}_N - \bar{y}_N), \\ D\psi_n(y_N^n) &\to \frac{1}{\varepsilon} P_N A^{-\eta} P_N(\bar{x}_N - \bar{y}_N), \end{aligned} \tag{3.14}$$

$$D^2\varphi_n(x_N^n) \to Z_N, \qquad D^2\psi_n(y_N^n) \to Y_N, \tag{3.15}$$

where Z_N, Y_N satisfy (3.4). Consider finally the map $\Phi_N^n : X \times X \to \mathbb{R}$ defined as

$$\Phi_N^n(x,y) = u_1(x) - v_1(y) - \varphi_n(P_N x) + \psi_n(P_N y). \tag{3.16}$$

This map has the variables split and, by the definition of u_1 and v_1, attains its global maximum (which we can assume to be strict) at some point $(\hat{x}^n, \hat{y}^n)$. This point depends also on N but we will drop this dependence since N is now fixed. Setting now

$$\overline{\varphi}_{N,n}(x) \stackrel{def}{=} \frac{\langle x, Q_N A^{-\eta} Q_N(\bar{x}-\bar{y})\rangle}{\varepsilon} + \frac{\left|A^{-\frac{\eta}{2}} Q_N(x-\bar{x})\right|^2}{\varepsilon} + \varphi_n(P_N x),$$

we easily see that $\overline{\varphi}_{N,n} \in \tilde{C}^2_-(X)$ and we have from (3.16) that $u(x) - \overline{\varphi}_{N,n}(x) - \frac{\delta}{2}|x|^2$ has a maximum at $\hat{x}^n$. Therefore, by the definition of viscosity subsolution, $\hat{x}^n \in D(A^{\frac{1}{2}})$ and

$$\begin{aligned}
&\lambda u(\hat{x}^n) + \left\langle A^{\frac{1}{2}}\hat{x}^n, A^{\frac{1}{2}} D\varphi_n(P_N\hat{x}^n) + \frac{A^{\frac{1}{2}-\eta} Q_N(\bar{x}-\bar{y})}{\varepsilon} + \frac{2A^{\frac{1}{2}-\eta} Q_N(\hat{x}^n-\bar{x})}{\varepsilon}\right\rangle \\
&\quad + \delta|A^{\frac{1}{2}}\hat{x}^n|^2 + G\Bigg(\hat{x}^n, D\varphi_n(P_N\hat{x}^n) + \frac{A^{-\eta} Q_N(\bar{x}-\bar{y})}{\varepsilon} + \frac{2A^{-\eta} Q_N(\hat{x}^n-\bar{x})}{\varepsilon} \\
&\qquad\qquad + \delta\hat{x}^n, D^2\varphi_n(P_N\hat{x}^n) + \frac{2A^{-\eta}Q_N}{\varepsilon} + \delta I\Bigg) \le 0.
\end{aligned} \tag{3.17}$$

We now would like to pass to the limit as $n \to \infty$ in the above inequality keeping ε, δ, N fixed. To do this we have to justify a lot of convergencies. We start by observing that setting

$$\hat{x}^n = (P_N\hat{x}^n, Q_N\hat{x}^n), \qquad \hat{y}^n = (P_N\hat{y}^n, Q_N\hat{y}^n) \quad \text{for every } x_N^\perp,\ y_N^\perp \in X_N^\perp,$$

we have

$$\begin{aligned}
&\tilde{u}_1(P_N\hat{x}^n) - \tilde{v}_1(P_N\hat{y}^n) - \varphi_n(P_N\hat{x}^n) + \psi_n(P_N\hat{y}^n) \\
&\quad \ge u_1(P_N\hat{x}^n, Q_N\hat{x}^n) - v_1(P_N\hat{y}^n, Q_N\hat{y}^n) - \varphi_n(P_N\hat{x}^n) + \psi_n(P_N\hat{y}^n) \\
&\quad \ge u_1(x_N^n, x_N^\perp) - v_1(y_N^n, y_N^\perp) - \varphi_n(x_N^n) + \psi_n(y_N^n).
\end{aligned}$$

Therefore taking suprema over $x_N^\perp$ and $y_N^\perp$ in the above inequality we obtain

$$\begin{aligned}
&\tilde{u}_1(P_N\hat{x}^n) - \tilde{v}_1(P_N\hat{y}^n) - \varphi_n(P_N\hat{x}^n) + \psi_n(P_N\hat{x}^n) \\
&\quad \ge u_1(P_N\hat{x}^n, Q_N\hat{x}^n) - v_1(P_N\hat{y}^n, Q_N\hat{y}^n) - \varphi_n(P_N\hat{x}^n) + \psi_n(P_N\hat{y}^n) \\
&\quad \ge \tilde{u}_1(x_N^n) - \tilde{v}_1(y_N^n) - \varphi_n(x_N^n) + \psi_n(y_N^n).
\end{aligned}$$

This implies that

$$P_N\hat{x}^n = x_N^n, \qquad P_N\hat{y}^n = y_N^n, \qquad u_1(\hat{x}^n) = \tilde{u}_1(x_N^n), \qquad v_1(\hat{y}^n) = \tilde{v}_1(y_N^n),$$

which, together with (3.13) and (3.12), yields

$$u_1(\hat{x}^n) \longrightarrow u_1(\bar{x}), \qquad v_1(\hat{y}^n) \longrightarrow u_1(\bar{y}) \tag{3.18}$$

as $n \to +\infty$. Finally, since

$$u_1(\hat{x}^n) = \tilde{u}_1(P_N\hat{x}^n), \quad v_1(\hat{y}^n) = \tilde{v}_1(P_N\hat{y}^n)$$

and

$$u_1(\bar{x}) = \tilde{u}_1(P_N\bar{x}), \quad v_1(\bar{y}) = \tilde{v}_1(P_N\bar{y}),$$

formula (3.18), together with the weak upper semicontinuity of u_1 and the weak lower semicontinuity of v_1, implies

$$\hat{x}^n \longrightarrow \bar{x}, \qquad \hat{y}^n \longrightarrow \bar{y} \tag{3.19}$$

as $n \to +\infty$. Therefore, using (3.19), (3.18), (3.6), (A0), and (3.3), it follows from (3.17) that $|A^{\frac{1}{2}}\hat{x}^n|$ are bounded independently of n which implies, thanks to (3.19), that $\bar{x} \in D(A^{\frac{1}{2}})$ and

$$A^{\frac{1}{2}}\hat{x}^n \rightharpoonup A^{\frac{1}{2}}\bar{x} \tag{3.20}$$

as $n \to +\infty$. Since $A^{\frac{\beta-1}{2}}$ and $A^{-\frac{\eta}{2}}$ are compact we conclude that, as $n \to +\infty$,

$$A^{\frac{\beta}{2}}\hat{x}^n = A^{\frac{\beta-1}{2}}\big(A^{\frac{1}{2}}\hat{x}^n\big) \to A^{\frac{\beta}{2}}\bar{x} \qquad \text{and} \qquad A^{\frac{1-\eta}{2}}\hat{x}^n \to A^{\frac{1-\eta}{2}}\bar{x}. \tag{3.21}$$

Using (3.14), (3.15), (3.20), (3.21), and the weak lower semicontinuity of norm we thus obtain that

$$\begin{aligned}
&\left\langle A^{\frac{1-\eta}{2}}\bar{x}, \frac{A^{\frac{1-\eta}{2}}(\bar{x}-\bar{y})}{\varepsilon}\right\rangle + \delta|A^{\frac{1}{2}}\bar{x}|^2 \\
&\qquad \le \liminf_{n\to\infty}\left[\left\langle A^{\frac{1}{2}}\hat{x}^n, A^{\frac{1}{2}}D\varphi_n(P_N\hat{x}^n) + \frac{A^{\frac{1}{2}-\eta}Q_N(\bar{x}-\bar{y})}{\varepsilon} + \frac{2A^{\frac{1}{2}-\eta}Q_N(\hat{x}^n-\bar{x})}{\varepsilon}\right\rangle\right. \\
&\qquad\qquad \left. + \delta|A^{\frac{1}{2}}\hat{x}^n|^2\right].
\end{aligned}$$

Therefore, letting $n \to \infty$ in (3.17) yields

$$\begin{aligned}
&\lambda u(\bar{x}) + \left\langle A^{\frac{1-\eta}{2}}\bar{x}, \frac{A^{\frac{1-\eta}{2}}(\bar{x}-\bar{y})}{\varepsilon}\right\rangle + \delta|A^{\frac{1}{2}}\bar{x}|^2 \\
&\qquad + G\left(\bar{x}, \frac{A^{-\eta}(\bar{x}-\bar{y})}{\varepsilon} + \delta\bar{x}, Z_N + \frac{2\|A^{-\eta}\|Q_N}{\varepsilon} + \delta I\right) \le 0.
\end{aligned} \tag{3.22}$$

We now eliminate terms with δ and N. Using (A2) we have

$$\begin{aligned}
&G\left(\bar{x}, \frac{A^{-\eta}(\bar{x}-\bar{y})}{\varepsilon}, Z_N + \frac{2\|A^{-\eta}\|Q_N}{\varepsilon}\right) - \rho\Big(d\delta\big(1+|A^{\frac{\beta}{2}}\bar{x}|^2\big)\Big) \\
&\qquad \le G\left(\bar{x}, \frac{A^{-\eta}(\bar{x}-\bar{y})}{\varepsilon} + \delta\bar{x}, Z_N + \frac{2\|A^{-\eta}\|Q_N}{\varepsilon} + \delta I\right)
\end{aligned} \tag{3.23}$$

for some constant $d > 0$. Now, given $\tau > 0$, let K_τ be such that $\rho(s) \le \tau + K_\tau s$. Applying (3.3) with $\alpha = \beta/2$ and $\gamma = 1/2$ we obtain that

$$\rho\big(d\delta\big(1+|A^{\frac{\beta}{2}}\bar{x}|^2\big)\big) \le \frac{\delta}{2}|A^{\frac{1}{2}}\bar{x}|^2 + \delta C_\tau|\bar{x}|^2 + \tau + K_\tau\, d\delta$$

for some constant $C_\tau > 0$ independent of δ and ε. It then follows from (3.10) that

$$\limsup_{\delta\to 0}\Big(\rho\big(d\delta\big(1+|A^{\frac{\beta}{2}}\bar{x}|^2\big)\big) - \delta|A^{\frac{1}{2}}\bar{x}|^2\Big) \le 0. \tag{3.24}$$

Using this, (3.23), and (3.5) in (3.22), we therefore obtain

(3.25)
$$\lambda u(\bar{x}) + \left\langle A^{\frac{1-\eta}{2}}\bar{x}, \frac{A^{\frac{1-\eta}{2}}(\bar{x}-\bar{y})}{\varepsilon}\right\rangle + G\left(\bar{x}, \frac{A^{-\eta}(\bar{x}-\bar{y})}{\varepsilon}, Z_N\right) \le \omega_1(N;\delta,\varepsilon) + \omega_2(\delta,\varepsilon),$$

where $\lim_{N\to\infty}\omega_1(N;\delta,\varepsilon) = 0$ if ε, δ are fixed and ω_2 is a local modulus. Similarly we obtain

(3.26)
$$\lambda v(\bar{y}) + \left\langle A^{\frac{1-\eta}{2}}\bar{y}, \frac{A^{\frac{1-\eta}{2}}(\bar{x}-\bar{y})}{\varepsilon}\right\rangle + G\left(\bar{y}, \frac{A^{-\eta}(\bar{x}-\bar{y})}{\varepsilon}, Y_N\right) \ge -\omega_1(N;\delta,\varepsilon) - \omega_2(\delta,\varepsilon).$$

We now subtract (3.26) from (3.25), use (A3), and then let $N \to \infty$. We then conclude that

$$\lambda\big(u(\bar{x}) - v(\bar{y})\big) \le \omega\left(\big|A^{\frac{\beta}{2}}(\bar{x}-\bar{y})\big|\left(1 + \frac{|A^{\frac{\beta}{2}}(\bar{x}-\bar{y})|}{\varepsilon}\right)\right) - \frac{|A^{\frac{1-\eta}{2}}(\bar{x}-\bar{y})|^2}{\varepsilon} + 2\omega_2(\delta,\varepsilon).$$

Set $r = |A^{\frac{1-\eta}{2}}(\bar{x}-\bar{y})|$, and let $\sigma > 0$. Using the interpolation inequality (3.3), the fact that $|A^{\frac{\beta}{2}}(\bar{x}-\bar{y})| \le c|A^{\frac{1-\eta}{2}}(\bar{x}-\bar{y})|$ for some $c > 0$ and the property of the moduli, we have that, for all $\alpha,\ \sigma > 0$, there exist $C_\sigma,\ K_\alpha > 0$ such that

$$\lambda\big(u(\bar{x}) - v(\bar{y})\big) \le \alpha + cK_\alpha\left(\sigma\frac{r^2}{\varepsilon} + C_\sigma\frac{\big|A^{-\frac{\eta}{2}}(\bar{x}-\bar{y})\big|}{\varepsilon}r + r\right) - \frac{r^2}{\varepsilon} + 2\omega_2(\delta,\varepsilon).$$

For α fixed, we chose σ such that $cK_\alpha\sigma < 1$. Then, in the right-hand side of the previous inequality, we have a polynomial of order 2 in $r/\sqrt{\varepsilon}$ which is bounded from above and we get

$$\lambda\big(u(\bar{x}) - v(\bar{y})\big) \le \alpha + \frac{K_\alpha^2 c^2\left(\sqrt{\varepsilon} + C_\sigma\frac{|A^{-\frac{\eta}{2}}(\bar{x}-\bar{y})|}{\sqrt{\varepsilon}}\right)^2}{4(1 - K_\alpha c\sigma)} + 2\omega_2(\delta,\varepsilon). \tag{3.27}$$

By using (3.9), this yields

$$\limsup_{\varepsilon\to 0}\limsup_{\delta\to 0}\big(u(\bar{x}) - v(\bar{y})\big) \le \alpha$$

for all $\alpha > 0$, which implies $u \le v$ in X since for all $x \in X$, we have

$$\Phi(x,x) \le \Phi(\bar{x},\bar{y}) \le u(\bar{x}) - v(\bar{y}).$$

Now, let u be a solution. We can set $u = v$ in the preceding proof and we obtain that for all x, y in X

$$u(x) - u(y) - \frac{|A^{-\frac{\eta}{2}}(x-y)|^2}{2\varepsilon} = \limsup_{\delta\to 0}\Phi(x,y) \le \limsup_{\delta\to 0}\big(u(\bar{x}) - u(\bar{y})\big) \le \rho_1(\epsilon)$$

for some modulus ρ_1 in light of (3.27). This proves (3.7).

Existence. Consider the approximating equation

$$\lambda u_N + \langle Ax, Du_N\rangle + G(x, Du_N, D^2u_N) = 0 \quad \text{in } X_N. \tag{3.28}$$

We notice that (3.28) satisfies the assumptions of the comparison part of the theorem with constants and moduli independent of N. By the finite dimensional theory (3.28) has a solution u_N (for every N) such that $\|u_N\|_\infty \le K/\lambda$. We will prove that there exists a modulus σ independent of N such that

$$\big|u_N(x) - u_N(y)\big| \le \sigma\big(|x-y|_{-\eta}\big)$$

for all $x, y \in X_N$. To do this we adapt the technique of [30] which was also used in [43].

For every $\varepsilon > 0$ let K_ε be such that $\omega(r) \le \lambda\varepsilon/2 + K_\varepsilon r$. For $L > K/\lambda + 1$ we set

$$\psi_L(r) = 2L2^{1-\frac{1}{2L}} r^{\frac{1}{2L}}.$$

The function $\psi_L \in C^2(0,\infty)$ is increasing and concave, $\psi_L'(r) \ge 1$ for $0 < r \le 2$, $\psi_L(0) = 0$, $\psi_L(1) > 2(K/\lambda + 1)$, and

$$\psi_L(r) > L\big(\psi_L'(r)r + r\big) \quad \text{for } 0 \le r \le 2. \tag{3.29}$$

We will show that for every $\varepsilon > 0$ there exists $L = L_\varepsilon$ such that

$$u_N(x) - u_N(y) \le \psi_L\big(|A^{-\frac{\eta}{2}}(x-y)|\big) + \varepsilon \quad \text{for every } x, y \in X. \tag{3.30}$$

Denoting by $\Delta = \{(x,y) \in X \times X : |A^{-\frac{\eta}{2}}(x-y)| < 1\}$ it is clear, from the properties of ψ_L, that for $(x,y) \notin \Delta$ (3.30) is always satisfied independently of L. Assume now by contradiction that (3.30) is false. Then, given any $L > \frac{K}{\lambda} + 1$ we have, for small $\delta > 0$, that

$$\sup_{(x,y)\in X\times X} \left(u_N(x) - u_N(y) - \psi_L(|A^{-\frac{\eta}{2}}(x-y)|) - \varepsilon - \frac{\delta}{2}|x|^2 - \frac{\delta}{2}|y|^2\right) > 0 \tag{3.31}$$

and the supremum is attained at $(\bar x, \bar y) \in \Delta$ such that $\bar x \ne \bar y$. Denote $s = |A^{-\frac{\eta}{2}}(\bar x - \bar y)|$. Using Lemma 2.3 in [43] (see the proof of Proposition 2.5 in [43]) and then repeating arguments from the just-finished proof of comparison we obtain that there exist $Z, Y \in \Sigma(X_N)$ such that

$$\begin{pmatrix} Z & 0 \\ 0 & -Y \end{pmatrix} \le \frac{2\psi_L'(s)}{s} \begin{pmatrix} P_N A^{-\eta} P_N & -P_N A^{-\eta} P_N \\ -P_N A^{-\eta} P_N & P_N A^{-\eta} P_N \end{pmatrix}$$

and

$$\begin{aligned}
\lambda(u_N(\bar x) - u_N(\bar y)) &\le -\frac{\psi_L'(s)}{s}|A^{\frac{1-\eta}{2}}(\bar x - \bar y)|^2 + G\left(\bar y, \frac{\psi_L'(s)}{s}A^{-\eta}(\bar x - \bar y), Y\right) \\
&\quad - G\left(\bar x, \frac{\psi_L'(s)}{s}A^{-\eta}(\bar x - \bar y), Z\right) + \rho(\delta, L) \\
&\le -\frac{\psi_L'(s)}{s}\big|A^{\frac{1-\eta}{2}}(\bar x - \bar y)\big|^2 + \frac{\lambda\varepsilon}{2} \\
&\quad + K_\varepsilon\left(\big|A^{\frac{\beta}{2}}(\bar x - \bar y)\big|\left(1 + \frac{\psi_L'(s)}{s}\big|A^{\frac{\beta}{2}}(\bar x - \bar y)\big|\right)\right) + \rho(\delta, L)
\end{aligned}$$

for some local modulus ρ. Therefore, using (3.3) with a sufficiently small σ, it follows that

$$\begin{aligned}
\lambda(u_N(\bar x) - u_N(\bar y)) &\le -\frac{\psi_L'(s)}{2s}\big|A^{\frac{1-\eta}{2}}(\bar x - \bar y)\big|^2 + \frac{\lambda\varepsilon}{2} \\
&\quad + C_\varepsilon\big(\psi_L'(s)s + s\big) + \frac{c}{2}|A^{\frac{1-\eta}{2}}(\bar x - \bar y)| + \rho(\delta, L),
\end{aligned}$$

where C_ε depends only on K_ε and the interpolation constant (but not on L), and c is such that $|A^{\frac{1-\eta}{2}}x| \geq c|A^{-\frac{\eta}{2}}x|$ for all $x \in H$. Thus, we eventually have

$$\lambda\big(u_N(\bar{x}) - u_N(\bar{y})\big) \leq \frac{\lambda\varepsilon}{2} + C_\varepsilon\big(\psi_L'(s)s + s\big) + \rho(\delta, L),$$

which becomes, choosing $L = C_\varepsilon/\lambda$ and letting $\delta \to 0$,

$$u_N(\bar{x}) - u_N(\bar{y}) \leq \frac{\varepsilon}{2} + L\big(\psi_L'(s)s + s\big).$$

This leads to a contradiction in light of (3.29) since we obviously have by (3.31)

$$\psi_L(s) + \varepsilon \leq u_N(\bar{x}) - u_N(\bar{y}).$$

Hence, we obtain the required modulus of continuity.

Now set $v_N(x) = u_N(P_N x)$. Since $A^{-\frac{\eta}{2}}$ is compact we are in a position to apply the Arzelà–Ascoli theorem to find a subsequence (still denoted by v_N) converging uniformly on bounded sets of X to a function u that obviously satisfies the same estimates as u_N's (see [14, Part IV] and [43] for more). It remains to show that u solves the limiting equation (3.1). Let $\varphi \in C^2_-(X)$ and let $u(x) - \varphi(x) - \frac{\delta}{2}|x|^2$ have a maximum at $\hat{x}$ which we may assume to be strict. It follows that there exists a sequence $\hat{x}_N = P_N\hat{x}_N \to \hat{x}$ as $N \to \infty$ such that, for every $x \in X_N$,

$$v_N(x) - \varphi(x) - \frac{\delta}{2}|x|^2 \leq v_N(\hat{x}_N) - \varphi(\hat{x}_N) - \frac{\delta}{2}|\hat{x}_N|^2.$$

Therefore, since $AP_N = P_N A$,

$$\begin{aligned}\lambda u_N(\hat{x}_N) + \big\langle A^{\frac{1}{2}}\hat{x}_N, A^{\frac{1}{2}}D\varphi(\hat{x}_N)\big\rangle + \delta\big|A^{\frac{1}{2}}\hat{x}_N\big|^2 \\ + G\left(\hat{x}_N, P_N D\varphi(\hat{x}_N) + \delta\hat{x}_N, P_N(D^2\varphi(\hat{x}_N) + \delta I)P_N\right) \leq 0.\end{aligned} \tag{3.32}$$

Since $\hat{x}_N \in X_N$ and φ is a test function we have

$$\big|A^{\frac{1}{2}}D\varphi(\hat{x}_N)\big| \leq B + C\big|A^{\frac{1}{2}-\varepsilon}\hat{x}_N\big| \tag{3.33}$$

for some independent constants B, C. Also, by (3.33), (3.3), (3.6), and (3.8),

$$\begin{aligned}&\big|G\big(\hat{x}_N, P_N D\varphi(\hat{x}_N) + \delta\hat{x}_N, P_N(D^2\varphi(\hat{x}_N) + \delta I)P_N\big)\big| \\ &\qquad \leq C_1\Big(1 + |A^{\frac{\beta}{2}}\hat{x}_N|^2 + |A^{\frac{1}{2}-\varepsilon}\hat{x}_N|^2\Big) \leq C_2 + \frac{\delta}{4}|A^{\frac{1}{2}}\hat{x}_N|^2.\end{aligned}$$

Using this, (3.33), and the interpolation inequality (3.3), we therefore obtain from (3.32) that

$$\big|A^{\frac{1}{2}}\hat{x}_N\big| \leq C_3$$

for some constant C_3 independent of N. Thus, $A^{\frac{1}{2}}\hat{x}_N \rightharpoonup A^{\frac{1}{2}}\hat{x}$ (so $\hat{x} \in D(A^{\frac{1}{2}})$) and hence

$$A^{\frac{\beta}{2}}\hat{x}_N \to A^{\frac{\beta}{2}}\hat{x}, \qquad \text{and} \qquad A^{\frac{1}{2}}D\varphi(\hat{x}_N) \to A^{\frac{1}{2}}D\varphi(\hat{x}).$$

These convergencies and Lemma 2.8 in [43] allow us to pass to the limit in (3.32) as $N \to \infty$ to conclude that

$$\lambda u(\hat{x}) + \big\langle A^{\frac{1}{2}}\hat{x}, A^{\frac{1}{2}}D\varphi(\hat{x})\big\rangle + \delta\big|A^{\frac{1}{2}}\hat{x}\big|^2 + G\left(\hat{x}, D\varphi(\hat{x}) + \delta\hat{x}, D^2\varphi(\hat{x}) + \delta I\right) \leq 0. \qquad \square$$

Remark 3.9. If instead of (A2) we assume that there exist $\gamma < 1$ and a modulus ρ such that

$$\big|G(x,p,S_1) - G(x,q,S_2)\big| \le \rho\left(\big(1+|A^{\frac{\beta}{2}}x|^{\gamma}\big)\big|A^{\frac{\beta}{2}}(p-q)\big| + \big(1+|A^{\frac{\beta}{2}}x|^{2\gamma}\big)\|S_1-S_2\|\right)$$

for all $x,p,q \in D(A^{\frac{\beta}{2}})$ and $S_1, S_2 \in \Sigma(X)$, then the conclusion of the comparison part of Theorem 3.8 holds if we replace the assumption that $u, -v \le M$ by the assumption that

$$\big|u(x)-u(y)\big|, \big|v(x)-v(y)\big| \le m\big(|A^{-\frac{\eta}{2}}(x-y)|\big)$$

for some modulus m. The same proof applies except that an analogue of (3.24) has to be justified differently and instead of (3.10) we now obtain (by a standard argument) that

$$\frac{\big|A^{-\frac{\eta}{2}}(\bar{x}-\bar{y})\big|^2}{\varepsilon} \le t_\varepsilon \to 0 \quad \text{as } \varepsilon \to 0.$$

4. Viscosity solutions of the original equation. We now briefly explain how to define the viscosity solution of the original equation

$$\lambda v(x) + \big\langle Ax, Dv(x)\big\rangle + H\big(x, Dv(x), D^2v(x)\big) = 0, \quad x \in X, \tag{4.1}$$

where we assume that Hypothesis 3.1 and the following hold.

HYPOTHESIS 4.1. *The function $H : X \times D(A^\beta) \times A^{-\frac{\beta}{2}}\Sigma(X)A^{-\frac{\beta}{2}} \to \mathbb{R}$ is such that, if we define*

$$G_H(z,p,S) \stackrel{def}{=} H\left(A^{\frac{\beta}{2}}z, A^{-\frac{\beta}{2}}p, A^{-\frac{\beta}{2}}SA^{-\frac{\beta}{2}}\right),$$

then G_H satisfies Hypothesis 3.3.

Remark 4.2. The last hypothesis could also be written more explicitly by translating Hypothesis 3.3. We used the above formulation for brevity and also because it can be easily checked by the change of variable.

If we apply the change of variable $y = A^{-\frac{\beta}{2}}x$ and set $u(y) \stackrel{def}{=} v(A^{\frac{\beta}{2}}y)$, (4.1) for v formally becomes the following equation for u:

$$\lambda u(y) + \big\langle Ay, Du(y)\big\rangle + G_H\big(y, Du(y), D^2u(y)\big) = 0, \quad y \in X. \tag{4.2}$$

The function v is uniquely determined, once u has been characterized on $D(A^{\frac{\beta}{2}})$. Therefore we are driven to the following definition of viscosity solution for equation (4.1).

DEFINITION 4.3. *A bounded continuous function $v : X \to \mathbb{R}$ is said to be a viscosity solution of equation* (4.1) *if and only if the function*

$$u(y) \stackrel{def}{=} v\big(A^{\frac{\beta}{2}}y\big)$$

is a viscosity solution of the transformed equation (4.2). *Similarly we define a viscosity subsolution and a supersolution of* (4.1).

From Theorem 3.8 we have the following.

THEOREM 4.4. *Let Hypotheses* 3.1 *and* 4.1 *be satisfied. If*

$$\sup_{x\in X} |H(x,0,0)| = K < \infty, \tag{4.3}$$

then there exists a unique viscosity solution $v \in BUC(X_{-(\beta+\eta)})$ of (4.1).

5. The stochastic boundary control problem.

5.1. The state equation. Let X, U_1, and U_2 be real separable Hilbert spaces. Let $\tilde{U}_2$ be a given closed bounded subset of U_2 with $R = \sup_{h\in\tilde{U}_2}|h|$ and define $U \stackrel{def}{=} U_1 \times \tilde{U}_2$. Let $(\Omega, \mathcal{F}, (\mathcal{F}_t)_{t\geq 0}, \mathbf{P})$ and W be respectively the stochastic base and the cylindrical Brownian motion on X defined in section 2.

Consider a stochastic dynamical system governed by the following stochastic differential equation in X:

(5.1)
$$\begin{cases} dx(t) = \left[-Ax(t) + b(x(t),\alpha_1(t)) + A^\beta C\alpha_2(t)\right]dt + B(x(t),\alpha_1(t))dW(t), \quad t>0, \\ x(0) = x_0 \in X, \end{cases}$$

where we assume that Hypothesis 3.1 holds. We also assume the following.

HYPOTHESIS 5.1.

(i) $\alpha \stackrel{def}{=} (\alpha_1,\alpha_2) \in \mathcal{U}_{ad}(0,T;U) \stackrel{def}{=} M^0(0,T;U_1)\times M^0(0,T;\tilde{U}_2)$ *for every* $T>0$.

(ii) *The function* b *is continuous from* $X\times U_1$ *to* X *and there exists a constant* $c_0>0$ *such that*

$$|b(x,\alpha_1)| \leq c_0(1+|x|) \quad \textit{for all } x\in X,\ \alpha_1\in U_1,$$
$$|b(x_1,\alpha_1)-b(x_2,\alpha_1)| \leq c_0|x_1-x_2| \quad \textit{for all } x_1,x_2\in X,\ \alpha_1\in U_1.$$

(iii) $C\in\mathcal{L}(U_2,X)$ *and* $\beta\in(\frac{3}{4},1)$.

(iv) B *is a mapping from* $X\times U_1$ *into* $\mathcal{L}(X)$ *such that* $A^{-\frac{\beta}{2}}B : X\times U_1 \to \mathcal{L}_2(X)$ *is continuous, and moreover, there exists a constant* $K_1>0$ *such that*

$$\|A^{-\frac{\beta}{2}}B(x,\alpha_1)\|_{\mathcal{L}_2(X)} \leq K_1(1+|x|) \quad \textit{for all } x\in X,\ \alpha_1\in U_1,$$
$$\|A^{-\frac{\beta}{2}}[B(x_1,\alpha_1)-B(x_2,\alpha_1)]\|_{\mathcal{L}_2(X)} \leq K_1|x_1-x_2| \quad \textit{for all } x_1,x_2\in X,\ \alpha_1\in U_1.$$

(v) *For all* $x\in X$,

$$\lim_{N\to+\infty}\sup_{\alpha_1\in U_1}\left\|Q_N A^{-\frac{\beta}{2}}B(x,\alpha_1)\right\|_{\mathcal{L}_2(X)} = 0.$$

Remark 5.2.

- Hypotheses 5.1(iv), (v) are satisfied if we assume, for example, that there exists a constant $K_2>0$ such that

$$\|B(x,\alpha_1)\| \leq K_2(1+|x|) \quad \text{for all } x\in X,\ \alpha_1\in U_1,$$
$$\|B(x_1,\alpha_1)-B(x_2,\alpha_1)\| \leq K_2|x_1-x_2| \quad \text{for all } x_1,x_2\in X,\ \alpha_1\in U_1,$$

  and if the operator $A^{-\beta}$ is trace class.
- Hypothesis 5.1(v) is satisfied if, for instance, for every $x\in X$ there exists $\eta\in(0,\beta/2)$ such that $A^{-\eta}B(x,\alpha_1)$ is bounded in $\mathcal{L}_2(X)$ independently of $\alpha_1\in U_1$.

Equation (5.1) contains terms that are not well defined in general. However it is possible to give sense to its integral form. For this reason we introduce the concept of mild solution of (5.1) (see, e.g., [17]).

DEFINITION 5.3. *An* $\mathcal{F}_t$*-adapted process* $x(t)$, $t\geq 0$, *is said to be a mild solution of* (5.1) *if*

$$\mathbf{P}\left(\int_0^t |x(s)|^2 ds < +\infty\right) = 1 \quad \textit{for all } t\geq 0$$

and if it satisfies the integral equation

$$
\begin{aligned}
x(t) = e^{-tA}x_0 &+ \int_0^t e^{-(t-s)A} b\big(x(s), \alpha_1(s)\big)\, ds + A^\beta \int_0^t e^{-(t-s)A} C\alpha_2(s)\, ds \\
&+ \int_0^t e^{-(t-s)A} B\big(x(s), \alpha_1(s)\big)\, dW(s), \quad \mathbf{P}\text{a.s., } t \geq 0.
\end{aligned} \tag{5.2}
$$

Equation (5.2) *will be called the mild form of* (5.1).

If, given $T \geq 0$, we denote by $\mathcal{H}_{2,T}$ the Banach space of all (equivalence classes of) predictable X-valued processes $z(t)$, $t \geq 0$ such that

$$
\|z\|_{\mathcal{H}_{2,T}} \stackrel{def}{=} \sup_{s\in[0,T]} \big(\mathbb{E}|z(s)|^2\big)^{\frac{1}{2}} < +\infty,
$$

then we have the following result (see [17, Chapter 7] and [18, Chapter 5]).

THEOREM 5.4. *Assume that Hypotheses* 3.1 *and* 5.1(i)–(iv) *hold. Then, for any initial condition x_0 there exists a unique mild solution $x(\cdot) = x(\cdot; x_0, \alpha)$ of* (5.1) *such that for every $T > 0$ $x(\cdot) \in \mathcal{H}_{2,T}$ and there exists a constant C_T independent of x_0 such that*

$$
\sup_{s\in[0,T]} \mathbb{E}|x(s)|^2 \leq C_T\big(1 + |x_0|^2\big).
$$

Finally, the solution has continuous trajectories $\mathbf{P}$*-a.s.*

Proof. The proof is completely similar to the one given in [17, Chapter 7] and [18, Chapter 5]. The only differences between (5.2) and those treated in the results above are the presence of the term

$$
A^\beta \int_0^t e^{-(t-s)A} C\alpha_2(s)\, ds \tag{5.3}
$$

and the nonstandard assumption 5.1(iv). The term (5.3) does not influence the proof of the theorem, since, when α_2 is bounded a.s., we have for some constant $C_0 > 0$,

$$
\mathbb{E}\left| A^\beta \int_0^t e^{-(t-s)A} C\alpha_2(s)\, ds \right|^2 \leq C_0 R^2 \left(\int_0^t \frac{e^{-a(t-s)}}{(t-s)^\beta}\, ds \right)^2.
$$

Moreover, the stochastic convolution can be estimated by writing

$$
\int_0^t e^{-(t-s)A} B\big(x(s), \alpha_1(s)\big)\, dW(s) = \int_0^t A^{\frac{\beta}{2}} e^{-(t-s)A} A^{-\frac{\beta}{2}} B\big(x(s), \alpha_1(s)\big)\, dW(s)
$$

so that, thanks to (3.2),

$$
\begin{aligned}
\mathbb{E}\left| \int_0^t e^{-(t-s)A} B\big(x(s), \alpha_1(s)\big)\, dW(s) \right|^2 & \\
\leq \int_0^t \frac{M_{\frac{\beta}{2}}^2 e^{-2a(t-s)}}{(t-s)^\beta} \mathbb{E}\big\| A^{-\frac{\beta}{2}} B\big(x(s), \alpha_1(s)\big)\big\|^2_{\mathcal{L}_2(X)}\, ds. &
\end{aligned}
$$

In particular this ensures also that (2.1) holds as well as the continuity of trajectories of the solution. □

5.2. Change of variables and main estimates. As we said in the introduction, we transform (5.1) by the change of variables

$$y(t) = A^{-\frac{\beta}{2}} x(t). \tag{5.4}$$

Then the function $y(\cdot) = y(\cdot; y_0, \alpha) = A^{-\frac{\beta}{2}} x(\cdot; x_0, \alpha)$ satisfies the equation

$$\begin{cases} dy(t) = \big[-Ay(t) + A^{-\frac{\beta}{2}} b(A^{\frac{\beta}{2}} y(t), \alpha_1(t)) + A^{\frac{\beta}{2}} C\alpha_2(t) \big] dt \\ \qquad\quad + A^{-\frac{\beta}{2}} B\big(A^{\frac{\beta}{2}} y(t), \alpha_1(t)\big)\, dW(t), \\ y(0) = y_0 = A^{-\frac{\beta}{2}} x_0 \in X. \end{cases}$$

Again, the above equation has to be understood in its mild form

$$\begin{aligned} y(t) &= e^{-tA} y_0 + A^{-\frac{\beta}{2}} \int_0^t e^{-(t-s)A} b\big(A^{\frac{\beta}{2}} y(s), \alpha_1(s)\big)\, ds \\ &\quad + A^{\frac{\beta}{2}} \int_0^t e^{-(t-s)A} C\alpha_2(s)\, ds + A^{-\frac{\beta}{2}} \int_0^t e^{-(t-s)A} B\big(A^{\frac{\beta}{2}} y(s), \alpha_1(s)\big)\, dW(s). \end{aligned} \tag{5.5}$$

When $y_0 \in D(A^{\frac{\beta}{2}})$, the last equation is nothing else but a different form of (5.1) obtained by the change of variables (5.4), and so the function

$$y(\cdot\,; y_0, \alpha) = A^{-\frac{\beta}{2}} x\big(\cdot\,; A^{\frac{\beta}{2}} y_0, \alpha\big)$$

is a mild solution of (5.5), and the following estimate holds:

$$\big\| A^{\frac{\beta}{2}} y \big\|^2_{\mathcal{H}_{2,T}} \le C_T \big(1 + |A^{\frac{\beta}{2}} y_0|^2\big).$$

However, one can also find a solution of (5.5) for a general initial condition $y_0 \in X$.

Denote by $M^2(a, b; D(A^\gamma))$, for $\gamma \in [0, 1]$, the Banach space of all (equivalence classes of) predictable X-valued processes $z(t)$, for $t \ge 0$, such that

$$\|z\|_{2,\gamma} \stackrel{def}{=} \left(\mathbb{E} \int_a^b |A^\gamma z(s)|^2 ds \right)^{\frac{1}{2}} < +\infty.$$

By a careful application of the arguments used in the proof of Theorem 5.4 we get the following result.

PROPOSITION 5.5. *Assume that Hypotheses* 3.1 *and* 5.1 *hold. Then for every* $y_0 \in X$, *equation* (5.5) *has a unique solution* $y(\cdot) = y(\cdot\,; y_0, \alpha)$ *such that for every* $T > 0$, $y \in M^2(0, T; D(A^{\frac{\beta}{2}}))$. *Moreover,* y *has continuous trajectories to* 0.

Proof. The proof of existence and uniqueness follows by applying the contraction mapping principle in a suitable way. Let $y_0 \in X$; for $y \in M^2(0, T; D(A^{\frac{\beta}{2}}))$, we define a map Λ on $M^2(0, T; D(A^{\frac{\beta}{2}}))$ by

$$\begin{aligned} [\Lambda y](t) &= e^{-tA} y_0 + A^{-\frac{\beta}{2}} \int_0^t e^{-(t-s)A} b\big(A^{\frac{\beta}{2}} y(s), \alpha_1(s)\big)\, ds + A^{\frac{\beta}{2}} \int_0^t e^{-(t-s)A} C\alpha_2(s)\, ds \\ &\quad + A^{-\frac{\beta}{2}} \int_0^t e^{-(t-s)A} B\big(A^{\frac{\beta}{2}} y(s), \alpha_1(s)\big)\, dW(s). \end{aligned}$$

First, we prove that Λ maps $M^2(0,T;D(A^{\frac{\beta}{2}}))$ into itself. For $y \in M^2(0,T;D(A^{\frac{\beta}{2}}))$, we have

$$\begin{aligned} A^{\frac{\beta}{2}}[\Lambda y](t) &= A^{\frac{\beta}{2}} e^{-tA} y_0 + \int_0^t e^{-(t-s)A} b\big(A^{\frac{\beta}{2}} y(s), \alpha_1(s)\big)\, ds \\ &\quad + A^{\beta} \int_0^t e^{-(t-s)A} C\alpha_2(s)\, ds \\ &\quad + A^{\frac{\beta}{2}} \int_0^t e^{-(t-s)A} A^{-\frac{\beta}{2}} B\big(A^{\frac{\beta}{2}} y(s), \alpha_1(s)\big)\, dW(s) \end{aligned}$$

so that, for suitable constants $C_1, C_2 > 0$,

$$\begin{aligned} \big|A^{\frac{\beta}{2}}[\Lambda y](t)\big| &\le \big|A^{\frac{\beta}{2}} e^{-tA} y_0\big| + C_1 \int_0^t e^{-a(t-s)} \big[1 + |A^{\frac{\beta}{2}} y(s)|\big]\, ds \\ &\quad + C_2 R \int_0^t \frac{e^{-a(t-s)}}{(t-s)^{\beta}} ds + \left| \int_0^t A^{\frac{\beta}{2}} e^{-(t-s)A} A^{-\frac{\beta}{2}} B\big(A^{\frac{\beta}{2}} y(s), \alpha_1(s)\big)\, dW(s) \right|. \end{aligned}$$

Then, taking the mean value of the square of the terms of this last inequality and using the isometry formula for stochastic integrals, we get

$$\begin{aligned} \mathbb{E}\Big|A^{\frac{\beta}{2}}[\Lambda y](t)\Big|^2 &\le C_3 \Big[\big|A^{\frac{\beta}{2}} e^{-tA} y_0\big|^2 + \int_0^t \big[1 + \mathbb{E}\big|A^{\frac{\beta}{2}} y(s)\big|^2\big]\, ds \\ &\quad + R^2 + \int_0^t \frac{1}{(t-s)^{\beta}} \big[1 + \mathbb{E}\big|A^{\frac{\beta}{2}} y(s)\big|^2\big]\, ds \Big], \end{aligned} \tag{5.6}$$

where $C_3 > 0$ is a suitable constant, since

$$\int_0^t \frac{e^{-a(t-s)}}{(t-s)^{\beta}}\, ds$$

is bounded independently of $t > 0$. Integrating over $[0,T]$ we obtain, for suitable $C_4 > 0$,

$$\begin{aligned} \|\Lambda y\|^2_{M^2\left(0,T;D\left(A^{\frac{\beta}{2}}\right)\right)} &= \int_0^T \mathbb{E}\left|A^{\frac{\beta}{2}}[\Lambda y](t)\right|^2 dt \\ &\le C_4 \Big[|y_0|^2 T^{1-\beta} + T^2 + T + T^{2-\beta} \\ &\quad + \int_0^T \int_0^t \left[1 + \frac{1}{(t-s)^{\beta}}\right] \mathbb{E}\big|A^{\frac{\beta}{2}} y(s)\big|^2\, ds \Big]. \end{aligned}$$

Moreover, since

$$\begin{aligned} \int_0^T \int_0^t \left[1 + \frac{1}{(t-s)^{\beta}}\right] \mathbb{E}\big|A^{\frac{\beta}{2}} y(s)\big|^2 ds\, dt &= \int_0^T \mathbb{E}\big|A^{\frac{\beta}{2}} y(s)\big|^2 \int_s^T \left[1 + \frac{1}{(t-s)^{\beta}}\right] dt\, ds \\ &= \int_0^T \mathbb{E}|A^{\frac{\beta}{2}} y(s)|^2 \left[T - s + \frac{(T-s)^{1-\beta}}{1-\beta}\right] ds, \end{aligned}$$

it finally follows that, for some modulus ω,

$$\|\Lambda y\|^2_{M^2\left(0,T;D\left(A^{\frac{\beta}{2}}\right)\right)} \le \omega(T) \left[1 + |y_0|^2 + \|y\|^2_{M^2\left(0,T;D\left(A^{\frac{\beta}{2}}\right)\right)}\right].$$

We now prove that Λ is a contraction on $M^2(0,T;D(A^{\frac{\beta}{2}}))$ when T is sufficiently small. Let $y_1(\cdot), y_2(\cdot) \in M^2(0,T;D(A^{\frac{\beta}{2}}))$. Then, arguing as above, we have

$$\mathbb{E}\left|A^{\frac{\beta}{2}}\big(\Lambda y_1(t) - \Lambda y_2(t)\big)\right|^2 \le C_4 \int_0^t \left(1 + \frac{1}{(t-s)^\beta}\right) \mathbb{E}\left|A^{\frac{\beta}{2}}\big(y_1(s) - y_2(s)\big)\right|^2 ds$$

for some constant $C_4 > 0$, which implies

$$\|\Lambda y_1 - \Lambda y_2\|_{M^2\left(0,T;D\left(A^{\frac{\beta}{2}}\right)\right)} \le \omega_1(T)\, \|y_1 - y_2\|_{M^2\left(0,T;D\left(A^{\frac{\beta}{2}}\right)\right)}$$

for some modulus ω_1. This implies that Λ is a contraction on $M^2(0,T;D(A^{\frac{\beta}{2}}))$ for small T. The existence of a unique solution in $M^2(0,T;D(A^{\frac{\beta}{2}}))$ for general $T > 0$ follows by repeating the argument a finite number of times. Finally this implies the existence of a unique solution $y(\cdot\,; y_0, \alpha)$ of (5.5) defined on $[0,+\infty)$ and such that $y(\cdot\,; y_0, \alpha) \in M^2(0,T;D(A^{\frac{\beta}{2}}))$ for every $T > 0$. The proof of continuity of trajectories is exactly the same as in the proof of Theorem 5.4. □

We conclude this section with some estimates for the solution of (5.5). First we recall a well-known generalization of the Gronwall lemma (see [29, p. 188]).

LEMMA 5.6. *Let $T \in (0,+\infty]$ and let $a_0 : [0,T] \to \mathbb{R}$ be a locally integrable function. Assume that $\varphi : [0,T] \to \mathbb{R}$ is a locally integrable function satisfying the inequality, for some $\eta \in (0,1)$,*

$$\varphi(t) \le a_0(t) + b_0 \int_0^t \Big[1 + \frac{1}{(t-s)^\eta}\Big]\varphi(s)\, ds$$

for a given constant $b_0 > 0$. Then there exists a positive constant $C_0 = C_0(b_0,\eta)$ such that

$$\varphi(t) \le a_0(t) + C_0 \int_0^t \frac{e^{C_0(t-s)}}{(t-s)^\eta} a_0(s)\, ds.$$

PROPOSITION 5.7. *Assume that Hypothesis 5.1 holds. Let $T > 0$, $y_0 \in X$. Then there exists a constant $C(T,y_0)$ such that, for all $t \in (0,T]$ and all $\alpha \in \mathcal{U}_{ad}(0,+\infty;U)$, we have*

$$\mathbb{E}\big|A^{\frac{\beta}{2}} y(t;y_0,\alpha)\big|^2 \le C(T,y_0)\left(1 + \frac{1}{t^\beta}\right). \tag{5.7}$$

Moreover, for every $\gamma \in (0, 1-\beta)$, there exists a constant $C_\gamma(T,y_0) > 0$ such that, for all $t \in (0,T]$ and $\alpha \in \mathcal{U}_{ad}(0,+\infty;U)$, we have

$$\int_0^t \left(1 + \frac{1}{(t-s)^{\beta+\gamma}}\right) \mathbb{E}\big|A^{\frac{\beta}{2}} y(s;y_0,\alpha)\big|^2 ds \le C_\gamma(T,y_0)\left(1 + \frac{1}{t^\beta}\right). \tag{5.8}$$

Proof. We use the arguments from the proof of Proposition 5.5. Let $y_0 \in X$, $T > 0$, $\alpha \in \mathcal{U}_{ad}(0,T;U)$, $y(\cdot) = y(\cdot;y_0,\alpha)$ and $t \in (0,T]$. The estimate (5.6) implies that there exist some $C > 0$ and then some $C(T) > 0$ such that

$$\begin{aligned}\mathbb{E}\big|A^{\frac{\beta}{2}} y(t)\big|^2 &\le C\left[1 + t + t^{1-\beta} + \big|A^{\frac{\beta}{2}} e^{-tA} y_0\big|^2 + \int_0^t \left(1 + \frac{1}{(t-s)^\beta}\right) \mathbb{E}\big|A^{\frac{\beta}{2}} y(s)\big|^2 ds\right] \\ &\le C(T)\left(1 + \frac{|y_0|^2}{t^\beta} + \int_0^t \left(1 + \frac{1}{(t-s)^\beta}\right) \mathbb{E}\big|A^{\frac{\beta}{2}} y(s)\big|^2 ds\right).\end{aligned}$$

We apply Lemma 5.6 and we obtain for a suitable $\overline{C}(T, y_0) > 0$ that

$$\begin{aligned}\mathbb{E}\big|A^{\frac{\beta}{2}} y(t)\big|^2 &\le \overline{C}(T, y_0)\left(1 + \frac{1}{t^\beta} + \int_0^t \frac{1}{(t-s)^\beta}\left(1 + \frac{1}{s^\beta}\right) ds\right) \\ &= \overline{C}(T, y_0)\left(1 + \frac{1}{t^\beta} + \frac{t^{1-\beta}}{1-\beta} + \int_0^t \frac{1}{(t-s)^\beta}\frac{1}{s^\beta}\, ds\right).\end{aligned}$$

However,

$$\int_0^t \frac{1}{(t-s)^\beta}\frac{1}{s^\beta}\, ds = t^{1-2\beta}\int_0^1 \frac{1}{(1-s)^\beta}\frac{1}{s^\beta}\, ds,$$

and this last integral is bounded so that, since $t^{1-2\beta} \le t^{-\beta}T^{1-\beta}$, we obtain the desired result.

Now, let $\gamma \in (0, 1-\beta)$. We have by (5.7)

$$\int_0^t \frac{1}{(t-s)^{\beta+\gamma}}\mathbb{E}\big|A^{\frac{\beta}{2}} y(s)\big|^2\, ds \le C(T, y_0)\int_0^t \frac{1}{(t-s)^{\beta+\gamma}}\left(1 + \frac{1}{s^\beta}\right) ds.$$

We proceed as above and the result follows. □

5.3. Properties of the value function. We now consider the stochastic optimal control problem of minimizing the functional

$$J(x_0; \alpha) = \mathbb{E}\int_0^{+\infty} e^{-\lambda t} L\big(x(t; x_0, \alpha), \alpha(t)\big)\, dt, \qquad x_0 \in X \tag{5.9}$$

over all functions $\alpha \in \mathcal{U}_{ad}(0, +\infty; U)$ (called controls). Here $x(\cdot; x_0, \alpha)$ is the mild solution of equation (5.1), i.e., the solution of the integral equation (5.2). The discount factor λ is positive and L satisfies the following assumptions:

$$\begin{aligned}&\text{(i)} \quad L \in C(X \times U), \ \ |L(x, \alpha)| \le C_L \quad \text{for all } (x, \alpha) \in X \times U; \\ &\text{(ii)} \quad |L(x_1, \alpha) - L(x_2, \alpha)| \le \omega_L(|x_1 - x_2|) \quad \text{for all } \alpha \in U, \ x_1, x_2 \in X,\end{aligned} \tag{5.10}$$

for some positive constant C_L and some modulus ω_L.

The value function

$$v(x_0) = \inf\big\{J(x_0; \alpha);\ \alpha \in \mathcal{U}_{ad}(0, +\infty; U)\big\}, \quad x_0 \in X \tag{5.11}$$

should solve the associated Hamilton–Jacobi equation

$$\lambda v(x) + \big\langle Ax, Dv(x)\big\rangle + H\big(x, Dv(x), D^2 v(x)\big) = 0 \quad \text{for } x \in X, \tag{5.12}$$

where the Hamiltonian $H : X \times D(A^\beta) \times A^{-\frac{\beta}{2}}\Sigma(X)A^{-\frac{\beta}{2}} \to \mathbb{R}$ is given by

$$\begin{aligned}H(x, p, S) = \sup_{\alpha \in U}\Big\{ &-\frac{1}{2}\operatorname{Tr}\big[B^*(x, \alpha_1)SB(x, \alpha_1)\big] \\ &- \big\langle b(x, \alpha_1), p\big\rangle - \big\langle C\alpha_2, A^\beta p\big\rangle - L(x, \alpha)\Big\}.\end{aligned}$$

The connection between the control problem introduced above and stochastic boundary control problems is discussed in section 5.4.

By inserting the change of variable (5.4) in the cost functional (5.9), we obtain a new optimal control problem whose value function u is given by

$$u(y_0) = \inf_{\alpha(\cdot)\in\mathcal{U}_{ad}(0,+\infty;U)} \mathbb{E}\int_0^{+\infty} e^{-\lambda t} L\big(A^{\frac{\beta}{2}} y(t;y_0,\alpha),\alpha(t)\big)\,dt. \tag{5.13}$$

The value functions u and v are related by the formula

$$v(x) = u\big(A^{-\frac{\beta}{2}}x\big) \quad \text{for all } x\in X. \tag{5.14}$$

But if v solves (5.12) then u should solve

$$\lambda u(y) + \big\langle Ay, Du(y)\big\rangle + H\big(A^{\frac{\beta}{2}}y, A^{-\frac{\beta}{2}}Du(y), A^{-\frac{\beta}{2}}D^2u(y)A^{-\frac{\beta}{2}}\big) = 0. \tag{5.15}$$

By setting $G_H(y,q,S) = H(A^{\frac{\beta}{2}}y, A^{-\frac{\beta}{2}}q, A^{-\frac{\beta}{2}}SA^{-\frac{\beta}{2}})$, (5.15) can be written as

$$\lambda u(y) + \big\langle Ay, Du(y)\big\rangle + G_H\big(y, Du(y), D^2u(y)\big) = 0 \tag{5.16}$$

where now $G_H : D(A^{\frac{\beta}{2}}) \times D(A^{\frac{\beta}{2}}) \times \Sigma(X) \to \mathbb{R}$ and

$$G_H(y,q,S) = \sup_{\alpha\in U}\Big\{ -\frac{1}{2}\operatorname{Tr}\Big[\Big(A^{-\frac{\beta}{2}}B\big(A^{\frac{\beta}{2}}y,\alpha_1\big)\Big)^* S\Big(A^{-\frac{\beta}{2}}B\big(A^{\frac{\beta}{2}}y,\alpha_1\big)\Big)\Big] - \big\langle b\big(A^{\frac{\beta}{2}}y,\alpha_1\big), A^{-\frac{\beta}{2}}q\big\rangle - \big\langle C\alpha_2, A^{\frac{\beta}{2}}q\big\rangle - L\big(A^{\frac{\beta}{2}}y,\alpha_1\big)\Big\}.$$

The goal of this section is to prove that v is the unique viscosity solution of (5.12). In fact we will prove more. We will prove that u is the unique viscosity solution of (5.16) and that it satisfies the dynamic programming principle in X. To obtain the latter for v we require only that the dynamic programming principle for u be satisfied on $D(A^{\frac{\beta}{2}})$. Therefore in fact we prove that the transformed problem itself has a control interpretation and we think that this fact may be of independent interest.

THEOREM 5.8. *Assume that Hypotheses* 3.1 *and* 5.1 *and* (5.10) *hold. Then the value function v defined in* (5.11) *is the unique $BUC(X_{-\eta})$ viscosity solution (for every $\eta\in(0,1)$) of the Hamilton–Jacobi equation* (5.12). *Moreover, the dynamic programming principle holds for u and v, i.e., for $x\in X$ and all $T>0$,*

$$v(x) = \inf_{\alpha\in\mathcal{U}_{ad}(0,+\infty;U)} \mathbb{E}\left\{\int_0^T e^{-\lambda t}L\big(x(t;x,\alpha),\alpha(t)\big)\,dt + e^{-\lambda T}v\big(x(T;x,\alpha)\big)\right\}$$

and

$$u(y) = \inf_{\alpha\in\mathcal{U}_{ad}(0,+\infty;U)} \mathbb{E}\left\{\int_0^T e^{-\lambda t}L\big(A^{\frac{\beta}{2}}y(t;y,\alpha),\alpha(t)\big)\,dt + e^{-\lambda T}u\big(y(T;y,\alpha)\big)\right\}.$$

We will prove this theorem by the approximation argument used in the proof of the existence of Theorem 3.8. We need a preliminary lemma.

DEFINITION 5.9. *Let $N\ge 1$ and $y_0\in X$. For a given $\alpha\in\mathcal{U}_{ad}(0,+\infty;U)$, we define $y_N(\cdot;y_0,\alpha)\in M^2(0,+\infty;D(A^{\frac{\beta}{2}}))$ to be the solution of*

$$\begin{aligned} y_N(t) &= e^{-tA}P_Ny_0 + A^{-\frac{\beta}{2}}\int_0^t e^{-(t-s)A}P_Nb\big(A^{\frac{\beta}{2}}y_N(s),\alpha_1(s)\big)\,ds \\ &\quad + A^{\frac{\beta}{2}}\int_0^t e^{-(t-s)A}P_NC\alpha_2(s)\,ds \\ &\quad + A^{-\frac{\beta}{2}}\int_0^t e^{-(t-s)A}P_NB(A^{\frac{\beta}{2}}y_N(s),\alpha_1(s))\,dW(s). \end{aligned} \tag{5.17}$$

LEMMA 5.10. *Let $y_0 \in X$, and for $\alpha \in \mathcal{U}_{ad}(0,+\infty;U)$ denote by $y(\cdot;y_0,\alpha)$ the solution of (5.5). Then, for all $T>0$, we have*

$$\lim_{N\to+\infty} \sup_{\alpha\in\mathcal{U}_{ad}(0,+\infty;U)} \left\|y_N(\cdot;y_0,\alpha)-y(\cdot;y_0,\alpha)\right\|_{M^2\left(0,T;D\left(A^{\frac{\beta}{2}}\right)\right)} = 0.$$

Proof. Let $\alpha \in \mathcal{U}_{ad}(0,+\infty;U)$, $y_N(\cdot)=y_N(\cdot;y_0,\alpha)$, and $y(\cdot)=y(\cdot;y_0,\alpha)$. Fix γ in $(0,1-\beta)$. Then, for $t\in(0,T)$, we have

$$\begin{aligned}
A^{\frac{\beta}{2}}\big(y_N(t)-y(t)\big) &= -A^{\frac{\beta}{2}}e^{-tA}Q_Ny_0 - Q_NA^{-\frac{\gamma}{2}}\int_0^t A^{\frac{\gamma}{2}}e^{-(t-s)A}Q_Nb\big(A^{\frac{\beta}{2}}y(s),\alpha_1(s)\big)\,ds\\
&\quad - Q_NA^{-\frac{\gamma}{2}}\int_0^t A^{\beta+\frac{\gamma}{2}}e^{-(t-s)A}Q_NC\alpha_2(s)\,ds\\
&\quad - Q_NA^{-\frac{\gamma}{2}}\int_0^t A^{\frac{\beta+\gamma}{2}}e^{-(t-s)A}A^{-\frac{\beta}{2}}Q_NB\big(A^{\frac{\beta}{2}}y(s),\alpha_1(s)\big)\,dW(s)\\
&\quad + \int_0^t e^{-(t-s)A}P_N\big[b\big(A^{\frac{\beta}{2}}y_N(s),\alpha_1(s)\big)-b\big(A^{\frac{\beta}{2}}y(s),\alpha_1(s)\big)\big]\,ds\\
&\quad + \int_0^t A^{\frac{\beta}{2}}e^{-(t-s)A}A^{-\frac{\beta}{2}}P_N\big[B\big(A^{\frac{\beta}{2}}y_N(s),\alpha_1(s)\big)\\
&\qquad\qquad - B\big(A^{\frac{\beta}{2}}y(s),\alpha_1(s)\big)\big]\,dW(s)
\end{aligned}$$

which yields, for a suitable $C_\gamma(T)>0$,

$$\begin{aligned}
&\mathbb{E}\big|A^{\frac{\beta}{2}}\big(y_N(t)-y(t)\big)\big|^2\\
&\quad\le C_\gamma(T)\left[\frac{1}{t^\beta}\big|Q_Ny_0\big|^2 + \big\|Q_NA^{-\frac{\gamma}{2}}\big\|^2\left(1+\int_0^t\left(1+\frac{1}{(t-s)^{\beta+\gamma}}\right)\mathbb{E}|A^{\frac{\beta}{2}}y(s)|^2ds\right)\right.\\
&\qquad\left.+\int_0^t\left(1+\frac{1}{(t-s)^\beta}\right)\mathbb{E}\big|A^{\frac{\beta}{2}}\big(y_N(s)-y(s)\big)\big|^2ds\right].
\end{aligned}$$

Since $A^{-\gamma/2}$ is compact, $\|Q_NA^{-\gamma/2}\|\to 0$ as $N\to+\infty$, and by using (5.8) we can rewrite the above inequality as follows:

$$\begin{aligned}
\mathbb{E}\big|A^{\frac{\beta}{2}}\big(y_N(t)-y(t)\big)\big|^2 &\le C_{\gamma,N}(T,y_0)\left(1+\frac{1}{t^\beta}\right)\\
&\quad+\int_0^t\left(1+\frac{1}{(t-s)^\beta}\right)\mathbb{E}\big|A^{\frac{\beta}{2}}\big(y_N(s)-y(s)\big)\big|^2\,ds,
\end{aligned}$$

where $C_{\gamma,N}(T,y_0)\to 0$ as $N\to+\infty$. Using Lemma 5.6 and arguing as in the proof of Proposition 5.7, we finally obtain, for a suitable $C>0$,

$$\mathbb{E}\big|A^{\frac{\beta}{2}}\big(y_N(t)-y(t)\big)\big|^2 \le C_{\gamma,N}(T,y_0)C\left(1+\frac{1}{t^\beta}\right), \tag{5.18}$$

so that

$$\sup_{\alpha\in\mathcal{U}_{ad}(0,+\infty;U)}\|y_N-y\|_{M^2\left(0,T;D\left(A^{\frac{\beta}{2}}\right)\right)}\to 0 \quad\text{as } N\to+\infty. \qquad \square$$

Proof of Theorem 5.8. We notice that under our assumptions (5.16) has a unique viscosity solution in $BUC(X_{-\eta})$ for every $\eta\in(0,1-\beta)$. To verify that the value

function (5.13) is the solution, we consider the approximating problems

$$\lambda u_N + \langle Ax, Du_N\rangle + G_H(x, Du_N, D^2u_N) = 0 \quad \text{in } X_N. \tag{5.19}$$

Equation (5.19) is the one used in the proof of Theorem 3.8 and it is easy to see that it is the equation in X_N corresponding to the control problem with evolution given by (5.17). Therefore, by the finite dimensional theory (see [24, 35, 38] for results and techniques that adapt to our situation to obtain the dynamic programming principle and Theorem 3.8), the function

$$u_N(y_0) = \inf_{\alpha\in\mathcal{U}_{ad}(0,+\infty;U)} \mathbb{E}\int_0^{+\infty} e^{-\lambda t} L\big(A^{\frac{\beta}{2}} y_N(t; y_0, \alpha), \alpha(t)\big)\, dt \tag{5.20}$$

is the unique viscosity solution of (5.19) in X_N and the dynamic programming principle holds for u_N, i.e., for every $y_0 \in X$, $T \geq 0$,

$$\begin{aligned} u_N(y_0) = \inf_{\alpha\in\mathcal{U}_{ad}(0,+\infty;U)} \mathbb{E}\Big\{ &\int_0^T e^{-\lambda t} L\big(A^{\frac{\beta}{2}} y_N(t; y_0, \alpha), \alpha(t)\big)\, dt \\ &+ e^{-\lambda T} u_N\big(y_N(T; y_0, \alpha)\big)\Big\}. \end{aligned} \tag{5.21}$$

Since for every $y_0 \in X$ $y_N(t; y_0, \alpha) = P_N y_N(t; P_N y_0, \alpha)$, extending u_N to X by putting $u_N(y) = u_N(P_N y)$ we obtain (5.20) and (5.19) for every $y_0 \in X$. Moreover, from the proof of existence of Theorem 3.8, we know that for every $\eta \in (0, 1-\beta)$ and $N \geq 1$,

$$\|u_N\|_\infty \leq \frac{K}{\lambda}, \qquad \big|u_N(x) - u_N(y)\big| \leq \sigma_\eta\big(|x-y|_{-\eta}\big) \tag{5.22}$$

for some modulus σ_η and $u_N \to \overline{u}$ uniformly on bounded sets, where $\overline{u}$ is the unique viscosity solution of (5.16). Therefore it remains to show that $u = \overline{u}$ which is an immediate consequence of the following lemma.

LEMMA 5.11. *u_N converges pointwise to u as $N \to \infty$.*

Proof. Let $y_0 \in X$, $N \in \mathbb{N}$. For every $T > 0$,

$$\begin{aligned} &\big|u_N(y_0) - u(y_0)\big| \\ &\quad\leq \sup_{\alpha\in\mathcal{U}_{ad}(0,+\infty;U)} \int_0^T e^{-\lambda t}\mathbb{E}\,\omega_L\big(|A^{\frac{\beta}{2}}(y_N(t; y_0, \alpha) - y(t; y_0, \alpha))|\big)\, dt + 2C_L\frac{e^{-\lambda T}}{\lambda}. \end{aligned}$$

Let $\varepsilon > 0$. There exists $T_\varepsilon > 0$ such that, for all N,

$$\begin{aligned} &\big|u_N(y_0) - u(y_0)\big| \\ &\quad\leq \sup_{\alpha\in\mathcal{U}_{ad}(0,+\infty;U)} \int_0^{T_\varepsilon} e^{-\lambda t}\mathbb{E}\,\omega_L\big(|A^{\frac{\beta}{2}}(y_N(t; y_0, \alpha) - y(t; y_0, \alpha))|\big)\, dt + \varepsilon. \end{aligned}$$

Now, by using the properties of moduli and Schwarz inequality, we know that for all $\sigma > 0$, there exists $C_\sigma > 0$ such that

$$\begin{aligned} &\int_0^{T_\varepsilon} e^{-\lambda t}\mathbb{E}\,\omega_L\big(|A^{\frac{\beta}{2}}(y_N(t; y_0, \alpha) - y(t; y_0, \alpha))|\big)\, dt \\ &\qquad\leq \frac{\sigma}{\lambda} + C_\sigma\sqrt{T_\varepsilon}\big\|y_N(\cdot; y_0, \alpha) - y(\cdot; y_0, \alpha)\big\|_{M^2\left(0,T_\varepsilon;D\left(A^{\frac{\beta}{2}}\right)\right)} \end{aligned}$$

for all $N \in \mathbb{N}$ and all $\alpha \in \mathcal{U}_{ad}(0,+\infty;U)$. By letting N go to infinity, then σ to 0, and finally ε to 0, we conclude the proof of the Lemma 5.11. □

To finish the proof of Theorem 5.8 we only have to show the dynamic programming principle for u. By (5.21), we have

$$\begin{aligned}\Big|u_N(y_0) - \inf_{\alpha\in\mathcal{U}_{ad}} \mathbb{E}\Big\{ \int_0^T e^{-\lambda t} L\big(A^{\frac{\beta}{2}} y(t;y_0,\alpha),\alpha(t)\big)\,dt + e^{-\lambda T} u\big(y(T;y_0,\alpha)\big)\Big\}\Big| \\ \le \sup_{\alpha\in\mathcal{U}_{ad}} \mathbb{E}\int_0^T e^{-\lambda t}\omega_L\big(|A^{\frac{\beta}{2}}(y_N(t;y_0,\alpha) - y(t;y_0,\alpha))|\big)\,dt \\ + e^{-\lambda T}\sup_{\alpha\in\mathcal{U}_{ad}} \mathbb{E}\big|u_N\big(y_N(T;y_0,\alpha)\big) - u\big(y(T;y_0,\alpha)\big)\big|.\end{aligned}$$

As in the proof of Lemma 5.11, the first term of the right-hand side converges to 0 when N goes to infinity. For the second term, we proceed as follows:

$$\begin{aligned}\mathbb{E}\big|u_N\big(y_N(T;y_0,\alpha)\big) - u\big(y(T;y_0,\alpha)\big)\big| \le \mathbb{E}\big|u_N\big(y_N(T;y_0,\alpha)\big) - u_N\big(y(T;y_0,\alpha)\big)\big| \\ + \mathbb{E}\big|u_N\big(y(T;y_0,\alpha)\big) - u\big(y(T;y_0,\alpha)\big)\big|.\end{aligned}$$

The first term of the right-hand side converges uniformly to 0 when N goes to infinity by (5.18) and (5.22). It remains to prove that

$$\sup_{\alpha\in\mathcal{U}_{ad}(0,+\infty;U)} \mathbb{E}\big|u_N\big(y(T;y_0,\alpha)\big) - u\big(y(T;y_0,\alpha)\big)\big|$$

goes to 0 when N goes to infinity. By Proposition 5.7, estimate (5.7), $\mathbb{E}|y(T;y_0,\alpha)|^2$ is bounded by a constant $C(T,y_0) > 0$ which does not depend on $\alpha \in \mathcal{U}_{ad}(0,+\infty;U)$. Hence, for all $R > 0$,

$$\mathbf{P}\big\{\big|y(T;y_0,\alpha)\big| > R\big\} \le \frac{C(T,y_0)}{R^2}.$$

Let $\varepsilon > 0$ and choose $R_\varepsilon > 0$ sufficiently large so that this probability will be smaller than ε. Then

$$\sup_{\alpha\in\mathcal{U}_{ad}(0,+\infty;U)} \mathbb{E}\big|u_N\big(y(T;y_0,\alpha)\big) - u\big(y(T;y_0,\alpha)\big)\big| \le \frac{2C_L}{\lambda}\varepsilon + \sup_{|y|\le R_\varepsilon}\big|u_N(y) - u(y)\big|.$$

We conclude by letting $N \to +\infty$ since ε was arbitrary. □

5.4. Examples of stochastic boundary control problems. We now present examples of problems where the operator A is the Laplacian with Dirichlet boundary conditions, reminding that our results also hold true for the case of Laplacian with Neumann boundary conditions (which is in some sense easier to treat since it gives rise to a lower exponent β in the Bellman equation). For further examples we refer to the book of Lasiecka and Triggiani [36]. This book deals with deterministic boundary control problems. However, it can be checked that our results apply to suitable stochastic perturbations of examples belonging to the "first abstract class" treated in the book.

We refer the reader to the papers of Tessitore [47, 48] and Duncan, Maslowski, and Pasik-Duncan [19, 20] for more about stochastic boundary control problems and to Da Prato and Ichikawa [16], Flandoli [22, 23], Fattorini [21], and Zabczyk [49] for deterministic boundary control problems. However, to the best of our knowledge, the results presented in our paper are the first of this kind in the literature.

Example. Let $\Omega \subset \mathbb{R}^N$ be an open, connected, and bounded set with smooth boundary. Consider, as in the introduction, the following stochastic controlled PDE

(5.23)
$$\begin{cases} \dfrac{\partial x}{\partial t}(t,\xi) = \Delta_\xi x(t,\xi) \\ \qquad\qquad +f_1\left(x(t,\xi),\alpha_1(t,\xi)\right) + f_2\left(x(t,\xi),\alpha_1(t,\xi)\right)\dot{W}_Q(t,\xi) \quad \text{in } (0,\infty)\times\Omega, \\ x(0,\xi) = x_0(\xi) \qquad \text{on } \Omega, \\ x(t,\xi) = \alpha_2(t,\xi) \quad \text{on } (0,\infty)\times\partial\Omega, \end{cases}$$

where W_Q is a Wiener process with values in $L^2(\Omega)$ and with covariance operator Q, $x_0 \in L^2(\Omega)$, and the controls are $\alpha_1 \in M^0(0,\infty;L^2(\Omega))$, $\alpha_2 \in M^0(0,\infty;L^2(\partial\Omega))$, with $\|\alpha_2(s)\|_{L^2(\partial\Omega)} \le C_1$ for a given $C_1 > 0$. Moreover, $f_1 : \mathbb{R}^2 \to \mathbb{R}$ is a continuous function, Lipschitz continuous in the first variable, uniformly with respect to the second one. Assumptions on f_2 and Q will be specified later. Here we have two control functions. The first, indicated by α_1, is the "distributed" control, while the other, indicated by α_2, is the "boundary" control.

Using a standard procedure which can be found, e.g., in [3] and [10] for the deterministic case (the stochastic case does not need any substantial change), problem (5.23) may be rewritten in an abstract form by finding an equation similar to the one treated in section 5.1. Let $X = L^2(\Omega)$ be the state space, let $U_1 = L^2(\Omega), U_2 = L^2(\partial\Omega)$ be the spaces of control parameters, and let $\tilde{U}_2$ be the closed ball of U_2 of radius C_1. Let $U = U_1 \times \tilde{U}_2$. Then (5.23) becomes

(5.24)
$$\begin{cases} dx(s) = \left[-Ax(s) + b\big(x(s),\alpha_1(s)\big) + A\overline{C}\alpha_2(s)\right]ds + B\big(x(s),\alpha_1(s)\big)\,dW(s), \\ \qquad 0 < s < +\infty, \\ x(0) = x_0, \quad x_0 \in X, \end{cases}$$

where $\alpha \in M^0(0,\infty;U)$, A is the Laplace operator with zero Dirichlet boundary conditions, and $b : X \times U_1 \to X$ and $B : X \times U_1 \to \mathcal{L}(X)$ are defined as

$$\begin{aligned} b(x,\alpha_1)(\xi) &= f_1\big(x(\xi),\alpha_1(\xi)\big), \\ \big[B(x,\alpha_1)y\big](\xi) &= f_2\big(x(\xi),\alpha_1(\xi)\big)\big[[Q^{\frac{1}{2}}y](\xi)\big] \quad \text{for all } \xi \in \Omega,\ y \in L^2(\Omega), \end{aligned}$$

and finally $\overline{C} : U_2 \to X$ is a continuous linear operator, the Dirichlet operator.

Since f_1 is Lipschitz continuous in the first variable, uniformly with respect to the second one, b satisfies Hypothesis 5.1(ii). As for the diffusion term, we assume here that f_2 and Q are such that Hypotheses 5.1(iv) and (v) are satisfied. This is true, for example, if $N = 1, Q = I$, and f_2 is Lipschitz continuous in the first variable, uniformly with respect to the second one. Indeed, taking the orthonormal basis $\{e_k\}$ of eigenvectors of A (which is uniformly bounded in $L^\infty(\Omega)$), we have, for suitable $C_0 > 0$,

$$\left\|A^{-\frac{\beta}{2}}B(x,\alpha_1)\right\|^2_{\mathcal{L}_2(X)} = \sum_{k=1}^{+\infty}\left|A^{-\frac{\beta}{2}}B(x,\alpha_1)e_k\right|^2 = C_0\sum_{k=1}^{+\infty}\sum_{h=1}^{+\infty}h^{-2\beta}\big\langle B(x,\alpha_1)e_k,e_h\big\rangle^2.$$

Now, since B is self-adjoint,

$$\begin{aligned}\left\|A^{-\frac{\beta}{2}}B(x,\alpha_1)\right\|^2_{\mathcal{L}_2(X)} &= C_0\sum_{h=1}^{+\infty} h^{-2\beta}\left|B(x,\alpha_1)e_h\right|^2\\ &= C_0\sum_{h=1}^{+\infty} h^{-2\beta}\int_\Omega \left|f_2(x(\xi),\alpha_1(\xi))e_h(\xi)\right|^2 d\xi\end{aligned}$$

so, since $\|e_h(\xi)\|_\infty \le C_1$ for some positive constant C_1, we easily get that, for suitable $C_2 > 0$,

$$\left\|A^{-\frac{\beta}{2}}B(x,\alpha_1)\right\|^2_{\mathcal{L}_2(X)} \le C_2\big[1+|x|^2\big],$$

since $\beta > 3/4$. We can check similarly the other assumptions in Hypotheses 5.1(iv), (v). For $N \ge 2$ we need stronger assumptions on f_2. Furthermore, since Im $\overline{C} \subset D(A^\eta)$ (see, e.g., [40]) for $\eta \in (0, \frac{1}{4})$, setting $C = A^\eta\overline{C}$ we have that C is a continuous linear operator and $A\overline{C}\alpha_2 = A^\beta C\alpha_2$ for $\beta \in (\frac{3}{4}, 1)$, $\beta = 1-\eta$, so that we obtain an equation that falls into the class discussed in section 5.1.

Equation (5.24) makes sense only in the mild form, as explained in section 5.1, due to the presence of the unbounded term $A^\beta C\alpha_2$. Given $\lambda > 0$ and $l \in BUC(\mathbb{R}^2)$, we now define

$$L(x,\alpha_1) = \int_\Omega l\big(x(\xi),\alpha_1(\xi)\big)\,d\xi,$$

(assumptions (5.10) are then satisfied) and consider the problem of minimizing the cost functional

$$J(x_0;\alpha_1,\alpha_2) = \mathbb{E}\int_0^{+\infty} e^{-\lambda t}L\left(x(t;x_0,\alpha_1,\alpha_2),\alpha_1(t)\right)\,dt, \tag{5.25}$$

where $x(\cdot\,;x_0,\alpha)$ is the mild solution of the stochastic differential equation 5.24. The cost functional can be more general. The associated Hamilton–Jacobi–Bellman equation is now (1.1) where the Hamiltonian H is given by

$$\begin{aligned}&H(x,p,S)\\ &= \sup_{\alpha=(\alpha_1,\alpha_2)\in U}\left\{-\frac{1}{2}\mathrm{Tr}\ B^*(x,\alpha_1)SB(x,\alpha_1) - \langle b(x,\alpha_1),p\rangle - \langle C\alpha_2, A^\beta p\rangle - L(x,\alpha_1)\right\}.\end{aligned}$$

When f_2 is constant (and thus so is B) Hypothesis 5.1 holds in the following case.

HYPOTHESIS 5.12. *Let $\{e_k\}$ be an orthonormal basis in X, let B be linear and independent of x and α_1. Let A and B satisfy*

$$Ae_k = -\alpha_k e_k, \qquad BB^*e_k = \lambda_k e_k, \quad k \in \mathbb{N},$$

where $\{\alpha_k\}$ is a sequence of positive numbers increasing to $+\infty$ while $\{\lambda_k\}$ is a bounded sequence of nonnegative real numbers.

PROPOSITION 5.13. *Assume that Hypothesis* 5.12 *holds. Then Hypothesis* 5.1 *is satisfied if*

$$\sum_{k=1}^{\infty}\frac{\lambda_k}{\alpha_k^\beta} < +\infty. \tag{5.26}$$

For $\Omega = [0,\pi]^N$, the Laplace operator A satisfies Hypothesis 5.12 by taking, for $(n_1,\dots,n_N) \in \mathbb{N}^N$,

$$e_{n_1,\dots,n_N}(\xi) = \left(\frac{2}{\pi}\right)^{\frac{N}{2}} \sin n_1\xi_1 \cdots \sin n_N\xi_N \qquad \text{and} \qquad \alpha_{n_1,\dots,n_N}(\xi) = n_1^2 + \cdots + n_N^2$$

so that, by ordering the eigenvalues, we obtain

$$\alpha_k \approx k^{\frac{2}{N}} \quad \text{as } k \to +\infty.$$

Therefore condition 5.26 is fulfilled if for some small $\varepsilon > 0$ $\lambda_k = o(k^{\frac{2\beta}{N}-1-\varepsilon})$. When B is invertible this is possible only for $N = 1$. We remark that, differently from the case studied in [27], we can allow B to vary in a wider class, in particular B can have finite rank.

5.5. Concluding remarks. Second order Hamilton–Jacobi equations with second order terms being trace class have been studied in various papers (see, e.g., [2, 15, 28]). In some cases (e.g., when the second order term is linear and hypoelliptic) it is possible to prove existence and uniqueness of differentiable solutions, while in the general fully nonlinear case a theory of viscosity solutions is available (see [31, 39, 43]).

When the second order terms are not trace class there are no results about existence and uniqueness of viscosity solutions. However, in the papers [4, 7, 25] and [26] for the evolution case and [27] for the stationary case, it is proved that in some special cases these equations can be solved by using regularizing properties of transition semigroups associated with suitable Markov processes: in this way existence and uniqueness of differentiable solutions is proved and the theory finds applications to stochastic control problems allowing to prove existence of feedback controls. The stationary case is also treated in [11] where existence and uniqueness of a certain kind of weak solution is proved using variational methods in Gauss–Sobolev spaces. The assumptions in [11] are different from ours. The question arises naturally if it is possible to extend the theory of viscosity solutions to include the equations studied in [11, 27], and if the viscosity solution coincides with solutions proposed there. An answer to this question, though incomplete, is given by Theorem 5.14 below. The types of equations studied in [27] and here overlap, but neither of them contains the other. The main advantage of our approach is that it allows us to handle fully nonlinear equations. However, we pay the price of rather strong assumptions on the operator A.

To be more precise, we recall that in [27] the equation

$$\lambda v(x) = \frac{1}{2}\big[QD^2v(x)\big] + \big\langle Ax + F(x), Dv(x)\big\rangle - H_0\big(Dv(x)\big) + \psi(x), \quad x \in X \tag{5.27}$$

was considered (see [27] for the precise setting). By combining the results and techniques of this paper and [27] the following result holds, which we state here without a proof.

THEOREM 5.14. *Assume that $-A$ satisfies Hypothesis* 3.1 *and moreover that the operator $A^{-\beta}Q$ is nuclear for some $\beta \in (0,1)$. Assume also that F is Lipschitz continuous and H_0 is uniformly continuous on bounded subsets of X. Then, for every $\lambda > 0$ and $\psi \in BUC(X)$, there exists a unique $BUC(X_{-\eta})$ viscosity solution (for every $\eta < 1$) of 5.27. Moreover, if the hypotheses of Theorems* 3.3 *and* 3.11 *in* [27]

hold then the viscosity solution is Fréchet differentiable with respect to x and coincides with the mild-strong one of the above theorems.

It is proved in [27, Theorem 5.7] that the solution v can be used to construct an optimal control in feedback form given by

$$\alpha^*(s) = DH_0\big(Dv\big(y^*(t)\big)\big),$$

where $y^*(\cdot)$ solves a closed-loop equation. To our knowledge this is the only nontrace class infinite dimensional case when the solution is regular enough to allow construction of an optimal control. Moreover there are no general results available about optimality conditions (see [37] for the finite dimensional case). It is an open problem if such results can be proved in cases of boundary control or stronger nonlinearities in the Bellman equation, i.e., the cases tractable only by viscosity solutions introduced here. We plan to come back to these questions in the future.

Acknowledgments. The authors would like to thank Prof. G. Da Prato for his hospitality and for providing excellent conditions for work in Scuola Normale Superiore di Pisa. We also wish to thank Prof. J. Zabczyk for helpful suggestions and remarks on the paper.

REFERENCES

[1] S. Albeverio and M. Röckner, *Stochastic differential equations in infinite dimensions: solutions via Dirichlet forms*, Probab. Theory Related Fields, 89 (1991), pp. 347–86.

[2] V. Barbu and G. Da Prato, *Hamilton–Jacobi Equations in Hilbert Spaces*, Res. Notes Math., Pitman, Boston, 1983.

[3] A. Bensoussan, G. Da Prato, M. C. Delfour, and S. K. Mitter, *Representation and Control of Infinite Dimensional Systems*, Birkhäuser, Boston, Cambridge, MA, 1992.

[4] P. Cannarsa and G. Da Prato, *Direct solution of a second order Hamilton–Jacobi equation in Hilbert spaces*, in Stochastic Partial Differential Equations and Applications, G. Da Prato and L. Tubaro, eds., Pitman Res. Notes in Math. 268, Longman, Harlow, UK, 1992, pp. 72–85.

[5] P. Cannarsa and G. Da Prato, *On a functional analysis approach to parabolic equations in infinite dimensions*, J. Funct. Anal., 118 (1993), pp. 22–42.

[6] P. Cannarsa and G. da prato, *Second order elliptic and parabolic equations with infinitely many variables*, in Proceedings of the Amer. Math. Soc. Conference in honor of Lax and Nirenberg, Venice, 1996.

[7] P. Cannarsa and G. Da Prato, *Second-order Hamilton–Jacobi equations in infinite dimensions*, SIAM J. Control Optim., 29 (1991), pp. 474–492.

[8] P. Cannarsa, F. Gozzi, and H. M. soner, *A dynamic programming approach to nonlinear boundary control problems of parabolic type*, J. Funct. Anal., 117 (1992), pp. 25–61.

[9] P. Cannarsa and M. E. Tessitore, *Cauchy problem for the dynamic programming equation of boundary control*, in Proceedings IFIP Workshop on Boundary Control and Boundary Variation, Marcel Dekker, New York, Basel, 1993.

[10] P. Cannarsa and M. E. Tessitore, *Infinite-dimensional Hamilton–Jacobi equations and Dirichlet boundary control problems of parabolic type*, SIAM J. Control Optim., 34 (1996), pp. 1831–1847.

[11] P. L. Chow and J. L. Menaldi, *Infinite dimensional Hamilton–Jacobi–Bellman equations in Gauss–Sobolev spaces*, Nonlinear Anal., 27 (1997), pp. 415–426.

[12] M. G. Crandall, H. Ishii, and P.-L. Lions, *User's guide to viscosity solutions of second order partial differential equations*, Bull. Amer. Math. Soc., 27 (1992), pp. 1–67.

[13] M. G. Crandall, M. kocan, and A. Święch, *On partial sup-convolutions, a lemma of P.-L. Lions and viscosity solutions in Hilbert spaces*, Adv. Math. Sci. Appl., 3 (1993/4), pp. 1–15.

[14] M. G. Crandall and P.-L. Lions, *Hamilton–Jacobi equations in infinite dimensions. Part* I: *Uniqueness of viscosity solutions*, J. Funct. Anal., 62 (1985), pp. 379–396. *Part* II: *Existence of viscosity solutions*, J. Funct. Anal., 65 (1986), pp. 368–405. *Part* III, J. Funct. Anal., 68 (1986), pp. 214–247. *Part* IV: *Hamiltonians with unbounded linear terms*, J. Funct. Anal., 90 (1990), pp. 237–283. *Part* V: *Unbounded linear terms and B-continuous*

solutions, J. Funct. Anal., 97 (1991), pp. 417–465. *Part* VI: *Nonlinear A and Tataru's method refined*, Evolution Equations, Control Theory, and Biomathematics (Han sur Lesse, 1991), Lecture Notes in Pure and Appl. Math. 155, Dekker, New York, 1994, pp. 51–89. *Part* VII: *The HJB equation is not always satisfied*, J. Funct. Anal., 125 (1994), pp. 111–148.

[15] G. Da Prato, *Some results on Bellman equation in Hilbert spaces*, SIAM J. Control Optim., 23 (1985), pp. 61–71.

[16] G. Da Prato and A. Ichikawa, *Riccati equations with unbounded coefficient*, Ann. Mat. Pura Appl., 140 (1985), pp. 209–221.

[17] G. Da Prato and J. Zabczyk, *Stochastic Equations in Infinite Dimensions*, Encyclopedia of Mathematics and Its Applications, Cambridge University Press, Cambridge, UK, 1992.

[18] G. Da Prato and J. Zabczyk, *Ergodicity for Infinite Dimensional Systems*, London Math. Soc. Lecture Note Ser. 229, Cambridge University Press, Cambridge, UK, 1996.

[19] T. E. Duncan, B. Maslowski, and B. Pasik-Duncan, *Adaptive boundary and point control of linear stochastic distributed parameter systems*, SIAM J. Control Optim., 32 (1994), pp. 648–672.

[20] T. E. Duncan, B. Maslowski, and B. Pasik-Duncan, *Ergodic boundary/point control of stochastic semilinear systems*, SIAM J. Control Optim., 36 (1998), pp. 1020–1047.

[21] H. O. Fattorini, *Boundary control systems*, SIAM J. Control Optim., 6 (1968), pp. 349–385.

[22] F. Flandoli, *Riccati equations arising in a boundary control problem with distributed parameters*, SIAM J. Control Optim., 22 (1984), pp. 76–86.

[23] F. Flandoli, *A counterexample in the boundary control of parabolic systems*, Appl. Math. Lett., 3 (1990), pp. 47–50.

[24] W. H. Fleming and H. M. Soner, *Controlled Markov Processes and Viscosity Solutions*, Springer-Verlag, Berlin, New York, 1993.

[25] F. Gozzi, *Regularity of solutions of a second order Hamilton–Jacobi equation and application to a control problem*, Comm. Partial Differential Equations, 20 (1995), pp. 775–826.

[26] F. Gozzi, *Global regular solutions of second order Hamilton–Jacobi equations in Hilbert spaces with locally Lipschitz nonlinearities*, J. Math. Anal. Appl., 198 (1996), pp. 399–443.

[27] F. Gozzi and E. Rouy, *Regular solutions of second order stationary Hamilton–Jacobi equations*, J. Differential Equations 130 (1996), pp. 201–234.

[28] T. Havarneanu, *Existence for the dynamic programming equation of control diffusion processes in Hilbert space*, Nonlinear Anal., 9 (1985), pp. 619–629.

[29] D. Henry, *Geometric Theory of Semilinear Parabolic Equations*, Lecture Notes in Math., 840, Springer-Verlag, Berlin, 1981.

[30] H. Ishii, *On uniqueness and existence of viscosity solutions of fully nonlinear second-order elliptic PDE's*, Comm. Pure Appl. Math., 42 (1989), pp. 15–45.

[31] H. Ishii, *Viscosity solutions for a class of Hamilton–Jacobi equations in Hilbert spaces*, J. Funct. Anal., 105 (1992), pp. 301–341.

[32] H. Ishii, *Viscosity solutions of nonlinear second-order partial differential equations in Hilbert spaces*, Comm. Partial Differential Equations, 18 (1993), pp. 601–651.

[33] G. Jona Lasinio and S. K. Mitter, *On the Stochastic Quantization of Field Theory*, Pitman, Boston, 1982.

[34] M. Kocan and A. Święch, *Second order unbounded parabolic equations in separated form*, Studia Math., 115 (1995), pp. 291–310.

[35] N. V. Krylov, *Controlled Diffusion Processes*, Springer-Verlag, New York, 1980.

[36] I. Lasiecka and R. Triggiani, *Differential and Algebraic Riccati Equations with Application to Boundary/Point Control Problems: Continuous Theory and Approximation Theory*, Lecture Notes in Control and Inform. Sci., 164, Springer Verlag, Berlin, 1991.

[37] X. Li, J. Yong, and X. Y. Zhou, *Stochastic verification theorems within the framework of viscosity solutions*, SIAM J. Control Optim., 35 (1997), pp. 243–253.

[38] P.-L. Lions, *Optimal control of diffusion processes and Hamilton–Jacobi–Bellman equations. Part* 1: *The dynamic programming principle and applications. Part* 2: *Viscosity solutions and uniqueness*, Comm. Partial Differential Equations, 8 (1983), pp. 1101–1174 and pp. 1229–1276.

[39] P.-L. Lions, *Viscosity solutions of fully nonlinear second-order equations and optimal stochastic control in infinite dimensions. Part* I: *The case of bounded stochastic evolution*, Acta Math., 161 (1988), pp. 243–278. *Part* II: *Optimal control of Zakai's equation*, in Stochastic Partial Differential Equations and Applications, Lecture Notes in Math. 1390, G. Da Prato and L. Tubaro, eds. Springer-Verlag, Berlin, 1989, pp. 147–170. *Part* III: *Uniqueness of viscosity solutions for general second order equations*, J. Funct. Anal., 86 (1989), pp. 1–18.

[40] J.-L. Lions and E. Magenes, *Problèmes aux limites non homogènes et applications* II, Dunod,

Paris, 1968.
[41] A. Pazy, *Semigroups of Linear Operators and Applications to Partial Differential Equations*, Springer-Verlag, Berlin, New York, 1983.
[42] H. M. Soner, *On the Hamilton–Jacobi–Bellman equations in Banach spaces*, J. Optim. Theory Appl., 57 (1988), pp. 429–437.
[43] A. Święch, *Unbounded second order partial differential equations in infinite dimensional Hilbert spaces*, Comm. Partial Differential Equations, 19 (1994), pp. 1999–2036.
[44] D. Tataru, *Viscosity solutions for Hamilton–Jacobi equations with unbounded nonlinear terms*, J. Math. Anal. Appl., 163 (1992), pp. 345–392.
[45] D. Tataru, *Viscosity solutions for Hamilton–Jacobi equations with unbounded nonlinear term: a simplified approach*, J. Differential Equations, 111 (1994), pp. 123–146.
[46] D. Tataru, *Viscosity solutions for the dynamic programming equations*, Appl. Math. Optim., 25 (1992), pp. 109–126.
[47] G. Tessitore, *Linear quadratic optical control for a stochastic system with control on the boundary and hyperbolic dynamics*, J. Math. Systems Estim. Control, 2 (1992), pp. 717–744.
[48] G. Tessitore, *Linear quadratic boundary control for an age equation perturbed by noise*, in Proceedings Trento, Dekker, New York, 1994, pp. 197–213.
[49] J. Zabczyk, *Stabilization of boundary control systems*, in International Symposium on Systems Optimization and Analysis, Lecture Notes in Control and Inform. Sci. 14, Springer-Verlag, Berlin, New York, 1978, pp. 321-332.

SIAM J. CONTROL OPTIM.
Vol. 38, No. 2, pp. 431–446

A MODIFIED FORWARD-BACKWARD SPLITTING METHOD FOR MAXIMAL MONOTONE MAPPINGS*

PAUL TSENG†

Abstract. We consider the forward-backward splitting method for finding a zero of the sum of two maximal monotone mappings. This method is known to converge when the inverse of the forward mapping is strongly monotone. We propose a modification to this method, in the spirit of the extragradient method for monotone variational inequalities, under which the method converges assuming only the forward mapping is (Lipschitz) continuous on some closed convex subset of its domain. The modification entails an additional forward step and a projection step at each iteration. Applications of the modified method to decomposition in convex programming and monotone variational inequalities are discussed.

Key words. maximal monotone mapping, forward-backward splitting method, extragradient method, variational inequality, convex programming, decomposition

AMS subject classifications. 49M45, 90C25, 90C33

PII. S0363012998338806

1. Introduction.

Let $\mathcal{H}$ be a real Hilbert space with inner product $\langle x, y\rangle$ and induced norm $\|x\| = \sqrt{\langle x, x\rangle}$ for $x, y \in \mathcal{H}$. (In the case of $\mathcal{H} = \Re^n$, the space of n-dimensional real column-vectors, $\langle \cdot, \cdot \rangle$ is the Euclidean inner product.) A set-valued mapping (also called "operator") T on $\mathcal{H}$, which is written as $T : \mathcal{H} \rightrightarrows \mathcal{H}$, associates each point $x \in \mathcal{H}$ with a subset $T(x)$ of $\mathcal{H}$. Denote $\mathrm{dom}T = \{x \in \mathcal{H} : T(x) \neq \emptyset\}$. The mapping T is called *monotone* if

$$\langle x - x', y - y' \rangle \geq 0 \qquad \text{for all } x, x' \in \mathrm{dom}T, y \in T(x), y' \in T(x'),$$

and T is called *maximal monotone* if its graph

$$\mathrm{gph}T = \{(x, y) \in \mathcal{H} \times \mathcal{H} : y \in T(x)\}$$

is not properly contained in the graph of any other monotone mapping on $\mathcal{H}$. The inverse mapping T^{-1} defined by $T^{-1}(y) = \{x \in \mathcal{H} : y \in T(x)\}$ is, by symmetry, maximal monotone on $\mathcal{H}$ whenever T is. Maximal monotone mappings have been well studied by Minty, Moreau, Rockafellar, and others (see [4, 36, 38, 61], [47, Chap. 12], and references therein). One well-known example of such a mapping is $T = \partial f$, where ∂f is the subdifferential of a proper closed convex function $f : \mathcal{H} \mapsto (-\infty, \infty]$ [43]. Then, $0 \in \partial f(x)$ if and only if x is a minimizer of f. Another example is $T = F + N_C$, where C is a nonempty closed convex set in $\mathcal{H}$, F is a maximal monotone mapping that is single-valued and continuous on C, and N_C is the normal cone mapping $N_C(x) = \{y \in \mathcal{H} : \langle y, x' - x\rangle \leq 0 \text{ for all } x' \in C\}$ for $x \in C$ and is empty otherwise. Then, $0 \in F(x) + N_C(x)$ if and only if $x \in C$ satisfies the variational inequalities of $\langle F(x), x' - x\rangle \geq 0$ for all $x' \in C$.

*Received by the editors May 18, 1998; accepted for publication (in revised form) July 9, 1999; published electronically January 11, 2000. This research was supported by National Science Foundation grant CCR-9731273.

http://www.siam.org/journals/sicon/38-2/33880.html

†Department of Mathematics, University of Washington, Box 354350, Seattle, WA 98195 (tseng@math.washington.edu).

In general, we are interested in finding, for a given maximal monotone mapping T on $\mathcal{H}$, an $x \in \mathcal{H}$ satisfying $0 \in T(x)$. A classical method for doing this is the proximal point algorithm, proposed by Martinet [27, 28] and generalized by Rockafellar [44, 45]:

$$x^{k+1} = (I + \alpha_k T)^{-1}(x^k), \qquad k = 0, 1, \ldots,$$

where $\alpha_k > 0$. This method and its dual version in the context of convex programming, the method of multipliers of Hesteness and Powell, have been extensively studied (see [1, 15, 18, 21] and references therein) and are known to yield as special cases decomposition methods such as the method of partial inverses [48, 53], the Douglas–Rachford splitting method, and the alternating direction method of multipliers [8, 9, 10, 22]. In the case of $T = A + B$, where A and B are maximal monotone mappings on $\mathcal{H}$ with A single-valued on $\mathrm{dom}A \supset \mathrm{dom}B$, the forward-backward (F-B) splitting method

$$x^{k+1} = (I + \alpha_k B)^{-1}(I - \alpha_k A)(x^k), \qquad k = 0, 1, \ldots,$$

where $\alpha_k > 0$, was proposed by Lions and Mercier [22], by Passty [37], and, in a dual form for convex programming, by Han and Lou [16]. In the case where $B = N_C$, with C a nonempty closed convex set in $\mathcal{H}$, this method reduces to a projection method proposed by Sibony [49] for monotone variational inequalities and, in the further case where F is the gradient of a differentiable convex function, it reduces to a gradient projection method of Goldstein and of Levitin and Polyak [1]. This method was extensively analyzed by Mercier [30] and Gabay [13] and was further studied in [5, 6, 20, 33, 34, 39, 56, 57]. In particular, Mercier and Gabay showed that if A^{-1} is strongly monotone with modulus $\gamma > 0$, then the iterates x^k converge weakly to a solution, provided, α_k is constant and less than 2γ. And if in addition A is strongly monotone, then x^k converge strongly to the unique solution [30, p. 24], [13, Thm. 6.1] (also see [12, Chap. 6], [34], and, for the case of nonconstant α_k, [6, 57]). If A is Lipschitz continuous on $\mathrm{dom}A = \mathcal{H}$ and T is strongly monotone, Chen and Rockafellar [6] showed that the iterates x^k converge strongly at a linear rate to the unique solution provided α_k is less than some threshold depending on the Lipschitz constant and the modulus of strong monotonicity (also see Chen's thesis [5] and, for the case A is Lipschitz continuous and strongly monotone, see [30, p. 26], [39, Prop. 3.2]). Moreover, their method incorporates scaling and they derive an explicit formula for the convergence ratio in terms of the constants and the stepsize, from which a minimum-ratio stepsize was calculated. If neither A^{-1} nor T is assumed to be strongly monotone, Passty [37] showed that a weighted average of the iterates x^k, weighted by the stepsize α_k, converges weakly to a solution, provided $A(x^k)$ is bounded (it need not be single-valued even) and α_k is square summable but not absolutely summable. However, such ergodic convergence does not seem very useful in practice. For further discussions of splitting methods and applications, see [9, 11, 14, 30, 34] and references therein.

A nice feature of the F-B method is that the backward (i.e., proximal) step involves B only, so the "dense" portion of T can be put into A to facilitate problem decomposition. And, as a referee noted, the splitting is even more attractive if in addition any ill-conditioned portion of T can be put into B. In contrast, other splitting methods have backward step(s) that involve A also, possibly limiting the level of problem decomposition achievable. Some studies of this in the context of discrete-time optimal control problems are given in [5, 9, 48, 56]. (Of course, problem decomposition is not the only consideration. Convergence rate is another.) On the other hand, the

F-B method has the drawback that it requires either A^{-1} to be strongly monotone, implying A is Lipschitz continuous on $\mathrm{dom}A = \mathcal{H}$ (see [4, Cor. 2.4], [41]), or A to be Lipschitz continuous on $\mathrm{dom}A = \mathcal{H}$ and T to be strongly monotone. These requirements limit the choice of the splitting $T = A + B$ and rule out the kind of applications considered in section 4. Furthermore, choosing a good stepsize may be difficult since it entails estimating the modulus of A^{-1} or the Lipschitz constant of A and the modulus of A or B.

In this paper, we propose a simple modification to the F-B method that removes the requirement of A^{-1} or T being strongly monotone for convergence. The modification is motivated by the extragradient method of Korpelevich [19] for monotone variational inequalities, which modifies the projection method of Sibony by performing an additional forward step and a projection step at each iteration. By adaptively choosing the stepsize, this method has been shown to converge for monotone continuous mappings in the case $\mathcal{H} = \Re^n$ [17, 26, 54]. Our proposed modified method is in the same spirit as the extragradient method, performing an additional forward step and projection step onto some closed convex set $X \subset \mathrm{dom}A$ at each iteration (see (2.3)) and using an adaptive stepsize rule (see (2.4)). The modified method, when specialized to the case of variational inequalities, coincides with the extragradient method when the problem is unconstrained but otherwise seems to be new (see Example 2 in section 4). We show that if (i) A is Lipschitz continuous on $X \cup \mathrm{dom}B$ or if (ii) A is locally (in the weak topology) uniformly continuous on $X \subset \mathrm{dom}B$ and $x \mapsto \min_{w \in T(x)} \|w\|$ is locally (in the weak topology) bounded on X, then the iterates generated by this method converge weakly to a solution. Under additional assumptions on T, namely, T is strongly monotone or T^{-1} has a (local) Lipschitzian property, linear rate of convergence is also shown (see Theorem 3.4). Applications of the method to decomposition in convex programming and variational inequalities are discussed in section 4.

Throughout this paper, we denote by I either the identity mapping on $\mathcal{H}$ or the identity matrix. By a closed set C in $\mathcal{H}$, we mean C is closed in the weak topology. For any nonempty closed convex set $X \subset \mathcal{H}$, we denote the nearest-point projection of $y \in \mathcal{H}$ onto X by $P_X[y] = \arg\min_{x \in X} \|x - y\|$. A mapping $S : \mathcal{H} \rightrightarrows \mathcal{H}$ is said to be strongly monotone with modulus $\gamma > 0$ if

$$\langle x - x', y - y' \rangle \geq \gamma \|x - x'\|^2 \qquad \text{for all } x, x' \in \mathrm{dom}S, y \in S(x), y' \in S(x').$$

A mapping $A : \mathcal{H} \rightrightarrows \mathcal{H}$ that is single-valued on some set $Y \subset \mathcal{H}$ is said to be Lipschitz continuous on a set X with constant $\lambda \geq 0$, where $X \subset Y$, if

$$\|A(x) - A(x')\| \leq \lambda \|x - x'\| \qquad \text{for all } x, x' \in X$$

and is said to be locally uniformly continuous on $X \subset Y$ if

$$\|A(x^k) - A(y^k)\| \to 0 \text{ whenever } x^k \in X, y^k \in Y \text{ converge weakly and } \|x^k - y^k\| \to 0. \tag{1.1}$$

A function $\phi : X \mapsto \Re$, where $X \subset \mathcal{H}$, is said to be locally bounded on X if $\phi(x^k)$ is bounded whenever $x^k \in X$ converges weakly. (Here $\{x^k\}$ and $\{y^k\}$ are sequences.)

2. Method description. Let $A : \mathcal{H} \rightrightarrows \mathcal{H}$ and $B : \mathcal{H} \rightrightarrows \mathcal{H}$ be maximal monotone mappings with A single-valued on $\mathrm{dom}A \supset \mathrm{dom}B$ and with $T^{-1}(0) \neq \emptyset$, where $T = A + B$. In this section we describe the modified F-B splitting method for finding an element of $T^{-1}(0)$. For convenience, we denote the F-B mapping by

$$J(x, \alpha) = (I + \alpha B)^{-1}(I - \alpha A)(x) \qquad \text{for all } x \in \mathrm{dom}A, \text{ for all } \alpha > 0. \tag{2.1}$$

It is well known (see Minty [32]) that $(I+\alpha B)^{-1}$ is a single-valued mapping from $\mathcal{H}$ to dom$B$. Thus, $J(\cdot,\alpha)$ is a single-valued mapping from dom$A$ to dom$B$.

Modified F-B splitting method. *Assume $X \subset \mathrm{dom}A$ is a closed convex set such that $X \cap T^{-1}(0) \neq \emptyset$ and either A is Lipschitz continuous on $X \cup \mathrm{dom}B$ or A is continuous (from the strong topology to the strong topology) on* dom$B$ *and $X \subset \mathrm{dom}B$. Choose any $x^0 \in X$. For $k = 0, 1, \ldots$, we generate x^{k+1} from x^k by choosing an $\alpha_k \in (0,\infty)$ and letting*

$$\bar{x}^k = J(x^k, \alpha_k), \tag{2.2}$$

$$x^{k+1} = P_X[\bar{x}^k - \alpha_k(A(\bar{x}^k) - A(x^k))]. \tag{2.3}$$

Since $J(\cdot,\alpha)$ maps dom$A$ to dom$B \subset$ dom$A$ and $x^0 \in X \subset$ dom$A$, an induction argument yields $x^k \in X$ and $\bar{x}^k \in$ dom$B$, so $A(x^k)$ and $A(\bar{x}^k)$ are nonempty for all $k = 0, 1, \ldots$. The projection onto X in (2.3) is needed to ensure that $A(x^k)$ is nonempty and that α_k can be chosen by the stepsize rule below (see the proof of Theorem 3.4(a)).

There are various choices for the set X. If A is Lipschitz continuous on a closed convex subset of dom$A$ that contains dom$B$, then we can choose X to be this set. If dom$B$ is closed, then a result of Minty [31] (also see [4, Rem. 2.1]) implies dom$B$ is convex and we can choose $X =$ dom$B$. This occurs, for example, when the constraints are explicit, so that $B = G + N_C$, with C a nonempty closed convex set in $\mathcal{H}$ and $G : \mathcal{H} \rightrightarrows \mathcal{H}$ a maximal monotone mapping with dom$G \supset C$. (See section 4 for specific choices of X in some applications.) In some cases, such as when dom$B$ is unbounded, there may be an advantage in choosing X to be a bounded subset of dom$B$. Also, we can more generally work with a dynamically changing set X^k, provided $(\cap_{k=0}^{\infty} X^k) \cap T^{-1}(0) \neq \emptyset$. This does not affect our convergence result and allows for X^k to be adjusted so to better approximate $T^{-1}(0)$. For example, if A is Lipschitz continuous on dom$A = \mathcal{H}$ with constant $\lambda \geq 0$, then following a cutting-plane approach of Solodov and Svaiter [50], we can choose X^k to be the half-space

$$X^k = \{x \in \mathcal{H} : \langle \bar{w}^k, x - \bar{x}^k \rangle \leq 0\},$$

where $\bar{w}^k = (x^k - \bar{x}^k)/\alpha_k - A(x^k) + A(\bar{x}^k)$. Using $\bar{w}^k \in T(\bar{x}^k)$ (see (3.12)) and monotonicity of T, it can be seen that $T^{-1}(0) \subset X^k$, and if $\alpha_k < 1/\lambda$, then $x^k \notin X^k$.

Choosing α_k requires some care, for it cannot be too large (or the method might diverge) nor can it be too small (or the convergence might be too slow). If A is Lipschitz continuous on $X \cup$ dom$B$, then α_k can be chosen to be a constant (see Theorem 3.4(a)). However, it is more practical to choose α_k dynamically using an Armijo–Goldstein-type stepsize rule. Specifically, we will choose α_k to be the largest $\alpha \in \{\sigma, \sigma\beta, \sigma\beta^2, \ldots\}$ satisfying

$$\alpha \|A(J(x^k,\alpha)) - A(x^k)\| \leq \theta \|J(x^k,\alpha) - x^k\|, \tag{2.4}$$

where $\beta \in (0,1)$ and $\theta \in (0,1)$ and $\sigma > 0$ are constants. As a referee noted, in the optimization context, the acceptance criterion (2.4) involves function gradients, not function values, which distinguishes it from the classical Armijo–Goldstein criteria, e.g., [1, pp. 20, 57]. We will show that (2.4) is satisfied by all α sufficiently small, so α_k is well defined (see Theorem 3.4(a)). Alternatively, we can choose α_k to be the largest $\alpha \in \{\alpha_{k-1}, \alpha_{k-1}\beta, \alpha_{k-1}\beta^2, \ldots\}$ satisfying (2.4), with α_{-1} chosen arbitrarily. The resulting α_k, though more conservative, is cheaper to find since typically $\alpha = \alpha_{k-1}$ will satisfy (2.4). Our convergence results below hold for this alternative stepsize rule

also. The above stepsize rules are motivated by, but are simpler than, those given in [59] for an alternating projection-proximal method. These stepsize rules contrast with those for F-B splitting methods, which require the stepsize to be less than a constant depending on the modulus of A^{-1} or the Lipschitz constant of A and the modulus of A or B, so the latter need to be known or estimated. Related stepsize rules in the context of projection-type methods for variational inequalities are discussed in [17, 26, 52, 54]. In the case where $\mathrm{dom}A = \mathrm{dom}B = \mathcal{H} = X$, the modified F-B method may be viewed as an instance of a hybrid approximate extragradient-proximal point algorithm proposed recently by Solodov and Svaiter [51].

3. Convergence analysis. In this section we analyze the convergence and the rate of convergence of the method in the previous section. We begin with the following lemma, showing that the algorithmic mapping for the modified F-B method has a nonexpansive property analogous to those for projection and proximal methods.

LEMMA 3.1. *Consider any maximal monotone mappings $A : \mathcal{H} \rightrightarrows \mathcal{H}$ and $B : \mathcal{H} \rightrightarrows \mathcal{H}$ and any closed convex set $X \subset \mathrm{dom}A$ such that A is single-valued on $\mathrm{dom}A \supset \mathrm{dom}B$ and $\Sigma = X \cap (A+B)^{-1}(0)$ is nonempty. For any $\alpha > 0$, any $x \in \mathrm{dom}A$, and any $x^* \in \Sigma$, the vectors $\bar{x} = J(x,\alpha)$ and $z = \bar{x} - \alpha(A(\bar{x}) - A(x))$, where J is given by (2.1), together with some $\eta \geq 0$ satisfy*

$$\|P_X[z] - x^*\|^2 \leq \|z - x^*\|^2 = \|x - x^*\|^2 + \alpha^2\|A(\bar{x}) - A(x)\|^2 - \|\bar{x} - x\|^2 - 2\alpha\eta, \tag{3.1}$$

with η having the property that, if $A+B$ is strongly monotone on $\mathrm{dom}B$ with modulus $\gamma > 0$, then $\eta \geq \gamma\|\bar{x} - x^*\|^2$.

Proof. We have from the definition of $\bar{x}$ and z that

$$\bar{x} + \alpha\bar{v} = x - \alpha u, \quad z = \bar{x} - \alpha(\bar{u} - u), \quad u = A(x), \quad \bar{u} = A(\bar{x}), \quad \exists \bar{v} \in B(\bar{x}) \tag{3.2}$$

and from $0 \in A(x^*) + B(x^*)$ that

$$u^* + v^* = 0, \qquad u^* = A(u^*), \quad \exists v^* \in B(x^*). \tag{3.3}$$

Then,

$$\begin{aligned}
\|x - x^*\|^2 &= \|x - \bar{x} + \bar{x} - z + z - x^*\|^2 \\
&= \|x - \bar{x}\|^2 + \|\bar{x} - z\|^2 + \|z - x^*\|^2 + 2\langle x - \bar{x}, \bar{x} - x^*\rangle + 2\langle \bar{x} - z, z - x^*\rangle \\
&= \|x - \bar{x}\|^2 - \|\bar{x} - z\|^2 + \|z - x^*\|^2 + 2\langle x - z, \bar{x} - x^*\rangle \\
&= \|x - \bar{x}\|^2 - \alpha^2\|A(\bar{x}) - A(x)\|^2 + \|z - x^*\|^2 + 2\alpha\langle \bar{u} + \bar{v}, \bar{x} - x^*\rangle \\
&= \|x - \bar{x}\|^2 - \alpha^2\|A(\bar{x}) - A(x)\|^2 + \|z - x^*\|^2 + 2\alpha\langle \bar{u} - u^* + \bar{v} - v^*, \bar{x} - x^*\rangle,
\end{aligned}$$

where the fourth equality uses (3.2) and the fifth equality uses (3.3). This proves the equality in (3.1) with $\eta = \langle \bar{u} - u^* + \bar{v} - v^*, \bar{x} - x^*\rangle$. Since A and B are monotone, it follows from (3.2) and (3.3) that $\eta = \langle \bar{u} - u^* + \bar{v} - v^*, \bar{x} - x^*\rangle$ is nonnegative. And if in addition $A + B$ is strongly monotone on $\mathrm{dom}B$ with modulus $\gamma > 0$, then $\eta \geq \gamma\|\bar{x} - x^*\|^2$. The inequality in (3.1) follows from $x^* = P_X[x^*]$ (since $x^* \in X$) and the nonexpansive property of P_X [60, Eq. (1.8)]. □

The next lemma is well known (see [4, p. 27], [36, p. 105]) and a proof is included for completeness.

LEMMA 3.2. *Consider any maximal monotone mapping $S : \mathcal{H} \rightrightarrows \mathcal{H}$. If $\{x^k\}$ is a sequence in $\mathcal{H}$ bounded in norm and converging weakly to some x and $\{w^k\}$ is a sequence in $\mathcal{H}$ converging strongly to some w and $w^k \in S(x^k)$ for all k, then $w \in S(x)$.*

Proof. For any $x' \in \mathrm{dom}S$ and any $w' \in S(x')$, we have $0 \le \langle w'-w^k, x'-x^k\rangle = \langle w'-w^k, x'-x\rangle + \langle w'-w, x-x^k\rangle + \langle w-w^k, x-x^k\rangle \le \langle w'-w^k, x'-x\rangle + \langle w'-w, x-x^k\rangle + \|w-w^k\|\|x-x^k\| \to \langle w'-w, x'-x\rangle$. The maximality of S then implies $w \in S(x)$. □

In a Hilbert space, a weakly convergent sequence is automatically bounded in norm (see, e.g., [36, p. 2]), but we will not need this fact. The following lemma extends a basic result about the backward mapping $(I+\alpha B)^{-1}$ (see [4, Prop. 2.6]) to the F-B mapping J.

LEMMA 3.3. *Consider any maximal monotone mappings $A : \mathcal{H} \rightrightarrows \mathcal{H}$ and $B : \mathcal{H} \rightrightarrows \mathcal{H}$ such that A is single-valued on* $\mathrm{dom}A \supset \mathrm{dom}B$. *Let J be given by* (2.1). *Then,*

$$\|J(x,\alpha)-x\|/\alpha \le \min_{w\in A(x)+B(x)} \|w\| \qquad \textit{for all } x \in \mathrm{dom}B, \ \textit{for all } \alpha > 0. \tag{3.4}$$

Proof. Fix any $x \in \mathrm{dom}B$ and any $\alpha > 0$. Let $z = J(x,\alpha)$. Then, (2.1) implies $(x-z)/\alpha \in A(x)+B(z)$ so that, for any $w \in A(x)+B(x)$, the monotonicity of B implies

$$\langle ((x-z)/\alpha - A(x)) - (w - A(x)), z-x\rangle \ge 0.$$

Simplifying and rearranging terms give

$$\|x-z\|^2/\alpha \le \langle w, x-z\rangle \le \|w\|\|x-z\|,$$

and (3.4) is proven. (The minimum in (3.4) is attained uniquely since $A(x)+B(x)$ is a closed convex set.) □

The inequality in (3.4) is sharp as $\alpha \to 0$. To see this, note that, by (3.4), $J(x,\alpha)$ converges strongly to x as $\alpha \to 0$ and $w_\alpha = (x - J(x,\alpha))/\alpha$ is bounded in norm. Choose any sequence $\{\alpha_k\} \to 0$ such that $\|w_{\alpha_k}\| \to \liminf_{\alpha\to 0}\|w_\alpha\|$. By Alaoglu's theorem [36, p. 2], $\{w_{\alpha_k}\}$ has a weak cluster point and, by $w_\alpha \in A(x)+B(J(x,\alpha))$ and Lemma 3.2 with $S = B^{-1}$, any such weak cluster point w is in $A(x)+B(x)$. This together with $\|w_\alpha\|^2 = \|w_\alpha - w\|^2 + \|w\|^2 + 2\langle w_\alpha - w, w\rangle \ge \|w\|^2 + 2\langle w_\alpha - w, w\rangle \to \|w\|^2$ as $\alpha \to 0$ implies

$$\liminf_{\alpha\to 0} \|J(x,\alpha)-x\|/\alpha = \lim_k \|w_{\alpha_k}\| \ \ge \min_{w\in A(x)+B(x)} \|w\|. \tag{3.5}$$

By (3.4), the inequality in (3.5) holds with equality. The mapping $x \mapsto \arg\min_{w\in T(x)}\|w\|$ has been much studied; see, e.g., [4, Chap. 2].

Below we state and prove our main convergence result, showing that, under mild assumptions on A and B, the modified F-B method with α_k determined by the Armijo–Goldstein-type stepsize rule (2.4) generates well-defined iterates x^k that converge weakly to a solution. Moreover, if T is strongly monotone or T^{-1} has a (local) Lipschitzian property (see (3.7)) and if the stepsizes α_k are bounded away from zero, then the iterates have (local) linear rate of convergence. The proof entails using Lemma 3.3 to show that (2.4) holds for all α sufficiently small, so that α_k is well defined. Then, Lemmas 3.1 and 3.2 are used to show, respectively, that x^k are bounded in norm and every weak cluster point is a solution.

THEOREM 3.4. *Consider any maximal monotone mappings $A : \mathcal{H} \rightrightarrows \mathcal{H}$ and $B : \mathcal{H} \rightrightarrows \mathcal{H}$ such that A is single-valued on* $\mathrm{dom}A \supset \mathrm{dom}B$, *and $T = A + B$ is maximal monotone with $T^{-1}(0) \ne \emptyset$. Assume $X \subset \mathrm{dom}A$ is a closed convex set such*

that $\Sigma = X \cap T^{-1}(0) \neq \emptyset$ *and either* A *is Lipschitz continuous on* $X \cup \mathrm{dom}B$ *or* A *is continuous on* $\mathrm{dom}B \supset X$. *Let* $\{(x^k, \bar{x}^k)\}_{k=0,1,\ldots}$ *be generated by the modified F-B method* (2.2)–(2.3) *with* α_k *chosen to be the largest* $\alpha \in \{\sigma, \sigma\beta, \sigma\beta^2, \ldots\}$ *satisfying* (2.4), *where* $\beta \in (0,1)$, $\theta \in (0,1)$, *and* $\sigma > 0$. *Then the following hold.*

(a) α_k *is well defined for all* k. *If* A *is Lipschitz continuous on* $X \cup \mathrm{dom}B$ *with constant* $\lambda \geq 0$, *then* $\alpha_k \geq \min\{\sigma, \theta\beta/\lambda\}$.

(b) *For every* $x^* \in \Sigma$ *and every* $k \in \{0, 1, \ldots\}$, *we have*

$$\|x^{k+1} - x^*\|^2 \leq \|x^k - x^*\|^2 - (1-\theta^2)\|\bar{x}^k - x^k\|^2 - 2\alpha_k \eta^k \tag{3.6}$$

for some $\eta^k \geq 0$ *with the property that, if* T *is strongly monotone on* $\mathrm{dom}B$ *with modulus* $\gamma > 0$, *then* $\eta^k \geq \gamma\|\bar{x}^k - x^*\|^2$. *If either* (i) $\liminf_{k\to\infty}\alpha_k > 0$ *or* (ii) A *is locally uniformly continuous on* $X \subset \mathrm{dom}B$ *(see* (1.1)*) and the function* $x \mapsto \min_{w\in T(x)}\|w\|$ *is locally bounded on* X, *then* $\{x^k\}$ *converges weakly to an element of* Σ.

(c) *If* T *is strongly monotone on* $\mathrm{dom}B$ *with modulus* $\gamma > 0$, *then*

$$\|x^{k+1} - x^*\| \;\leq\; (1 - \min\{1-\theta^2, 2\alpha_k\gamma\}/2)^{1/2}\|x^k - x^*\|$$

for all k, *where* x^* *denotes the unique element of* Σ.

(d) *If there exist* $\tau > 0$ *and* $\delta > 0$ *such that*

$$(X + \delta\mathbb{B}) \cap T^{-1}(w) \subset \Sigma + \tau\|w\|\mathbb{B} \quad \textit{for all } w \in \mathcal{H} \textit{ with } \|w\| \leq \delta, \tag{3.7}$$

where $\mathbb{B} = \{x \in \mathcal{H} : \|x\| \leq 1\}$, *and either* $\liminf_{k\to\infty}\alpha_k > 0$ *or* $\delta = \infty$, *then there exists an index* $\bar{k}$ *such that*

$$d\left(x^{k+1}, \Sigma\right) \;\leq\; d\left(x^k, \Sigma\right)/(1 + \rho\alpha_k^2)^{1/2}$$

for all $k \geq \bar{k}$, *where we denote* $d(x, \Sigma) = \min_{x^*\in\Sigma}\|x - x^*\|$ *and* $\rho = (1 - \theta^2)/((\sigma+\tau)\theta + \tau)^2$. *If* $\delta = \infty$, *then* $\bar{k} = 0$. *(The minimum is attained uniquely since* Σ *is a closed convex set.)*

Proof. (a) Suppose A is Lipschitz continuous on $X \cup \mathrm{dom}B$ with constant $\lambda \geq 0$. For each $k \in \{0, 1, \ldots\}$, since $x^k \in X$ and $J(x^k, \alpha) \in \mathrm{dom}B$, it follows from the Lipschitz continuity of A that (2.4) holds for all $\alpha \leq \theta/\lambda$, so α_k is well defined. Moreover, either $\alpha_k = \sigma$ or else (2.4) fails to hold for $\alpha = \alpha_k/\beta$. In the latter case, we must have $\alpha_k/\beta > \theta/\lambda$.

Suppose A is continuous on $\mathrm{dom}B$ and $X \subset \mathrm{dom}B$. Fix any $k \in \{0, 1, \ldots\}$. If $x^k \in \Sigma$, then $\alpha_k = \sigma$ (since both sides of (2.4) equal zero for any $\alpha > 0$). Now suppose $x^k \notin \Sigma$. Since $x^k \in X \subset \mathrm{dom}B$, applying Lemma 3.3 with $x = x^k$ yields

$$\|J(x^k, \alpha) - x^k\|/\alpha \leq \min_{w\in T(x^k)}\|w\| \quad \text{for all } \alpha > 0, \tag{3.8}$$

so $J(x^k, \alpha)$ converges strongly to x^k as $\alpha \to 0$ and the continuity of A on $\mathrm{dom}B$ implies

$$\|A(J(x^k, \alpha)) - A(x^k)\| \to 0 \quad \text{as } \alpha \to 0.$$

By (2.1), we have $(x^k - J(x^k, \alpha))/\alpha \in A(x^k) + B(J(x^k, \alpha))$. If $\liminf_{\alpha\to 0}\|x^k - J(x^k, \alpha)\|/\alpha = 0$, then since $J(x^k, \alpha)$ converges strongly to x^k, Lemma 3.2 with $S = B$ would yield $0 \in A(x^k) + B(x^k)$, contradicting $x^k \notin \Sigma$. Thus,

$$\liminf_{\alpha\to 0}\|x^k - J(x^k, \alpha)\|/\alpha > 0.$$

The preceding two relations show that (2.4) holds whenever α is sufficiently small, so α_k is well defined.

(b) For every $x^* \in \Sigma$ and every $k \in \{0, 1, \ldots\}$, we have from (2.2), (2.3), and applying Lemma 3.1 that (3.1) holds with $\alpha = \alpha_k$, $x = x^k$, $\bar{x} = \bar{x}^k$, $P_X[z] = x^{k+1}$, and $\eta = \eta^k$, for some $\eta^k \geq 0$ having the desired property. Hence

$$\|x^{k+1} - x^*\|^2 \leq \|x^k - x^*\|^2 + (\alpha_k)^2 \|A(\bar{x}^k) - A(x^k)\|^2 - \|\bar{x}^k - x^k\|^2 - 2\alpha_k \eta^k.$$

Since $\alpha = \alpha_k$ satisfies (2.4) so that, by (2.2), $\alpha_k \|A(\bar{x}^k) - A(x^k)\| \leq \theta \|\bar{x}^k - x^k\|$, this yields (3.6). Thus, (3.6) holds for $k = 0, 1, \ldots$ and any $x^* \in \Sigma$. Then the sequence $\{x^k\}_{k=0,1,\ldots}$ is bounded in norm and, by Alaoglu's theorem [36, p. 2], has at least one weak cluster point.

Let x^∞ be any weak cluster point of $\{x^k\}$, and we will show that $x^\infty \in \Sigma$. Consider any subsequence $\{x^k\}_{k \in K}$ ($K \subset \{0, 1, 2, \ldots\}$) converging weakly to x^∞. Since $x^k \in X$ and X is closed, then $x^\infty \in X$. Suppose $\liminf_{k \to \infty} \alpha_k > 0$. Since (3.6) implies $\|\bar{x}^k - x^k\| \to 0$, this together with $\alpha = \alpha_k$ in (2.4) and (2.2) yields $\|A(\bar{x}^k) - A(x^k)\| \to 0$. By (2.1) and (2.2), we also have

$$(x^k - \bar{x}^k)/\alpha_k + A(\bar{x}^k) - A(x^k) \in A(\bar{x}^k) + B(\bar{x}^k) = T(\bar{x}^k) \tag{3.9}$$

for all k. It follows that the left-hand side of (3.9) converges strongly to 0 as $k \in K$, $k \to \infty$. Since $\{\bar{x}^k\}_{k \in K}$ is bounded in norm and converges weakly to x^∞, Lemma 3.2 with $S = T$ yields $0 \in T(x^\infty)$. Thus, $x^\infty \in \Sigma$. Suppose instead A is locally uniformly continuous on $X \subset \mathrm{dom} B$ and the function $x \mapsto \min_{w \in T(x)} \|w\|$ is locally bounded on X. If $\{\alpha_k\}_{k \in K}$ contains a subsequence that is bounded below by a positive scalar, then an argument analogous to that used above would yield $x^\infty \in \Sigma$. Otherwise, suppose $\{\alpha_k\}_{k \in K} \to 0$. Then for all $k \in K$ sufficiently large, we have $\alpha_k < \sigma$, so our choice of α_k implies (2.4) fails to hold for $\alpha = \bar{\alpha}_k$, where $\bar{\alpha}_k = \alpha_k / \beta$, i.e.,

$$\theta \|J(x^k, \bar{\alpha}_k) - x^k\| / \bar{\alpha}_k < \|A(J(x^k, \bar{\alpha}_k)) - A(x^k)\|. \tag{3.10}$$

Applying Lemma 3.3 with $x = x^k$ and $\alpha = \bar{\alpha}_k$ yields

$$\|J(x^k, \bar{\alpha}_k) - x^k\| / \bar{\alpha}_k \ \leq \ \min_{w \in T(x^k)} \|w\|$$

for all $k \in K$. Since x^k is in X and converges weakly to x^∞ as $k \in K, k \to \infty$, the right-hand side is bounded for all $k \in K$, implying $\{J(x^k, \bar{\alpha}_k)\}_{k \in K}$ is bounded in norm and converges weakly to x^∞. By (2.1) and $T = A + B$, we have

$$w^k = (x^k - J(x^k, \bar{\alpha}_k))/\bar{\alpha}_k + A(J(x^k, \bar{\alpha}_k)) - A(x^k) \in T(J(x^k, \bar{\alpha}_k))$$

for all $k \in K$. If $\liminf_{k \in K, k \to \infty} \|w^k\| = 0$, then Lemma 3.2 with $S = T$ would imply $0 \in T(x^\infty)$, so $x^\infty \in \Sigma$. Otherwise suppose $\liminf_{k \in K, k \to \infty} \|w^k\| > 0$. Since $x^k \in X$ and $J(x^k, \bar{\alpha}_k) \in \mathrm{dom} B$ converge weakly and $\|J(x^k, \bar{\alpha}_k) - x^k\| \to 0$ as $k \in K, k \to \infty$, we have from the local uniform continuity of A on $X \subset \mathrm{dom} B$ that $\{\|A(J(x^k, \bar{\alpha}_k)) - A(x^k)\|\}_{k \in K} \to 0$, and hence $\liminf_{k \in K, k \to \infty} \|x^k - J(x^k, \bar{\alpha}_k)\| / \bar{\alpha}_k = \liminf_{k \in K, k \to \infty} \|w^k\| > 0$. This contradicts the fact that (3.10) holds for all $k \in K$ sufficiently large.

We now show, by an argument used in [3] and in [44, p. 885], that $\{x^k\}$ has no more than one weak cluster point. Suppose x_1^∞ and x_2^∞ are two weak cluster points of $\{x^k\}$. Then, as we just showed above, for $i = 1, 2$, we have $x_i^\infty \in \Sigma$, so letting

$x^* = x_i^\infty$ in (3.6) yields $\{\|x^k - x_i^\infty\|^2\}$ is monotonically decreasing and hence converges to a limit, say, υ_i. Since

$$\|x^k - x_1^\infty\|^2 = \|x^k - x_2^\infty\|^2 + \|x_2^\infty - x_1^\infty\|^2 + 2\langle x^k - x_2^\infty, x_1^\infty - x_2^\infty\rangle$$

for all k, passing to the limit along any subsequence of $\{x^k\}$ converging weakly to x_2^∞ yields $\upsilon_1 = \upsilon_2 + \|x_2^\infty - x_1^\infty\|^2$. A symmetric argument yields $\upsilon_2 = \upsilon_1 + \|x_1^\infty - x_2^\infty\|^2$. Adding these two equalities gives $\|x_1^\infty - x_2^\infty\| = 0$.

(c) Suppose that T is strongly monotone on dom$B$ with modulus $\gamma > 0$. Then (b) implies (3.6) holds and $\eta^k \geq \gamma\|\bar{x}^k - x^*\|^2$, so that

$$\begin{aligned}\|x^{k+1} - x^*\|^2 &\leq \|x^k - x^*\|^2 - (1-\theta^2)\|\bar{x}^k - x^k\|^2 - 2\alpha_k\gamma\|\bar{x}^k - x^*\|^2 \\ &\leq \|x^k - x^*\|^2 - \min\{1-\theta^2, 2\alpha_k\gamma\}(\|\bar{x}^k - x^k\|^2 + \|\bar{x}^k - x^*\|^2) \\ &\leq \|x^k - x^*\|^2 - \frac{1}{2}\min\{1-\theta^2, 2\alpha_k\gamma\}\|x^k - x^*\|^2,\end{aligned}$$

where the last inequality follows from $\|x^k - x^*\|^2 \leq (\|x^k - \bar{x}^k\| + \|\bar{x}^k - x^*\|)^2 \leq 2\|x^k - \bar{x}^k\|^2 + 2\|\bar{x}^k - x^*\|^2$.

(d) Suppose that there exist $\tau > 0$ and $\delta > 0$ such that (3.7) holds. It can be seen that (3.7) is equivalent to

$$d(x, \Sigma) \leq \tau \min_{w \in T(x)} \|w\| \quad \text{for all } x \in \text{dom}T \cap (X + \delta\mathbb{B}) \text{ with } \min_{w \in T(x)} \|w\| \leq \delta. \tag{3.11}$$

(The minimum is attained uniquely since $T(x)$ is a closed convex set.) Also, we have from (3.9) that the vector $z^k = \bar{x}^k - \alpha_k(A(\bar{x}^k) - A(x^k))$ satisfies

$$(x^k - z^k)/\alpha_k = (x^k - \bar{x}^k)/\alpha_k - A(x^k) + A(\bar{x}^k) \in B(\bar{x}^k) + A(\bar{x}^k) = T(\bar{x}^k) \tag{3.12}$$

for all k. Suppose $\liminf_{k\to\infty}\alpha_k > 0$. Since (3.6) implies $\{\|\bar{x}^k - x^k\|\} \to 0$, we have from (2.4) with $\alpha = \alpha_k$ and (2.2) that $\{\|A(\bar{x}^k) - A(x^k)\|\} \to 0$. Thus, there exists an index $\bar{k}$ such that the left-hand side of (3.12) and $\bar{x}^k - x^k$ are both below δ in norm for all $k \geq \bar{k}$, in which case $x^k \in X$ and (3.11) yield

$$d\left(\bar{x}^k, \Sigma\right) \leq \tau\|x^k - z^k\|/\alpha_k.$$

Then, for $x^* \in \Sigma$ satisfying $\|\bar{x}^k - x^*\| = d(\bar{x}^k, \Sigma)$, we have

$$\begin{aligned}d\left(x^{k+1}, \Sigma\right) &\leq \|x^{k+1} - x^*\| \\ &= \|P_X[z^k] - P_X[x^*]\| \\ &\leq \|z^k - x^*\| \\ &\leq \|z^k - \bar{x}^k\| + \|\bar{x}^k - x^*\| \\ &\leq \|z^k - \bar{x}^k\| + \tau\|x^k - z^k\|/\alpha_k \\ &\leq \|z^k - \bar{x}^k\| + \tau(\|x^k - \bar{x}^k\| + \|\bar{x}^k - z^k\|)/\alpha_k \\ &= (\alpha_k + \tau)\|A(\bar{x}^k) - A(x^k)\| + \tau\|x^k - \bar{x}^k\|/\alpha_k \\ &\leq ((\alpha_k + \tau)\theta + \tau)\|x^k - \bar{x}^k\|/\alpha_k,\end{aligned}$$

where the first equality uses (2.3), the second inequality uses the nonexpansive property of P_X [60, Eq. (1.8)], and the last inequality uses (2.2) and (2.4) with $\alpha = \alpha_k$.

Then, for $x^* \in \Sigma$ satisfying $\|x^k - x^*\| = d(x^k, \Sigma)$, inequality (3.6) and the above inequality yield

$$\begin{aligned} d\left(x^{k+1}, \Sigma\right)^2 &\leq \|x^{k+1} - x^*\|^2 \\ &\leq \|x^k - x^*\|^2 - (1-\theta^2)\|\bar{x}^k - x^k\|^2 \\ &\leq \|x^k - x^*\|^2 - \frac{(1-\theta^2)\alpha_k^2}{((\alpha_k+\tau)\theta+\tau)^2} d\left(x^{k+1}, \Sigma\right)^2 \\ &= d\left(x^k, \Sigma\right)^2 - \frac{(1-\theta^2)\alpha_k^2}{((\alpha_k+\tau)\theta+\tau)^2} d\left(x^{k+1}, \Sigma\right)^2. \end{aligned}$$

This holds for all $k \geq \bar{k}$ and, since $\alpha_k \leq \sigma$, the desired inequality follows. The case of $\delta = \infty$ is treated similarly with $\bar{k} = 0$. □

Sufficient conditions for the sum of two maximal monotone mappings $T = A + B$ to be maximal are given in [42]. The assumption made in Theorem 3.4(b) on the local uniform continuity of A on $X \subset \mathrm{dom}B$ holds if either A is continuous (from the weak topology to the strong topology) on $\mathrm{dom}B$ or A is uniformly continuous (from the strong topology to the strong topology) on $\mathrm{dom}B$. The other assumption that $x \mapsto \min_{w \in T(x)} \|w\|$ be locally bounded on X is reasonably mild. If A is continuous, from the weak topology to the strong topology, on X and $B = G + N_C$, with $G : \mathcal{H} \rightrightarrows \mathcal{H}$ maximal monotone and C a nonempty closed convex subset of $\mathrm{int}(\mathrm{dom}G)$, then this assumption holds. This is because G is locally bounded on $\mathrm{int}(\mathrm{dom}G)$ [41] (also see [4, Prop. 2.9]). However, in general, this assumption may fail to hold. For example, for $B = \partial f$, where f is the proper closed convex function on $\Re^2$ defined by

$$f(x) = f(x_1, x_2) = \begin{cases} \max\{-\sqrt{1 - x_1^2 - x_2^2}, x_1 - 1\} & \text{if } x_1^2 + x_2^2 \leq 1, \\ \infty & \text{else,} \end{cases}$$

it can be checked that $x \mapsto \min_{v \in B(x)} \|v\|$ is not locally bounded at $(1, 0) \in \mathrm{dom}B$. The assumption (3.7) made in Theorem 3.4(d) is weaker than the Lipschitzian assumption made in [44, Eq. (3.1)] as it does not require $T^{-1}(0)$ to be a singleton. This assumption has been much studied and is known to hold when T is polyhedral [40] (also see [47, Chap. 9] for related discussions). In the case of variational inequalities, corresponding to $T = F + N_C$ with F single-valued on a nonempty closed convex set C, we have $\min_{w \in T(x)} \|w\| = \|F(x) + v\|$ for some $v \in N_C(x)$, implying

$$\|x - P_C[x - F(x)]\| = \|P_C[x+v] - P_C[x - F(x)]\| \leq \|(x+v) - (x - F(x))\| = \min_{w \in T(x)} \|w\|.$$

Thus, the Lipschitzian property (3.7) and its equivalent formulation as an error bound (3.11) can be inferred from corresponding results for the projection residual $R(x) = x - P_C[x - F(x)]$, as studied in [23, 24, 35, 58] and references therein. Also, as in [7, 8, 9, 44, 50], it may be worthwhile to consider inexact evaluation of the backward mapping $(I + \alpha B)^{-1}$.

In the case where A is Lipschitz continuous on $\mathrm{dom}A = \mathcal{H}$ with constant λ and T is strongly monotone with modulus γ, parts (a) and (c) of Theorem 3.4 imply that x^k converge strongly at linear rate to the unique element of $T^{-1}(0)$ and the convergence ratio is at most

$$(1 - \min\{1 - \theta^2, 2\gamma\sigma, 2\gamma\theta\beta/\lambda\}/2)^{1/2}.$$

Assuming λ and γ are known, one can choose $\sigma > 0$ and $\theta \in (0, 1)$ to minimize the above estimate and obtain

$$1/\sqrt{1 + \beta\kappa/(1 + (\beta\kappa)^2)^{1/2}},$$

where $\kappa = \gamma/\lambda$. This can be compared with an analogous estimate by Chen and Rockafellar [6, Cor. 2.5] for the F-B method:

$$1/\sqrt{1+\kappa^2/(1-(\gamma_1/\lambda)^2)},$$

where γ_1 is the modulus of A (with $\gamma_1 = 0$ allowed). Which estimate is smaller depends on κ, β, and γ_1/λ, and whether these estimates reflect the methods' behavior in practice remains to be seen.

4. Applications. Below we derive new decomposition methods by applying the modified F-B method appropriately to special cases of convex programming and variational inequalities. Throughout this paper, for any matrix $D \in \Re^{m\times n}$, we denote its transpose by D^T and its operator norm by $\|D\| = \max_{x\in\Re^n:\|x\|=1}\|Dx\|$. For any function $f : \Re^n \mapsto (-\infty,\infty]$, we denote its effective domain by $\mathrm{dom}f = \{x \in \Re^n : f(x) < \infty\}$.

Example 1. Consider the following convex program studied in [57, Sec. 4]:

$$\begin{array}{ll} \text{minimize} & f_1(x_1) + f_2(x_2) \\ \text{subject to} & Dx_1 + Ex_2 = b, \end{array} \tag{4.1}$$

where f_1 and f_2 are closed proper convex functions on, respectively, $\Re^{n_1}$ and $\Re^{n_2}$ and $D \in \Re^{m\times n_1}$, $E \in \Re^{m\times n_2}$, $b \in \Re^m$. We assume that f_1 is strictly convex and cofinite [43, p. 116], so that $(\partial f_1)^{-1}$ is single-valued and continuous on $\Re^{n_1}$. (Roughly speaking, f_1 being cofinite means it grows faster than linear. If f_1 is strongly convex or if $\mathrm{dom}f_1$ is bounded, then f_1 is cofinite.) We also assume that there exists $(x_1, x_2) \in \mathrm{ri}(\mathrm{dom}f_1) \times \mathrm{ri}(\mathrm{dom}f_2)$ satisfying $Dx_1 + Ex_2 = b$ and that there does not exist $x_2 \in \mathrm{dom}f_2$ and $z \in \Re^{n_2}$ satisfying $Ez = 0$ and $f_2(x_2 + tz) < f_2(x_2)$ for all $t > 0$. Then, it can be argued similarly as in [57] that the above convex program has an optimal solution and a Kuhn–Tucker vector $y \in \Re^m$ associated with the constraints $Dx_1 + Ex_2 = b$. Moreover, y is a Kuhn–Tucker vector if and only if y satisfies $0 \in A(y) + B(y)$, where A and B are the maximal monotone mappings:

$$A(y) = D(\partial f_1)^{-1}(D^T y), \qquad B(y) = E(\partial f_2)^{-1}(E^T y) - b.$$

Notice that A is single-valued and continuous on $\Re^m$. Applying the modified F-B method with this choice of A and B and with X being a suitable closed convex subset of $\mathrm{dom}B$ (or $X = \Re^m$, if $(\partial f_1)^{-1}$ is Lipschitz continuous on $\Re^{n_1}$), we obtain the following modification to the alternating minimization algorithm in [57]:

$$\begin{aligned} \bar{x}_1^k &= \arg\min_{x_1\in\Re^{n_1}} \{f_1(x_1) - \langle y^k, Dx_1\rangle\}, \\ \bar{x}_2^k &= \arg\min_{x_2\in\Re^{n_2}} \{f_2(x_2) - \langle y^k, Ex_2\rangle + \alpha_k\|D\bar{x}_1^k + Ex_2 - b\|^2/2\}, \\ \bar{y}^k &= y^k + \alpha_k(b - D\bar{x}_1^k - E\bar{x}_2^k), \\ x_1^{k+1} &= \arg\min_{x_1\in\Re^{n_1}} \{f_1(x_1) - \langle \bar{y}^k, Dx_1\rangle\}, \\ y^{k+1} &= P_X[\bar{y}^k - \alpha_k(Dx_1^{k+1} - D\bar{x}_1^k)], \end{aligned}$$

for $k = 0, 1, \ldots$ In contrast to the original method, the modified method does not require f_1 to be strongly convex for convergence, so f_1 can include functions such as $x \ln x$, which are strictly convex and cofinite, but not strongly convex. On the other hand, the modified method requires an additional minimization in x_1 and a projection

onto the set X. If f_2 has a separable structure and $E = -I$, as considered in [7], then X can be chosen to have a corresponding Cartesian product structure and the projection onto X would decompose accordingly. Alternatively, if f_2 is cofinite so that $\mathrm{dom}(\partial f_2)^{-1} = \Re^{n_2}$, then X can be chosen to be $\mathrm{dom}B = \Re^m$ and the projection would be vacuous.

Example 2. Consider the variational inequality problem of finding an $x \in \mathcal{H}$ satisfying $0 \in F(x) + N_C(x)$, where C is a nonempty closed convex set in $\mathcal{H}$ and F is a maximal monotone mapping that is single-valued and continuous on C. This is equivalent to $0 \in A(x) + B(x)$, where A and B are the maximal monotone mappings:

$$A(x) = F(x), \qquad B(x) = N_C(x).$$

Applying the modified F-B method with this choice of A and B and with $X = C$ (or $X = \mathcal{H}$, if A is Lipschitz continuous on $\mathrm{dom}A = \mathcal{H}$), we obtain the following (new) double-projection method:

$$\bar{x}^k = P_C[x^k - \alpha_k F(x^k)], \tag{4.2}$$

$$x^{k+1} = P_X[\bar{x}^k - \alpha_k(F(\bar{x}^k) - F(x^k))], \tag{4.3}$$

for $k = 0, 1, \ldots$ This method differs from the extragradient method [17, 19, 26, 54], whose second equation is

$$x^{k+1} = P_C[x^k - \alpha_k F(\bar{x}^k)]. \tag{4.4}$$

It also differs from a modified projection-type method [52] (also see [55] for a similar method) whose second equation is

$$x^{k+1} = x^k - \gamma_k(x^k - \bar{x}^k + \alpha_k(F(\bar{x}^k) - F(x^k))) \tag{4.5}$$

with $\gamma_k > 0$ a quantity depending on $x^k, \bar{x}^k, \alpha_k$. If $X = \mathcal{H}$, then (4.3) and (4.5) would coincide whenever $\gamma_k = 1$. If in addition $C = \mathcal{H}$, then (4.3) and (4.4) would also coincide. In this case, the three methods differ only in their stepsize rules for choosing α_k. While the method (4.2)–(4.3) has similar theoretical convergence properties as the other two methods, its practical performance remains to be determined from numerical testing. Notice that (4.3) involves projection onto X rather than onto C as in (4.4), which can be advantageous when X has a simpler structure than C. Other projection-type methods, including the method of Sibony, require stronger assumptions, such as F^{-1} being strongly monotone, for convergence (see [2, 25, 29, 49] and references therein).

Example 3. Consider the convex program (4.1), where f_1 and f_2 are closed proper convex functions on, respectively, $\Re^{n_1}$ and $\Re^{n_2}$ and $D \in \Re^{m\times n_1}$, $E \in \Re^{m\times n_2}$, $b \in \Re^m$. The case of $E = -I$ and $b = 0$ corresponds to the problem studied in [7]. We assume there exists $(x_1, x_2) \in \mathrm{ri}(\mathrm{dom}f_1)\times\mathrm{ri}(\mathrm{dom}f_2)$ satisfying $Dx_1 + Ex_2 = b$ but, in contrast to Example 1, we do not assume f_1 is strictly convex or cofinite. Then, this problem has an optimal solution if and only if $0 \in A(x_1, x_2, y) + B(x_1, x_2, y)$ has a solution [42, Chap. 28], where A and B are the maximal monotone mappings:

$$A(x_1, x_2, y) = (D^T y, E^T y, -Dx_1 - Ex_2), \qquad B(x_1, x_2, y) = \partial f_1(x_1) \times \partial f_2(x_2) \times \{b\}.$$

Notice that A is Lipschitz continuous on $\Re^{m+n_1+n_2}$ with constant

$$\lambda = \sqrt{\|D^T\|^2 + \|D\|^2 + \|E^T\|^2 + \|E\|^2}.$$

Applying the modified F-B method with this choice of A and B and with $X = \Re^{m+n_1+n_2}$, we obtain (in case of $E = -I$ and $b = 0$) the following variant of a decomposition method of Chen and Teboulle [7]:

$$\begin{aligned}
\bar{x}_1^k &= \arg\min_{x_1 \in \Re^{n_1}} \{f_1(x_1) + \langle y^k, Dx_1\rangle + \|x_1 - x_1^k\|^2/(2\alpha_k)\},\\
\bar{x}_2^k &= \arg\min_{x_2 \in \Re^{n_2}} \{f_2(x_2) + \langle y^k, Ex_2\rangle + \|x_2 - x_2^k\|^2/(2\alpha_k)\},\\
\bar{y}^k &= y^k + \alpha_k(Dx_1^k + Ex_2^k - b),\\
x_1^{k+1} &= \bar{x}_1^k - \alpha_k D^T(\bar{y}^k - y^k),\\
x_2^{k+1} &= \bar{x}_2^k - \alpha_k E^T(y^k - \bar{y}^k),\\
y^{k+1} &= \bar{y}^k + \alpha_k(D\bar{x}_1^k + E\bar{x}_2^k - Dx_1^k - Ex_2^k),
\end{aligned}$$

for $k = 0, 1, \ldots$ This method has the same minimization subproblems as the Chen–Teboulle method and, in particular, the subproblems have strongly convex objective function and decompose according to the separable structure of f_1 and f_2.

Example 4. Consider the inclusion problem $0 \in A(x_1, x_2, y) + B(x_1, x_2, y)$, where

$$A(x_1, x_2, y) = (D^T y, E^T y, -Dx_1 - Ex_2), \qquad B(x_1, x_2, y) = T_1(x_1) \times T_2(x_2) \times \{b\},$$

and T_1 and T_2 are maximal monotone mappings on $\Re^{n_1}$ and $\Re^{n_2}$, respectively, and $D \in \Re^{m\times n_1}$, $E \in \Re^{m\times n_2}$, $b \in \Re^m$. Then, A and B are maximal monotone and A is Lipschitz continuous on $\Re^{m+n_1+n_2}$ with constant

$$\lambda = \sqrt{\|D^T\|^2 + \|D\|^2 + \|E^T\|^2 + \|E\|^2}.$$

The special case where $T_1 = \partial f_1$, $T_2 = \partial f_2$ yields the convex program (4.1). The special case where $n_1 = n_2$, $D = -E = I$, and $b = 0$ yields the inclusion $0 \in T_1(x) + T_2(x)$. Applying the modified F-B method with this choice of A and B and with $X = \Re^{m+n_1+n_2}$, we obtain the following variant of a splitting method in [59, Ex. 3]:

$$\begin{aligned}
\bar{x}_1^k &= (I + \alpha_k T_1)^{-1}(x^k - \alpha_k D^T y^k),\\
\bar{x}_2^k &= (I + \alpha_k T_2)^{-1}(x^k - \alpha_k E^T y^k),\\
\bar{y}^k &= y^k + \alpha_k(Dx_1^k + Ex_2^k - b),\\
x_1^{k+1} &= \bar{x}_1^k - \alpha_k D^T(\bar{y}^k - y^k),\\
x_2^{k+1} &= \bar{x}_2^k - \alpha_k E^T(\bar{y}^k - y^k),\\
y^{k+1} &= \bar{y}^k + \alpha_k(D\bar{x}_1^k + E\bar{x}_2^k - Dx_1^k - Ex_2^k),
\end{aligned}$$

for $k = 0, 1, \ldots$.

Example 5. Consider the minimax problem:

$$\min_{x\in\Re^n} \max_{y\in Y} \{f(x) - g(y) + \langle y, Dx\rangle\},$$

where f is a closed proper convex functions on $\Re^n$, g is a continuously differentiable convex function on $\Re^m$, Y is a nonempty closed convex set in $\Re^m$, and $D \in \Re^{m\times n}$. Under a suitable constraint qualification [42, Chap. 37], this problem is equivalent to $0 \in A(x, y) + B(x, y)$, where A and B are the maximal monotone mappings

$$A(x, y) = (D^T y, \nabla g(y) - Dx), \qquad B(x, y) = \partial f(x) \times N_Y.$$

Applying the modified F-B method with this choice of A and B and with $X = X_1 \times Y$, where $X_1 = \Re^n$ if ∇g is Lipschitz continuous on $\Re^m$ and otherwise X_1 is a suitable closed convex subset of $\mathrm{dom}\partial f$, we obtain the following variant of a method in [59, Ex. 4]:

$$\begin{aligned}
\bar{x}^k &= \arg\min_{x\in\Re^n}\{f(x) + \langle y^k, Dx\rangle + \|x - x^k\|^2/(2\alpha_k)\},\\
\bar{y}^k &= P_Y[y^k + \alpha_k(Dx^k - \nabla g(y^k))],\\
x^{k+1} &= P_{X_1}[\bar{x}^k - \alpha_k D^T(\bar{y}^k - y^k)],\\
y^{k+1} &= P_Y[y^k + \alpha_k(D\bar{x}^k - \nabla g(\bar{y}^k) - Dx^k + \nabla g(y^k))],
\end{aligned}$$

for $k = 0, 1, \ldots$. As in the method of [59], if Y has a Cartesian product structure or f has a separable structure (e.g., $f(x_1, \ldots, x_n) = f_1(x_1) + \cdots + f_n(x_n)$ for some functions $f_1, \ldots, f_n$ on $\Re$), as in certain discrete-time deterministic optimal control problem [5, 46] and in the scheduling of hydroelectric power generation under uncertainty [48], then X_1 can be chosen to have a corresponding product structure and the computation of $\bar{x}^k$ and x^{k+1} decompose accordingly. See [59] for further discussions of the advantages of such decomposition methods.

Under additional assumptions on the problems, convergence and linear convergence of the methods in Examples 1–5 can be established by appropriately applying Theorem 3.4. (For $\mathcal{H} = \Re^n$, weak convergence and strong convergence are equivalent.) Additional applications are discussed in [5, 13, 56, 57]. Notice that the inclusion $0 \in A(x) + B(x)$ may be reformuated as $0 \in F(x, y_1, y_2) \times G(x, y_1, y_2)$, where

$$F(x, y_1, y_2) = (B(x) - y_2), \qquad G(x, y_1, y_2) = (A(y_1) + y_2) \times \{x - y_1\}.$$

Then, provided A is single-valued and continuous on $\mathcal{H} = \Re^n$, the method in [59] may be applied to this reformulated problem to obtain a method that has similar computation and convergence properties as, but is more complicated than, the modified F-B method. (The analysis in [59] is for the case $\mathcal{H} = \Re^n$, although extension to a Hilbert space setting seems possible.) Last, there recently has been much study of proximal point methods using a nonquadratic proximal term (see [18] and references therein), and it would be interesting to extend the modified F-B method to this setting.

REFERENCES

[1] D. P. Bertsekas, *Constrained Optimization and Lagrange Multiplier Methods*, Academic Press, New York, 1982.

[2] D. P. Bertsekas and E. M. Gafni, *Projection methods for variational inequalities with application to the traffic assignment problem*, Math. Programming Stud., 17 (1982), pp. 139–159.

[3] L. M. Bregman, *The method of successive projection for finding a common point of convex sets*, Akad. Nauk SSSR Doklady, 162 (1965), pp. 487–490; English translation in Soviet Math. Doklady, 162 (1965), pp. 688–692.

[4] H. Brézis, *Opérateurs Maximaux Monotones*, North-Holland, Amsterdam, 1973.

[5] H.-G. Chen, *Forward-Backward Splitting Techniques: Theory and Applications*, Ph.D. thesis, Department of Applied Mathematics, University of Washington, Seattle, WA, 1994.

[6] H.-G. Chen and R. T. Rockafellar, *Convergence rates in forward-backward splitting*, SIAM J. Optim., 7 (1997), pp. 421–444.

[7] G. Chen and M. Teboulle, *A proximal-based decomposition method for convex minimization problems*, Math. Programming, 64 (1994), pp. 81–101.

[8] J. Eckstein and D. P. Bertsekas, *On the Douglas–Rachford splitting method and the proximal point algorithm for maximal monotone operators*, Math. Programming, 55 (1992), pp. 293–318.

[9] J. Eckstein and M. C. Ferris, *Operator-splitting methods for monotone affine variational inequalities, with a parallel application to optimal control*, INFORMS J. Comput., 10 (1998), pp. 218–235.

[10] J. Eckstein and M. Fukushima, *Some reformulations and applications of the alternating direction method of multipliers*, in Large Scale Optimization: State of the Art, W. W. Hager, D. W. Hearn, and P. M. Pardalos, eds., Kluwer Academic Publishers, Dordrecht, The Netherlands, 1994, pp. 115–134.

[11] M. Fortin and R. Glowinski, Eds., *Augmented Lagrangian Methods: Applications to the Numerical Solution of Boundary Value Problems*, North-Holland, Amsterdam, 1983.

[12] D. Gabay, *Méthods Numériques pour l'Optimisation Nonlinéaire*, Thésis de Doctorat d'Etat et Science Mathématiques, Université Pierre et Marie Curie, Paris, 1979.

[13] D. Gabay, *Applications of the method of multipliers to variational inequalities*, in Augmented Lagrangian Methods: Applications to the Numerical Solution of Boundary Value Problems, M. Fortin and R. Glowinski, eds., North-Holland, Amsterdam, 1983, pp. 299–331.

[14] R. Glowinski and P. Le Tallec, *Augmented Lagrangian and Operator-Splitting Methods in Nonlinear Mechanics*, SIAM Stud. Appl. Math. 9, SIAM, Philadelphia, PA, 1989.

[15] O. Güler, *On the convergence of the proximal point algorithm for convex minimization*, SIAM J. Control Optim., 29 (1991), pp. 403–419.

[16] S. P. Han and G. Lou, *A parallel algorithm for a class of convex programs*, SIAM J. Control Optim., 26 (1988), pp. 345–355.

[17] A. N. Iusem, *An iterative algorithm for the variational inequality problem*, Mat. Apl. Comput., 13 (1994), pp. 103–114.

[18] K. C. Kiwiel, *Proximal minimization methods with generalized Bregman functions*, SIAM J. Control Optim., 35 (1997), pp. 1142–1168.

[19] G. M. Korpelevich, *The extragradient method for finding saddle points and other problems*, Matecon, 12 (1976), pp. 747–756.

[20] B. Lemaire, *Coupling optimization methods and variational convergence*, in Trends in Mathematical Optimization, K.-H. Hoffman, J.-B. Hiriart-Urruty, J. Zowe, and C. Lemarechal, eds., Birkhäuser-Verlag, Basel, 1988, pp. 163–179.

[21] B. Lemaire, *The proximal algorithm*, in New Methods in Optimization and Their Industrial Uses, Internat. Series Numer. Math. 87, Birkhäuser, Basel, Boston, 1989, pp. 73–87.

[22] P. L. Lions and B. Mercier, *Splitting algorithms for the sum of two nonlinear operators*, SIAM J. Numer. Anal., 16 (1979), pp. 964–979.

[23] Z.-Q. Luo and P. Tseng, *Error bounds and convergence analysis of feasible descent methods: A general approach*, Ann. Oper. Res., 46/47 (1993), pp. 157–178.

[24] X.-D. Luo and P. Tseng, *On a global projection-type error bound for the linear complementarity problem*, Linear Algebra Appl., 253 (1997), pp. 251–278.

[25] T. L. Magnanti and G. Perakis, *The orthogonality theorem and the strong-f-monotonicity condition for variational inequality algorithms*, SIAM J. Optim., 7 (1997), pp. 248-273.

[26] P. Marcotte, *Application of Khobotov's algorithm to variational inequalities and network equilibrium problems*, Inform. Systems Oper. Res., 29 (1991), pp. 258–270.

[27] B. Martinet, *Regularisation d'inéquations variationelles par approximations successives*, Rev. Française Informat. Recherche Opérationnelle, 4 (1970), pp. 154–158.

[28] B. Martinet, *Determination approchée d'un point fixe d'une application pseudo-contractante*, C. R. Acad. Sci. Paris Ser. A-B, 274 (1972), pp. 163–165.

[29] B. Mercier, *Lectures on Topics in Finite Element Solution of Elliptic Problems*, Lectures on Mathematics 63, Tata Institute of Fundamental Research, Bombay, 1979.

[30] B. Mercier, *Inéquations Variationnelles de la Mécanique*, Publ. Math. Orsay, 80.01, Université de Paris-Sud, Orsay, 1980.

[31] G. J. Minty, *On the maximal domain of a "monotone" function*, Michigan Math. J., 8 (1961), pp. 135–137.

[32] G. J. Minty, *Monotone (nonlinear) operators in Hilbert space*, Duke Math. J., 29 (1962), pp. 341–346.

[33] K. Mouallif, V. H. Nguyen, and J.-J. Strodiot, *A perturbed parallel decomposition method for a class of nonsmooth convex minimization problems*, SIAM J. Control Optim., 29 (1991), pp. 829–847.

[34] A. Moudafi and M. Théra, *Finding a zero of the sum of two maximal monotone operators*, J. Optim. Theory Appl., 94 (1997), pp. 425–448.

[35] J.-S. Pang, *A posteriori error bounds for the linearly-constrained variational inequality problem*, Math. Oper. Res., 12 (1987), pp. 474–484.

[36] D. Pascali and S. Sburlan, *Nonlinear Mappings of Monotone Type*, Editura Academiei, Bucharest, 1978.

[37] G. B. Passty, *Ergodic convergence to a zero of the sum of monotone operators in Hilbert space*, J. Math. Anal. Appl., 72 (1979), pp. 383–390.
[38] R. R. Phelps, *Convex Functions, Monotone Operators and Differentiability*, Springer-Verlag, New York, 1989.
[39] A. Renaud and G. Cohen, *Conditioning and regularization of nonsymmetric operators*, J. Optim. Theory Appl., 92 (1997), pp. 127–148.
[40] S. M. Robinson, *Some continuity properties of polyhedral multifunctions*, Math. Programming Stud., 14 (1981), pp. 206–214.
[41] R. T. Rockafellar, *Local boundedness of nonlinear monotone operators*, Michigan Math. J., 16 (1969), pp. 397–407.
[42] R. T. Rockafellar, *On the maximality of sums of nonlinear monotone operators*, Trans. Amer. Math. Soc., 149 (1970), pp. 75–88.
[43] R. T. Rockafellar, *Convex Analysis*, Princeton University Press, Princeton, NJ, 1970.
[44] R. T. Rockafellar, *Monotone operators and the proximal point algorithm*, SIAM J. Control Optim., 14 (1976), pp. 877–898.
[45] R. T. Rockafellar, *Augmented Lagrangians and applications of the proximal point algorithm in convex programming*, Math. Oper. Res., 1 (1976), pp. 97–116.
[46] R. T. Rockafellar and R. J.-B. Wets, *Generalized linear-quadratic problems of deterministic and stochastic optimal control in discrete time*, SIAM J. Control Optim., 28 (1990), pp. 810–822.
[47] R. T. Rockafellar, *Variational Analysis*, Springer-Verlag, Berlin, 1998.
[48] D. Salinger, *A Splitting Algorithm for Multistage Stochastic Programming with Application to Hydropower Scheduling*, Ph.D. thesis, Department of Applied Mathematics, University of Washington, Seattle, WA, 1996.
[49] M. Sibony, *Méthodes itératives pour les équations et inéquations aux dérivées partielles nonlinéares de type monotone*, Calcolo, 7 (1970), pp. 65–183.
[50] M. V. Solodov and B. F. Svaiter, *A hybrid projection-proximal point algorithm*, J. Convex Anal., 6 (1999), pp. 59–70.
[51] M. V. Solodov and B. F. Svaiter, *A Hybrid Approximate Extragradient-Proximal Point Algorithm Using the Enlargement of a Maximal Monotone Operator*, Report, Instituto de Matemática Pura e Aplicada, Rio de Janeiro, 1998; Set-Valued Anal., to appear.
[52] M. V. Solodov and P. Tseng, *Modified projection-type methods for monotone variational inequalities*, SIAM J. Control Optim., 34 (1996), pp. 1814-1830.
[53] J. E. Spingarn, *Applications of the method of partial inverses to convex programming decomposition*, Math. Programming, 32 (1985), pp. 199–223.
[54] D. Sun, *A new step-size skill for solving a class of nonlinear projection equations*, J. Comput. Math., 13 (1995), pp. 357–368.
[55] D. Sun, *A class of iterative methods for solving nonlinear projection equations*, J. Optim. Theory Appl., 91 (1996), pp. 123–140.
[56] P. Tseng, *Further applications of a splitting algorithm to decomposition in variational inequalities and convex programming*, Math. Programming, 48 (1990), pp. 249–263.
[57] P. Tseng, *Applications of a splitting algorithm to decomposition in convex programming and variational inequalities*, SIAM J. Control Optim., 29 (1991), pp. 119–138.
[58] P. Tseng, *On linear convergence of iterative methods for the variational inequality problem*, J. Comput. Appl. Math., 60 (1995), pp. 237–252.
[59] P. Tseng, *Alternating projection-proximal methods for convex programming and variational inequalities*, SIAM J. Optim., 7 (1997), pp. 951–965.
[60] E. H. Zarantonello, *Projections on convex sets in Hilbert space and spectral theory*, in Contributions to Nonlinear Functional Analysis, E. H. Zarantonello, ed., Academic Press, New York, 1971, pp. 237–424.
[61] E. Zeidler, *Nonlinear Functional Analysis and Its Applications: Nonlinear Monotone Operators*, II/B, Springer-Verlag, New York, 1990.

SIAM J. CONTROL OPTIM.
Vol. 38, No. 2, pp. 447–469

THE O.D.E. METHOD FOR CONVERGENCE OF STOCHASTIC APPROXIMATION AND REINFORCEMENT LEARNING*

V. S. BORKAR† AND S. P. MEYN‡

Abstract. It is shown here that stability of the stochastic approximation algorithm is implied by the asymptotic stability of the origin for an associated ODE. This in turn implies convergence of the algorithm. Several specific classes of algorithms are considered as applications. It is found that the results provide (i) a simpler derivation of known results for reinforcement learning algorithms; (ii) a proof for the first time that a class of asynchronous stochastic approximation algorithms are convergent without using any a priori assumption of stability; (iii) a proof for the first time that asynchronous adaptive critic and Q-learning algorithms are convergent for the average cost optimal control problem.

Key words. stochastic approximation, ODE method, stability, asynchronous algorithms, reinforcement learning

AMS subject classifications. 62L20, 93E25, 93E15

PII. S0363012997331639

1. Introduction.

The stochastic approximation algorithm considered in this paper is described by the d-dimensional recursion

$$X(n+1) = X(n) + a(n)\big[h\big(X(n)\big) + M(n+1)\big], \quad n \geq 0, \tag{1.1}$$

where $X(n) = [X_1(n), \ldots, X_d(n)]^T \in \mathbb{R}^d$, $h : \mathbb{R}^d \to \mathbb{R}^d$, and $\{a(n)\}$ is a sequence of positive numbers. The sequence $\{M(n) : n \geq 0\}$ is uncorrelated with zero mean.

Though more than four decades old, the stochastic approximation algorithm is now of renewed interest due to novel applications to reinforcement learning [20] and as a model of learning by boundedly rational economic agents [19]. Traditional convergence analysis usually shows that the recursion (1.1) will have the desired asymptotic behavior provided that the iterates remain bounded with probability one, or that they visit a prescribed bounded set infinitely often with probability one [3, 14]. Under such stability or recurrence conditions one can then approximate the sequence $\boldsymbol{X} = \{X(n) : n \geq 0\}$ with the solution to the ordinary differential equation (ODE)

$$\dot{x}(t) = h\big(x(t)\big) \tag{1.2}$$

with identical initial conditions $x(0) = X(0)$.

The recurrence assumption is crucial, and in many practical cases this becomes a bottleneck in applying the ODE method. The most successful technique for establishing stochastic stability is the stochastic Lyapunov function approach (see, e.g.,

*Received by the editors December 17, 1997; accepted for publication (in revised form) February 22, 1999; published electronically January 11, 2000.
http://www.siam.org/journals/sicon/38-2/33163.html

†School of Technology and Computer Science, Tata Institute of Fundamental Research, Mumbai 400005, India (borkar@tifr.res.in). The research of this author was supported in part by the Department of Science and Technology (Government of India) grant III5(12)/96-ET.

‡Department of Electrical and Computer Engineering and the Coordinated Sciences Laboratory, University of Illinois at Urbana-Champaign, Urbana, IL 61801 (s-meyn@uiuc.edu). The research of this author was supported in part by NSF grant ECS 940372 and JSEP grant N00014-90-J-1270. This research was completed while this author was a visiting scientist at the Indian Institute of Science under a Fulbright Research Fellowship.

[14]). One also has techniques based upon the contractive properties or homogeneity properties of the functions involved (see, e.g., [20] and [12], respectively).

The main contribution of this paper is to add to this collection another general technique for proving stability of the stochastic approximation method. This technique is inspired by the fluid model approach to stability of networks developed in [9, 10], which is itself based upon the multistep drift criterion of [15, 16]. The idea is that the usual stochastic Lyapunov function approach can be difficult to apply due to the fact that time-averaging of the noise may be necessary before a given positive valued function of the state process will decrease towards zero. In general such time-averaging of the noise will require infeasible calculation. In many models, however, it is possible to combine time-averaging with a limiting operation on the magnitude of the initial state, to replace the stochastic system of interest with a simpler deterministic process.

The scaling applied in this paper to approximate the model (1.1) with a deterministic process is similar to the construction of the fluid model of [9, 10]. Suppose that the state is scaled by its initial value to give $\widetilde{X}(n) = X(n)/\max(|X(0)|, 1)$, $n \geq 0$. We then scale time to obtain a continuous function $\phi : \mathbb{R}_+ \to \mathbb{R}^d$ which interpolates the values of $\{\widetilde{X}(n)\}$. At a sequence of times $\{t(j) : j \geq 0\}$ we set $\phi(t(j)) = \widetilde{X}(j)$, and for arbitrary $t \geq 0$, we extend the definition by linear interpolation. The times $\{t(j) : j \geq 0\}$ are defined in terms of the constants $\{a(j)\}$ used in (1.1). For any $r > 0$, the scaled function $h_r : \mathbb{R}^d \to \mathbb{R}^d$ is given by

$$h_r(x) = h(rx)/r, \quad x \in \mathbb{R}^d. \tag{1.3}$$

Then through elementary arguments we find that the stochastic process $\boldsymbol{\phi}$ approximates the solution $\widehat{\phi}$ to the associated ODE

$$\dot{x}(t) = h_r\big(x(t)\big), \quad t \geq 0, \tag{1.4}$$

with $\widehat{\phi}(0) = \phi(0)$ and $r = \max(|X(0)|, 1)$.

With our attention on stability considerations, we are most interested in the behavior of $\boldsymbol{X}$ when the magnitude of the initial condition $|X(0)|$ is large. Assuming that the limiting function $h_\infty = \lim_{r\to\infty} h_r$ exists, for large initial conditions we find that ϕ is approximated by the solution ϕ^∞ of the limiting ODE

$$\dot{x}(t) = h_\infty\big(x(t)\big), \tag{1.5}$$

where again we take identical initial conditions $\phi^\infty(0) = \phi(0)$.

Thus, for large initial conditions all three processes are approximately equal,

$$\phi \approx \widehat{\phi} \approx \phi^\infty.$$

Using these observations we find in Theorem 2.1 that the stochastic model (1.1) is stable in a strong sense provided the origin is asymptotically stable for the limiting ODE (1.5). Equation (1.5) is precisely the fluid model of [9, 10].

Thus, the major conclusion of this paper is that the ODE method can be extended to establish both the stability *and* convergence of the stochastic approximation method, as opposed to only the latter. The result [14, Theorem 4.1, p. 115] arrives at a similar conclusion: if the ODE (1.2) possesses a "global" Lyapunov function with bounded partial derivatives, then this will serve as a stochastic Lyapunov function, thereby establishing recurrence of the algorithm. Though similar in flavor, there are

significant differences between these results. First, in the present paper we consider a scaled ODE, not the usual ODE (1.2). The former retains only terms with dominant growth and is frequently simpler. Second, while it is possible that the stability of the scaled ODE and the usual one go hand in hand, this does not imply that a Lyapunov function for the latter is easily found. The reinforcement learning algorithms for ergodic-cost optimal control and asynchronous algorithms, both considered as applications of the theory in this paper, are examples where the scaled ODE is conveniently analyzed.

Though the assumptions made in this paper are explicitly motivated by applications to reinforcement learning algorithms for Markov decision processes, this approach is likely to find a broader range of applications.

The paper is organized as follows. The next section presents the main results for the stochastic approximation algorithm with vanishing stepsize or with bounded, nonvanishing stepsize. Section 2 also gives a useful error bound for the constant stepsize case and briefly sketches an extension to asynchronous algorithms, omitting details that can be found in [6]. Section 3 gives examples of algorithms for reinforcement learning of Markov decision processes to which this analysis is applicable. The proofs of the main results are collected together in section 4.

2. Main results. Here we collect together the main general results concerning the stochastic approximation algorithm. Proofs not included here may be found in section 4.

We shall impose the following additional conditions on the functions $\{h_r : r \geq 1\}$ defined in (1.3) and the sequence $\boldsymbol{M} = \{M(n) : n \geq 1\}$ used in (1.1). Some relaxations of assumption (A1) are discussed in section 2.4.

(A1) The function h is Lipschitz, and there exists a function $h_\infty : \mathbb{R}^d \to \mathbb{R}^d$ such that

$$\lim_{r\to\infty} h_r(x) = h_\infty(x), \quad x \in \mathbb{R}^d.$$

Furthermore, the origin in $\mathbb{R}^d$ is an asymptotically stable equilibrium for the ODE (1.5).

(A2) The sequence $\{M(n), \mathcal{F}_n : n \geq 1\}$, with $\mathcal{F}_n = \sigma(X(i), M(i), i \leq n)$, is a martingale difference sequence. Moreover, for some $C_0 < \infty$ and any initial condition $X(0) \in \mathbb{R}^d$,

$$\mathsf{E}\big[\|M(n+1)\|^2 \mid \mathcal{F}_n\big] \leq C_0\big(1 + \|X(n)\|^2\big), \quad n \geq 0.$$

The sequence $\{a(n)\}$ is deterministic and is assumed to satisfy one of the following two assumptions. Here TS stands for "tapering stepsize" and BS for "bounded stepsize."

(TS) The sequence $\{a(n)\}$ satisfies $0 < a(n) \leq 1,\ n \geq 0$, and

$$\sum_n a(n) = \infty, \qquad \sum_n a(n)^2 < \infty.$$

(BS) The sequence $\{a(n)\}$ satisfies for some constants $1 > \overline{\alpha} > \underline{\alpha} > 0$,

$$\underline{\alpha} \leq a(n) \leq \overline{\alpha}, \quad n \geq 0.$$

2.1. Stability and convergence. The first result shows that the algorithm is stabilizing for both bounded and tapering step sizes.

THEOREM 2.1. *Assume that* (A1) *and* (A2) *hold. Then we have the following:*

(i) *Under* (TS), *for any initial condition* $X(0) \in \mathbb{R}^d$,

$$\sup_n \|X(n)\| < \infty \qquad \textit{almost surely (a.s.).}$$

(ii) *Under* (BS) *there exist* $\alpha^* > 0$ *and* $C_1 < \infty$ *such that for all* $0 < \overline{\alpha} < \alpha^*$ *and* $X(0) \in \mathbb{R}^d$,

$$\limsup_{n\to\infty} \mathsf{E}\big[\big\|X(n)\big\|^2\big] \le C_1. \qquad \square$$

An immediate corollary to Theorem 2.1 is convergence of the algorithm under (TS). The proof is a standard application of the Hirsch lemma (see [11, Theorem 1, p. 339] or [3, 14]), but we give the details below for sake of completeness.

THEOREM 2.2. *Suppose that* (A1), (A2), *and* (TS) *hold and that the ODE* (1.2) *has a unique globally asymptotically stable equilibrium* x^*. *Then* $X(n) \to x^*$ *a.s. as* $n \to \infty$ *for any initial condition* $X(0) \in \mathbb{R}^d$.

Proof. We may suppose that $X(0)$ is deterministic without any loss of generality so that the conclusion of Theorem 2.1 (i) holds that the sample paths of $\boldsymbol{X}$ are bounded with probability one. Fixing such a sample path, we see that $\boldsymbol{X}$ remains in a bounded set H, which may be chosen so that $x^* \in \operatorname{int}(H)$.

The proof depends on an approximation of $\boldsymbol{X}$ with the solution to the primary ODE (1.2). To perform this approximation, first define $t(n) \uparrow \infty$, $T(n) \uparrow \infty$ as follows: Set $t(0) = T(0) = 0$ and for $n \ge 1$, $t(n) = \sum_{i=0}^{n-1} a(i)$. Fix $T > 0$ and define inductively

$$T(n+1) = \min\big\{t(j) : t(j) > T(n) + T\big\}, \quad n \ge 0.$$

Thus $T(n) = t(m(n))$ for some $m(n) \uparrow \infty$ and $T \le T(n+1) - T(n) \le T + 1$ for $n \ge 0$. We then define two functions from $\mathbb{R}_+$ to $\mathbb{R}^d$:

(a) $\{\psi(t), t > 0\}$ is defined by $\psi(t(n)) = X(n)$ with linear interpolation on $[t(n), t(n+1)]$ for each $n \ge 0$.

(b) $\{\widehat{\psi}(t), t > 0\}$ is piecewise continuous, defined so that, for any $j \ge 0$, $\widehat{\psi}$ is the solution to (1.2) for $t \in [T(j), T(j+1))$, with the initial condition $\widehat{\psi}(T(j)) = \psi(T(j))$.

Let $\epsilon > 0$ and let $B(\epsilon)$ denote the open ball centered at x^* of radius ϵ. We may then choose the following:

(i) $0 < \delta < \epsilon$ such that $x(t) \in B(\epsilon)$ for all $t \ge 0$ whenever $x(\,\cdot\,)$ is a solution of (1.2) satisfying $x(0) \in B(\delta)$.

(ii) $T > 0$ so large that for any solution of (1.2) with $x(0) \in H$ we have $x(t) \in B(\delta/2)$ for all $t \ge T$. Hence, $\widehat{\psi}(T(j)-) \in B(\delta/2)$ for all $j \ge 1$.

(iii) An application of the Bellman Gronwall lemma as in Lemma 4.6 below that leads to the limit

$$\big\|\psi(t) - \widehat{\psi}(t)\big\| \to 0 \qquad \text{a.s.}, \quad t \to \infty. \tag{2.1}$$

Hence we may choose $j_0 > 0$ so that we have

$$\big\|\psi\big(T(j) - \big) - \widehat{\psi}\big(T(j) - \big)\big\| \le \delta/2, \quad j \ge j_0.$$

Since $\psi(\,\cdot\,)$ is continuous, we conclude from (ii) and (iii) that $\psi(T(j)) \in B(\delta)$ for $j \ge j_0$. Since $\widehat{\psi}(T(j)) = \psi(T(j))$, it then follows from (i) that $\widehat{\psi}(t) \in B(\epsilon)$ for all $t \ge T(j_0)$. Hence by (2.1),

$$\limsup_{t\to\infty} \|\psi(t) - x^*\| \le \epsilon \qquad \text{a.s.}$$

This completes the proof since $\epsilon > 0$ was arbitrary. □

We now consider (BS), focusing on the absolute error defined by

$$e(n) := \|X(n) - x^*\|, \qquad n \ge 0. \tag{2.2}$$

THEOREM 2.3. *Assume that* (A1), (A2), *and* (BS) *hold, and suppose that* (1.2) *has a globally asymptotically stable equilibrium point* x^*.

Then for any $0 < \alpha \le \alpha^*$, *where* α^* *is introduced in Theorem* 2.1 (ii),

(i) *for any* $\epsilon > 0$, *there exists* $b_1 = b_1(\epsilon) < \infty$ *such that*

$$\limsup_{n\to\infty} \mathsf{P}\big(e(n) \ge \epsilon\big) \le b_1 \overline{\alpha};$$

(ii) *if* x^* *is a globally exponentially asymptotically stable equilibrium for the ODE* (1.2), *then there exists* $b_2 < \infty$ *such that for every initial condition* $X(0) \in \mathbb{R}^d$,

$$\limsup_{n\to\infty} \mathsf{E}\big[e(n)^2\big] \le b_2 \overline{\alpha}. \qquad \square$$

2.2. Rate of convergence. A uniform bound on the mean square error $\mathsf{E}[e(n)^2]$ for $n \ge 0$ can be obtained under slightly stronger conditions on $\boldsymbol{M}$ via the theory of ψ-irreducible Markov chains. We find that this error can be bounded from above by a sum of two terms: the first converges to zero as $\alpha \downarrow 0$, while the second decays to zero exponentially as $n \to \infty$.

To illustrate the nature of these bounds, consider the linear recursion

$$X(n+1) = X(n) + \alpha\big[-\big(X(n) - x^*\big) + W(n+1)\big], \quad n \ge 0,$$

where $\{W(n)\}$ is independently and identically distributed (i.i.d.) with mean zero and variance σ^2. This is of the form (1.1) with $h(x) = -(x - x^*)$ and $M(n) = W(n)$. The error $e(n+1)$ defined in (2.2) may be bounded as follows:

$$\begin{aligned}\mathsf{E}\big[e(n+1)^2\big] &\le \alpha^2\sigma^2 + (1-\alpha)^2\mathsf{E}\big[e(n)^2\big] \\ &\le \alpha\sigma^2/(2-\alpha) + \exp(-2\alpha n)\mathsf{E}\big[e(0)^2\big], \quad n \ge 0.\end{aligned}$$

For a deterministic initial condition $X(0) = x$ and any $\epsilon > 0$, we thus arrive at the formal bound,

$$\mathsf{E}[e(n)^2 \mid X(0) = x] \le B_1(\alpha) + B_2\big(\|x\|^2 + 1\big)\exp\big(-\epsilon_0(\alpha)n\big), \tag{2.3}$$

where B_1, B_2, and ϵ_0 are positive-valued functions of α. The bound (2.3) is of the form that we seek: the first term on the right-hand side (r.h.s.) decays to zero with α, while the second decays exponentially to zero with n. However, the rate of convergence for the second term becomes vanishingly small as $\alpha \downarrow 0$. Hence to maintain a small probability of error the variable α should be neither too small nor too large. This recalls the well-known trade-off between mean and variance that must be made in the application of stochastic approximation algorithms.

A bound of this form carries over to the nonlinear model under some additional conditions. For convenience, we take a Markov model of the form

$$X(n+1) = X(n) + \alpha\big[h\big(X(n)\big) + m\big(X(n), W(n+1)\big)\big], \tag{2.4}$$

where again $\{W(n)\}$ is i.i.d. and also independent of the initial condition $X(0)$. We assume that the functions $h : \mathbb{R}^d \to \mathbb{R}^d$ and $m : \mathbb{R}^d \times \mathbb{R}^q \to \mathbb{R}^d$ are smooth (C^1) and that assumptions (A1) and (A2) continue to hold. The recursion (2.4) then describes a Feller–Markov chain with stationary transition kernel to be denoted by P.

Let $V\colon \mathbb{R}^d \to [1,\infty)$ be given. The Markov chain $\boldsymbol{X}$ with transition function P is called *V-uniformly ergodic* if there is a unique invariant probability π, an $R < \infty$, and $\rho < 1$ such that for any function g satisfying $|g(x)| \le V(x)$,

$$\big|\mathsf{E}\big[g\big(X(n)\big) \mid X(0) = x\big] - \mathsf{E}_\pi\big[g\big(X(n)\big)\big]\big| \le RV(x)\rho^n, \qquad x \in \mathbb{R}^d, \quad n \ge 0, \tag{2.5}$$

where $\mathsf{E}_\pi[g(X(n))] = \int g(x)\,\pi(dx)$, $n \ge 0$.

The following result establishes bounds of the form (2.3) using V-ergodicity of the model. Assumptions (2.6) and (2.7) below are required to establish ψ-irreducibility of the model in Lemma 4.10.

There exists a $w^* \in \mathbb{R}^q$ with $m(x^*, w^*) = 0$, and for a continuous function $p : \mathbb{R}^q \to [0,1]$ with $p(w^*) > 0$,

$$\mathsf{P}\big(W(1) \in A\big) \ge \int_A p(z)dz, \quad A \in \mathcal{B}(\mathbb{R}^q). \tag{2.6}$$

The pair of matrices (F, G) is controllable with

$$F = \frac{d}{dx}h(x^*) + \frac{\partial}{\partial x}m(x^*, w^*) \quad \textit{and} \quad G = \frac{\partial}{\partial w}m(x^*, w^*). \tag{2.7}$$

THEOREM 2.4. *Suppose that* (A1), (A2), (2.6), *and* (2.7) *hold for the Markov model* (2.4) *with $0 < \alpha \le \alpha^*$. Then the Markov chain $\boldsymbol{X}$ is V-uniformly ergodic, with $V(x) = \|x\|^2 + 1$, and we have the following bounds:*

(i) *There exist positive-valued functions A_1 and ϵ_0 of α and a constant A_2 independent of α, such that*

$$\mathsf{P}\big\{e(n) \ge \epsilon \mid X(0) = x\big\} \le A_1(\alpha) + A_2\big(\|x\|^2 + 1\big)\exp\big(-\epsilon_0(\alpha)n\big).$$

The functions satisfy $A_1(\alpha) \to 0$, $\epsilon_0(\alpha) \to 0$ as $\alpha \downarrow 0$.

(ii) *If in addition the ODE* (1.2) *is exponentially asymptotically stable, then the stronger bound* (2.3) *holds, where again $B_1(\alpha) \to 0$, $\epsilon_0(\alpha) \to 0$ as $\alpha \downarrow 0$, and B_2 is independent of α.*

Proof. The V-uniform ergodicity is established in Lemma 4.10.

From Theorem 2.3 (i) we have, when $X(0) \sim \pi$,

$$\mathsf{P}_\pi\big(e(n) \ge \epsilon\big) = \mathsf{P}_\pi\big(e(0) \ge \epsilon\big) \le b_1\overline{\alpha},$$

and hence from V-uniform ergodicity,

$$\begin{aligned}\mathsf{P}\big(e(n) \ge \epsilon \mid X(0) = x\big) &\le \mathsf{P}_\pi\big(e(n) \ge \epsilon\big) + \big|\mathsf{P}\big(e(n) \ge \epsilon \mid X(0) = x\big) - \mathsf{P}_\pi\big(e(n) \ge \epsilon\big)\big| \\ &\le b_1\alpha + RV(x)\rho^n, \quad n \ge 0.\end{aligned}$$

This and the definition of V establishes (i). The proof of (ii) is similar.

The fact that $\rho = \rho_\alpha \to 1$ as $\alpha \downarrow 0$ is discussed in section 4.3. ∎

2.3. The asynchronous case. The conclusions above also extend to the model of asynchronous stochastic approximation analyzed in [6]. We now assume that each component of $X(n)$ is updated by a separate processor. We postulate a set-valued process $\{Y(n)\}$ taking values in the set of subsets of $\{1, 2, \dots, d\}$, with the interpretation: $Y(n) = \{\text{indices of the components updated at time } n\}$. For $n \geq 0$, $1 \leq i \leq d$, define

$$\nu(i, n) = \sum_{m=0}^{n} I\{i \in Y(m)\},$$

the number of updates executed by the ith processor up to time n. A key assumption is that there exists a deterministic $\Delta > 0$ such that for all i,

$$\liminf_{n\to\infty} \frac{\nu(i, n)}{n} \geq \Delta \qquad \text{a.s.}$$

This ensures that all components are updated comparably often. Furthermore, if

$$N(n, x) = \min\left\{m > n : \sum_{k=n+1}^{m} a(n) > x\right\}$$

for $x > 0$, the limit

$$\lim_{n\to\infty} \frac{\sum_{k=v(i,n)}^{v(i,N(n,x))} a(k)}{\sum_{k=v(j,n)}^{v(j,N(n,x))} a(k)}$$

exists a.s. for all i, j.

At time n, the kth processor has available the following data:

(i) Processor (k) is given $\nu(k, n)$, but it may not have n, the "global clock."

(ii) There are interprocessor communication delays $\tau_{kj}(n), 1 \leq k, j \leq d, n \geq 0$, so that at time n, processor (k) may use the data $X_j(m)$ only for $m \leq n - \tau_{kj}(n)$.

We assume that $\tau_{kk}(n) = 0$ for all n and that $\{\tau_{kj}(n)\}$ have a common upper bound $\overline{\tau} < \infty$ ([6] considers a slightly more general situation).

To relate the present work to [6], we recall that the "centralized" algorithm of [6] is

$$X(n+1) = X(n) + a(n) f\big(X(n), W(n+1)\big),$$

where $\{W(n)\}$ are i.i.d. and $\{f(\cdot, y)\}$ are uniformly Lipschitz. Thus $F(x) := \mathsf{E}[f(x, W(1))]$ is Lipschitz. The correspondence with the present set up is obtained by setting $h(x) = F(x)$ and

$$M(n+1) = f\big(X(n), W(n+1)\big) - F\big(X(n)\big)$$

for $n \geq 0$. The asynchronous version then is

$$\begin{aligned} X_i(n+1) = X_i(n) + a\big(\nu(i,n)\big) f\big(&X_1(n - \tau_{i1}(n)), X_2\big(n - \tau_{i2}(n)\big), \\ &\dots, X_d\big(n - \tau_{id}(n)\big), W(n+1)\big) I\{i \in Y(n)\}, \quad n \geq 0, \end{aligned} \tag{2.8}$$

for $1 \leq i \leq d$. Note that this can be executed by the ith processor without any knowledge of the global clock which, in fact, can be a complete artifice as long as causal relationships are respected.

The analysis presented in [6] depends upon the following additional conditions on $\{a(n)\}$:

(i) $a(n+1) \le a(n)$ eventually;

(ii) for $x \in (0,1)$, $\sup_n a([xn])/a(n) < \infty$;

(iii) for $x \in (0,1)$,

$$\left(\sum_{i=0}^{[xn]} a(i)\right) \Big/ \left(\sum_{i=0}^{n} a(i)\right) \to 1,$$

where $[\,\cdot\,]$ stands for "the integer part of $(\,\cdot\,)$."

A fourth condition is imposed in [6], but this becomes irrelevant when the delays are bounded. Examples of $\{a(n)\}$ satisfying (i)–(iii) are $a(n) = 1/(n+1)$ or $1/(1 + n\log(n+1))$.

As a first simplifying step, it is observed in [6] that $\{Y(n)\}$ may be assumed to be singletons without any loss of generality. We shall do likewise. What this entails is simply unfolding a single update at time n into $|Y(n)|$ separate updates, each involving a single component. This blows up the delays at most d-fold, which does not affect the analysis in any way.

The main result of [6] is the analog of our Theorem 2.2 *given* that the conclusions of our Theorem 2.1 hold. In other words, stability implies convergence. Under (A1) and (A2), our arguments above can be easily adapted to show that the conclusions of Theorem 2.2 also hold for the asynchronous case. One argues exactly as above and in [6] to conclude that the suitably interpolated and rescaled trajectory of the algorithm tracks an appropriate ODE. The only difference is a scalar factor $1/d$ multiplying the r.h.s. of the ODE (i.e., $\dot{x}(t) = (1/d)h(x(t))$). This factor, which reflects the asynchronous sampling, amounts to a time-scaling that does not affect the qualitative behavior of the ODE.

THEOREM 2.5. *Under the conditions of Theorem* 2.2 *and the above hypotheses on* $\{a(n)\}$, $\{Y(n)\}$, *and* $\{\tau_{ij}(n)\}$, *the asynchronous iterates given by* (3.7) *remain a.s. bounded and (therefore) converge to* x^* *a.s.* □

2.4. Further extensions. Although satisfied in all of the applications treated in section 3, in some other models assumption (A1) that $h_r \to h_\infty$ pointwise may be violated. If this convergence does not hold, then we may abandon the fluid model and replace (A1) by

(A1′) The function h is Lipschitz, and there exists $T > 0$, $R > 0$ such that

$$\left|\widehat{\phi}(t)\right| \le \frac{1}{2}, \quad t \ge T,$$

for any solution to (1.4) with $r \ge R$ and with initial condition satisfying $|\widehat{\phi}(0)| = 1$.

Under the Lipschitz condition on h, at worst we may find that the pointwise limits of $\{h_r : r \ge 1\}$ will form a family Λ of Lipschitz functions on $\mathbb{R}^d$. That is, $h_\infty \in \Lambda$ if and only if there exists a sequence $\{r_i\} \uparrow \infty$ such that

$$h_{r_i}(x) \to h_\infty(x), \quad i \to \infty,$$

where the convergence is uniform for x in compact subsets of $\mathbb{R}^d$. Under (A1′) we then find, using the same arguments as in the proof of Lemma 4.1, that the family Λ is uniformly stable.

LEMMA 2.6. *Under* (A1′) *the family of ODEs defined via* Λ *is uniformly exponentially asymptotically stable in the following sense. For some* $b < \infty$, $\delta > 0$, *and any solution* ϕ^∞ *to the ODE* (1.5) with $h_\infty \in \Lambda$,

$$|\phi^\infty(t)| \le be^{-\delta t}|\phi^\infty(0)|, \quad t \ge 0. \qquad \square$$

Using this lemma the development of section 4 goes through with virtually no changes, and hence Theorems 2.1–2.5 are valid with (A1) replaced by (A1′).

Another extension is to broaden the class of scalings. Consider a nonlinear scaling defined by a function $g\colon \mathbb{R}_+ \to \mathbb{R}_+$ satisfying $g(r)/r \to \infty$ as $r \to \infty$, and suppose that $h_r(\cdot)$ redefined as $h_r(x) = h(rx)/g(r)$ satisfies

$$h_r(x) \to h_\infty(x) \text{ uniformly on compacts as } r \to \infty.$$

Then, assuming that the a.s. boundedness of rescaled iterates can be separately established, a completely analogous development of the stochastic algorithm is possible. An example would be a "stochastic gradient" scheme, where $h(\cdot)$ is the gradient of an even degree polynomial, with degree, say, $2n$. Then $g(r) = r^{2n-1}$ will do. We do not pursue this further because the reinforcement learning algorithms we consider below do conform to the case $g(r) = r$.

3. Reinforcement learning. As both an illustration of the theory and an important application in its own right, in this section we analyze reinforcement learning algorithms for Markov decision processes. The reader is referred to [4] for a general background of the subject and to other references listed below for further details.

3.1. Markov decision processes. We consider a Markov decision process $\boldsymbol{\Phi} = \{\Phi(t) : t \in \mathbb{Z}\}$ taking values in a finite state space $S = \{1, 2, \dots, s\}$ and controlled by a control sequence $\boldsymbol{Z} = \{Z(t) : t \in \mathbb{Z}\}$ taking values in a finite action space $A = \{a_0, \dots, a_r\}$. We assume that the control sequence is *admissible* in the sense that $Z(n) \in \sigma\{\Phi(t) : t \le n\}$ for each n. We are most interested in stationary policies of the form $Z(t) = w(\Phi(t))$, where the *feedback law* w is a function $w\colon S \to A$. The controlled transition probabilities are given by $p(i, j, a)$ for $i, j \in S, a \in A$.

Let $c : S \times A \to R$ be the one-step cost function, and consider first the infinite horizon discounted cost control problem of minimizing over all admissible $\boldsymbol{Z}$ the total discounted cost

$$J(i, \boldsymbol{Z}) = \mathsf{E}\left[\sum_{t=0}^{\infty} \beta^t c\big(\Phi(t), Z(t)\big) \mid \Phi(0) = i\right],$$

where $\beta \in (0, 1)$ is the discount factor. The minimal value function is defined as

$$V(i) = \min J(i, \boldsymbol{Z}),$$

where the minimum is over all admissible control sequences $\boldsymbol{Z}$. The function V satisfies the dynamic programming equation

$$V(i) = \min_a \left[c(i, a) + \beta \sum_j p(i, j, a)V(j)\right], \qquad i \in S,$$

and the optimal control minimizing J is given as the stationary policy defined through the feedback law w^* given as any solution to

$$w^*(i) := \arg\min_a \left[c(i, a) + \beta \sum_j p(i, j, a)V(j)\right], \quad i \in S.$$

The value iteration algorithm is an iterative procedure to compute the minimal value function. Given an initial function $V_0 \colon S \to \mathbb{R}_+$ one obtains a sequence of functions $\{V_n\}$ through the recursion

$$V_{n+1}(i) = \min_a \left[c(i,a) + \beta \sum_j p(i,j,a) V_n(j) \right], \quad i \in S, \quad n \geq 0. \tag{3.1}$$

This recursion is convergent for any initialization $V_0 \geq 0$. If we define *Q-values* via

$$Q(i,a) = c(i,a) + \beta \sum_j p(i,j,a) V(j), \qquad i \in S, \quad a \in A,$$

then $V(i) = \min_a Q(i,a)$ and the matrix Q satisfies

$$Q(i,a) = c(i,a) + \beta \sum_j p(i,j,a) \min_b Q(j,b), \qquad i \in S, \quad a \in A.$$

The matrix Q can also be computed using the equivalent formulation of value iteration,

$$Q_{n+1}(i,a) = c(i,a) + \beta \sum_j p(i,j,a) \min_b Q_n(j,b), \qquad i \in S, \quad a \in A, \quad n \geq 0, \tag{3.2}$$

where $Q_0 \geq 0$ is arbitrary.

The value iteration algorithm is initialized with a function $V_0 \colon S \to \mathbb{R}_+$. In contrast, the *policy iteration algorithm* is initialized with a feedback law w^0 and generates a sequence of feedback laws $\{w^n : n \geq 0\}$. At the nth stage of the algorithm a feedback law w^n is given and the value function for the resulting control sequence $\boldsymbol{Z}^n = \{w^n(\Phi(0)), w^n(\Phi(1)), w^n(\Phi(2)), \dots\}$ is computed to give

$$J_n(i) = J\big(i, \boldsymbol{Z}^n\big), \qquad i \in S.$$

Interpreted as a column vector in $\mathbb{R}^s$, the vector J_n satisfies the equation

$$\big(I - \beta P_n\big) J_n = c_n, \tag{3.3}$$

where the $s \times s$ matrix P_n is defined by $P_n(i,j) = p(i,j,w^n(i))$, $i,j \in S$, and the column vector c_n is given by $c_n(i) = c(i, w^n(i))$, $i \in S$. Equation (3.3) can be solved for fixed n by the "fixed-policy" version of value iteration given by

$$J_n(i+1) = \beta P_n J_n(i) + c_n, \quad i \geq 0, \tag{3.4}$$

where $J_n(0) \in \mathbb{R}^s$ is given as an initial condition. Then $J_n(i) \to J_n$, the solution to (3.3), at a geometric rate as $i \to \infty$.

Given J_n, the next feedback law w^{n+1} is then computed via

$$w^{n+1}(i) = \arg\min_a \left[c(i,a) + \beta \sum_j p(i,j,a) J_n(j) \right], \quad i \in S. \tag{3.5}$$

Each step of the policy iteration algorithm is computationally intensive for large state spaces since the computation of J_n requires the inversion of the $s \times s$ matrix $I - \beta P_n$.

In the average cost optimization problem one seeks to minimize over all admissible $\boldsymbol{Z}$,

$$\limsup_{n \to \infty} \frac{1}{n} \sum_{t=0}^{n-1} \mathsf{E}\big[c\big(\Phi(t), Z(t)\big)\big]. \tag{3.6}$$

The policy iteration and value iteration algorithms to solve this optimization problem remain unchanged with three exceptions. One is that the constant β must be set equal to unity in (3.1) and (3.5). Second, in the policy iteration algorithm the value function J_n is replaced by a solution J_n to Poisson's equation

$$\sum p\big(i,j,w^n(i)\big)J_n(j) = J_n(i) - c\big(i,w^n(i)\big) + \eta_n, \quad i \in S,$$

where η_n is the steady state cost under the policy w^n. The computation of J_n and η_n again involves matrix inversions via

$$\pi_n\big(I - P_n + ee'\big) = e', \qquad \eta_n = \pi_n c_n, \qquad \big(I - P_n + ee'\big)J_n = c_n,$$

where $e \in \mathbb{R}^s$ is the column vector consisting of all ones and the row vector π_n is the invariant probability for P_n. The introduction of the outer product ensures that the matrix $(I - P_n + ee')$ is invertible, provided that the invariant probability π_n is unique.

Lastly, the value iteration algorithm is replaced by the "relative value iteration," where a common scalar offset is subtracted from all components of the iterates at each iteration (likewise for the Q-value iteration). The choice of this offset term is not unique. We shall be considering one particular choice, though others can be handled similarly (see [1]).

3.2. *Q*-learning. If the matrix Q defined in (3.2) can be computed via value iteration or some other scheme, then the optimal control is found through a simple minimization. If transition probabilities are unknown so that value iteration is not directly applicable, one may apply a stochastic approximation variant known as the *Q-learning algorithm* of Watkins [1, 20, 21]. This is defined through the recursion

$$Q_{n+1}(i,a) = Q_n(i,a) + a(n)\Big[\beta \min_b Q_n(\Psi_{n+1}(i,a),b) + c(i,a) - Q_n(i,a)\Big],$$

$i \in S, a \in A$, where $\Psi_{n+1}(i,a)$ is an independently simulated S-valued random variable with law $p(i,\cdot,a)$.

Making the appropriate correspondences with our set up, we have $X(n) = Q_n$ and $h(Q) = [h_{ia}(Q)]_{i,a}$ with

$$h_{ia}(Q) = \beta \sum_j p(i,j,a) \min_b Q(j,b) + c(i,a) - Q(i,a), \qquad i \in S, \quad a \in A.$$

The martingale is given by $M(n+1) = [M_{ia}(n+1)]_{i,a}$ with

$$M_{ia}(n+1) = \beta \left(\min_b Q_n(\Psi_{n+1}(i,a),b) - \sum_j p(i,j,a)\Big(\min_b Q_n(j,b)\Big) \right), \quad i \in S, \quad a \in A.$$

Define $F(Q) = [F_{ia}(Q)]_{i,a}$ by

$$F_{ia}(Q) = \beta \sum_j p(i,j,a) \min_b Q(j,b) \ + \ c(i,a).$$

Then $h(Q) = F(Q) - Q$ and the associated ODE is

$$\dot{Q} = F(Q) - Q := h(Q). \tag{3.7}$$

The map $F : \mathbb{R}^{s\times(r+1)} \to \mathbb{R}^{s\times(r+1)}$ is a contraction with respect to the max norm $\|\cdot\|_\infty$. The global asymptotic stability of its unique equilibrium point is a special case of the results of [8]. This $h(\cdot)$ fits the framework of our analysis, with the (i,a)th component of $h_\infty(Q)$ given by

$$\beta \sum_j p(i,j,a) \min_b Q(j,b) - Q(i,a), \quad i \in S, \quad a \in A.$$

This also is of the form $h_\infty(Q) = F_\infty(Q) - Q$ where $F_\infty(\cdot)$ is an $\|\cdot\|_\infty$- contraction, and thus the asymptotic stability of the unique equilibrium point of the corresponding ODE is guaranteed (see [8]). We conclude that assumptions (A1) and (A2) hold, and hence Theorems 2.1–2.4 also hold for the Q-learning model.

3.3. Adaptive critic algorithm. Next we shall consider the *adaptive critic algorithm*, which may be considered as the reinforcement learning analog of policy iteration (see [2, 13] for a discussion). There are several variants of this, one of which, taken from [13], is as follows. For $i \in S$, we define

$$V_{n+1}(i) = V_n(i) + b(n)\big[c\big(i,\psi_n(i)\big) + \beta V_n\big(\Psi_n\big(i,\psi_n(i)\big)\big) - V_n(i)\big], \tag{3.8}$$

from which the policies are updated according to

$$\widehat{w}_{n+1}(i) = \Gamma\left\{\widehat{w}_n(i) + a(n)\sum_{\ell=1}^{r}\Big(\big[c(i,a_0) + \beta V_n\big(\eta_n(i,a_0)\big)\big] - [c(i,a_\ell) + \beta V_n(\eta_n(i,a_\ell))]e_\ell\Big)\right\}. \tag{3.9}$$

Here $\{V_n\}$ are s-vectors and for each $i, \{\widehat{w}_n(i)\}$ are r-vectors lying in the simplex $\{x \in \mathbb{R}^r \mid x = [x_1,\dots,x_r], x_i \ge 0, \sum_i x_i \le 1\}$. $\Gamma(\cdot)$ is the projection onto this simplex. The sequences $\{a(n)\}, \{b(n)\}$ satisfy

$$\sum_n a(n) = \sum_n b(n) = \infty, \qquad \sum_n \big(a(n)^2 + b(n)^2\big) < \infty, \qquad a(n) = o\big(b(n)\big).$$

The rest of the notation is as follows. For $1 \le \ell \le r, e_\ell$ is the unit r-vector in the ℓth coordinate direction. For each i, n, $w_n(i) = w_n(i,\cdot)$ is a probability vector on A defined by the following. For $\widehat{w}_n(i) = [\widehat{w}_n(i,1),\dots,\widehat{w}_n(i,r)]$,

$$w_n(i,a_\ell) = \begin{cases} \widehat{w}_n(i,\ell) & \text{for } \ell \ne 0, \\ 1 - \displaystyle\sum_{j\ne 0}\widehat{w}_n(i,j) & \text{for } \ell = 0. \end{cases}$$

Given $w_n(i)$, $\psi_n(i)$ is an A-valued random variable independently simulated with law $w_n(i)$. Likewise, $\Psi_n(i,\psi_n(i))$ are S-valued random variables which are independently simulated (given $\psi_n(i)$) with law $p(i,\cdot,\psi_n(i))$ and $\{\eta_n(i,a_\ell)\}$ are S-valued random variables independently simulated with law $p(i,\cdot,a_\ell)$, respectively.

To see why this is based on policy iteration, recall that policy iteration alternates between two steps. One step solves the linear system of (3.3) to compute the fixed-policy value function corresponding to the current policy. We have seen that solving (3.3) can be accomplished by performing the fixed-policy version of value iteration given in (3.4). The first step (3.8) in the above iteration is indeed the "learning" or

"simulation-based stochastic approximation" analog of this fixed-policy value iteration. The second step in policy iteration updates the current policy by performing an appropriate minimization. The second iteration (3.9) is a particular search algorithm for computing this minimum over the simplex of probability measures on A. This search algorithm is by no means unique; the paper [13] gives two alternative schemes. However, the first iteration (3.8) is common to all.

The different choices of stepsize schedules for the two iterations (3.8) and (3.9) induces the "two time-scale" effect discussed in [5]. The first iteration sees the policy computed by the second as nearly static, thus justifying viewing it as a fixed-policy iteration. In turn, the second sees the first as almost equilibrated, justifying the search scheme for minimization over A. See [13] for details.

The boundedness of $\{\widehat{w}_n\}$ is guaranteed by the projection $\Gamma(\,\cdot\,)$. For $\{V_n\}$, the fact that $b(n) = o(a(n))$ allows one to treat $\widehat{w}_n(i)$ as constant, say, $\overline{w}(i)$; see, e.g., [13]. The appropriate ODE then turns out to be

$$\dot{v} = G(v) - v := h(v), \tag{3.10}$$

where $G : \mathbb{R}^s \to \mathbb{R}^s$ is defined by

$$G_i(x) = \sum_{\ell} \overline{w}(i, a_\ell) \left[\beta \sum_j p(i, j, a_\ell) x_j + c(i, a_\ell) \right] - x_i, \quad i \in S.$$

Once again, $G(\,\cdot\,)$ is an $\|\,\cdot\,\|_\infty$-contraction and it follows from the results of [8] that (3.10) is globally asymptotically stable. The limiting function $h_\infty(x)$ is again of the form $h_\infty(x) = G_\infty(x) - x$ with $G_\infty(x)$ defined so that its ith component is

$$\sum_{\ell} \overline{w}(i, a_\ell) \left[\beta \sum_j p(i, j, a_\ell) x_j \right] - x_i.$$

We see that G_∞ is also a $\|\,\cdot\,\|_\infty$-contraction and the global asymptotic stability of the origin for the corresponding limiting ODE follows as before from the results of [8].

3.4. Average cost optimal control. For the average cost control problem, we impose the additional restriction that the chain $\mathbf{\Phi}$ has a *unique* invariant probability measure under any stationary policy so that the steady state cost (3.6) is independent of the initial condition.

For the average cost optimal control problem, the Q-learning algorithm is given by the recursion

$$Q_{n+1}(i, a) = Q_n(i, a) + a(n)\Big(\min_b Q_n(\Psi_n(i, a), b) + c(i, a) - Q_n(i, a) - Q_n(i_0, a_0)\Big),$$

where $i_0 \in S$, $a_0 \in A$ are fixed a priori. The appropriate ODE now is (3.7) with $F(\,\cdot\,)$ redefined as $F_{ia}(Q) = \sum_j p(i, j, a) \min_b Q(j, b) + c(i, a) - Q(i, a) - Q(i_0, a_0)$. The global asymptotic stability for the unique equilibrium point for this ODE has been established in [1]. Once again this fits our framework with $h_\infty(x) = F_\infty(x) - x$ for F_∞ defined the same way as F, except for the terms $c(\cdot, \cdot)$ which are dropped. We conclude that (A1) and (A2) are satisfied for this version of the Q-learning algorithm.

Another variant of Q-learning for average cost, based on a "stochastic shortest path" formulation, is presented in [1]. This also can be handled similarly.

In [13], three variants of the adaptive critic algorithm for the average cost problem are discussed, differing only in the $\{\widehat{w}_n\}$ iteration. The iteration for $\{V_n\}$ is common to all and is given by

$$V_{n+1}(i) = V_n(i) + b(n)\big[c\big(i,\psi_n(i)\big) + V_n\big(\Psi_n\big(i,\psi_n,(i)\big)\big) - V_n(i) - V_n(i_0)\big], \quad i \in S,$$

where $i_0 \in S$ is a fixed state prescribed beforehand. This leads to the ODE (3.10) with G redefined as

$$G_i(x) = \sum_{\ell} \overline{w}(i,a_\ell)\left(\sum_j p(i,j,a_\ell)x_j + c(i,a_\ell)\right) - x_i - x_{i_0}, \quad i \in S.$$

The global asymptotic stability of the unique equilibrium point of this ODE has been established in [7]. Once more, this fits our framework with $h_\infty(x) = G_\infty(x) - x$ for G_∞ defined just like G, but without the $c(\cdot,\cdot)$ terms.

Asynchronous versions of all the above can be written down along the lines of (3.7). Then by Theorem 2.5, they have bounded iterates a.s. The important point to note here is that to date, a.s. boundedness for Q-learning and adaptive critic is proved by other methods for centralized algorithms [1, 12, 20]. For asynchronous algorithms, it is proved for discounted cost only [1, 13, 20] or by introducing a projection to enforce stability [14].

4. Derivations. Here we provide proofs for the main results given in section 2. Throughout this section we assume that (A1) and (A2) hold.

4.1. Stability. The functions $\{h_r, r \geq 1\}$ and the limiting function h_∞ are Lipschitz with the same Lipschitz constant as h under (A1). It follows from Ascoli's theorem that the convergence $h_r \to h_\infty$ is uniform on compact subsets of $\mathbb{R}^d$. This observation is the basis of the following lemma.

LEMMA 4.1. *Under* (A1)*, the ODE* (1.5) *is globally exponentially asymptotically stable.*

Proof. The function h_∞ satisfies

$$h_\infty(cx) = ch_\infty(x), \qquad c > 0, \quad x \in \mathbb{R}^d.$$

Hence the origin $\theta \in \mathbb{R}^d$ is an equilibrium for (1.5), i.e., $h_\infty(\theta) = \theta$. Let $B(\epsilon)$ be the closed ball of radius ϵ centered at θ with ϵ chosen so that $x(t) \to \theta$ as $t \to \infty$ uniformly for initial conditions in $B(\epsilon)$. Thus there exists a $T > 0$ such that $\|x(T)\| \leq \epsilon/2$ whenever $\|x(0)\| \leq \epsilon$. For an arbitrary solution $x(\,\cdot\,)$ of (1.5), $y(\,\cdot\,) = \epsilon x(\,\cdot\,)/\|x(0)\|$ is another, with $\|y(0)\| = \epsilon$. Hence $\|y(T)\| < \epsilon/2$, implying $\|x(T)\| \leq \frac{1}{2}\|x(0)\|$. The global exponential asymptotic stability follows. ∎

With the scaling parameter r given by $r(j) = \max(1, \|X(m(j))\|)$, $j \geq 0$, we define three piecewise continuous functions from $\mathbb{R}_+$ to $\mathbb{R}^d$ as in the introduction:

(a) $\{\phi(t) : t \geq 0\}$ is an interpolated version of $\boldsymbol{X}$ defined as follows. For each $j \geq 0$, define a function ϕ_j on the interval $[T(j), T(j+1)]$ by

$$\phi_j\big(t(n)\big) = X(n)/r(j), \quad m(j) \leq n \leq m(j+1),$$

with $\phi_j(\,\cdot\,)$ defined by linear interpolation on the remainder of $[T(j), T(j+1)]$ to form a piecewise linear function.

We then define ϕ to be the piecewise continuous function

$$\phi(t) = \phi_j(t), \qquad t \in \big[T(j), T(j+1)\big), \quad j \geq 0.$$

(b) $\{\widehat{\phi}(t) : t \geq 0\}$ is continuous on each interval $[T(j), T(j+1))$, and on this interval it is the solution to the ODE

$$\dot{x}(t) = h_{r(j)}\big(x(t)\big), \tag{4.1}$$

with initial condition $\widehat{\phi}(T(j)) = \phi(T(j))$, $j \geq 0$.

(c) $\{\phi^\infty(t) : t \geq 0\}$ is also continuous on each interval $[T(j), T(j+1))$, and on this interval it is the solution to the "fluid model" (1.5) with the same initial condition

$$\phi^\infty\big(T(j)\big) = \widehat{\phi}\big(T(j)\big) = \phi\big(T(j)\big) \quad j \geq 0.$$

Boundedness of $\widehat{\phi}(\,\cdot\,)$ and $\phi^\infty(\,\cdot\,)$ is crucial in deriving useful approximations.

LEMMA 4.2. *Under* (A1) *and* (A2) *and either* (TS) *or* (BS), *there exists* $\bar{C} < \infty$ *such that for any initial condition* $X(0) \in \mathbb{R}^d$

$$\widehat{\phi}(t) \leq \bar{C} \quad \textit{and} \quad \phi^\infty(t) \leq \bar{C}, \qquad t \geq 0.$$

Proof. To establish the first bound use the Lipschitz continuity of h to obtain the bound

$$\frac{d}{dt}\big\|\widehat{\phi}(t)\big\|^2 = 2\widehat{\phi}(t)^T h_{r(j)}\big(\widehat{\phi}(t)\big) \leq C\big(\big\|\widehat{\phi}(t)\big\|^2 + 1\big), \quad T(j) \leq t < T(j+1),$$

where C is a deterministic constant, independent of j. The claim follows with $\bar{C} = 2\exp((T+1)C)$ since $\|\widehat{\phi}(T(j))\| \leq 1$. The proof of the second bound is therefore identical. ∎

The following version of the Bellman Gronwall lemma will be used repeatedly.

LEMMA 4.3.

(i) *Suppose* $\{\alpha(n)\}$, $\{A(n)\}$ *are nonnegative sequences and* $\beta > 0$ *such that*

$$A(n+1) \leq \beta + \sum_{k=0}^{n} \alpha(k)A(k), \quad n \geq 0.$$

Then for all $n \geq 1$,

$$A(n+1) \leq \exp\left(\sum_{k=1}^{n} \alpha(k)\right)\big(\alpha(0)A(0) + \beta\big).$$

(ii) *Suppose* $\{\alpha(n)\}$, $\{A(n)\}$, $\{\gamma(n)\}$ *are nonnegative sequences such that*

$$A(n+1) \leq \big(1 + \alpha(n)\big)A(n) + \gamma(n), \quad n \geq 0.$$

Then for all $n \geq 1$,

$$A(n+1) \leq \exp\left(\sum_{k=1}^{n} \alpha(k)\right)\big(\big(1 + \alpha(0)\big)A(0) + \beta(n)\big),$$

where $\beta(n) = \sum_0^n \gamma(k)$.

Proof. Define $\{R(n)\}$ inductively by $R(0) = A(0)$ and

$$R(n+1) = \beta + \sum_{k=0}^{n} \alpha(k)R(k), \quad n \geq 0.$$

A simple induction shows that $A(n) \le R(n), \ \ n \ge 0$. An alternative expression for $R(n)$ is

$$R(n) = \left(\prod_{k=1}^{n} (1 + \alpha(k)) \right) \big(\alpha(0)A(0) + \beta \big).$$

The inequality (i) then follows from the bound $1 + x \le e^x$.

To see (ii) fix $n \ge 0$ and observe that on summing both sides of the bound

$$A(k+1) - A(k) \le \alpha(k)A(k) + \gamma(k)$$

over $0 \le k \le \ell$ we obtain for all $0 \le \ell < n$,

$$A(\ell + 1) \le A(0) + \beta(n) + \sum_{k=0}^{\ell} \alpha(k)A(k).$$

The result then follows from (i). □

The following lemmas relate the three functions $\phi(\cdot)$, $\widehat{\phi}(\cdot)$, and $\phi^\infty(\cdot)$.

LEMMA 4.4. *Suppose that* (A1) *and* (A2) *hold. Given any $\epsilon > 0$, there exist $T, R < \infty$ such that for any $r > R$ and any solution to the ODE* (1.4) *satisfying $\|x(0)\| \le 1$, we have $\|x(t)\| \le \epsilon$ for $t \in [T, T+1]$.*

Proof. By global asymptotic stability of (1.5) we can find $T > 0$ such that $\|\phi^\infty(t)\| \le \epsilon/2$, $t \ge T$, for solutions $\phi^\infty(\cdot)$ of (1.5) satisfying $\|\phi^\infty(0)\| \le 1$.

With T fixed, choose R so large that $|\widehat{\phi}(t) - \phi^\infty(t)| \le \epsilon/2$ whenever $\widehat{\phi}$ is a solution to (1.4) satisfying $\widehat{\phi}(0) = \phi^\infty(0)$; $|\widehat{\phi}(0)| \le 1$; and $r \ge R$. This is possible since, as we have already observed, $h_r \to h_\infty$ as $r \to \infty$ uniformly on compact sets. The claim then follows from the triangle inequality. □

Define the following: For $j \ge 0$, $m(j) \le n < m(j+1)$,

$$\widetilde{X}(n) := X(n)/r(j),$$
$$\widetilde{M}(n+1) := M(n+1)/r(j),$$

and for $n \ge 1$,

$$\xi(n) := \sum_{m=0}^{n-1} a(m)\widetilde{M}(m+1).$$

LEMMA 4.5. *Under* (A1), (A2), *and either* (TS) *or* (BS), *for each initial condition $X(0) \in \mathbb{R}^d$ satisfying $\mathsf{E}[\|X(0)\|^2] < \infty$, we have the following:*

(i) $\sup_{n \ge 0} \mathsf{E}[\|\widetilde{X}(n)\|^2] < \infty$.

(ii) $\sup_{j \ge 0} \mathsf{E}[\|X(m(j+1))/r(j)\|^2] < \infty$.

(iii) $\sup_{j \ge 0, T(j) \le t \le T(j+1)} \mathsf{E}[\|\phi(t)\|^2] < \infty$.

(iv) *Under* (TS) *the sequence $\{\xi(n), \mathcal{F}_n\}$ is a square integrable martingale with*

$$\sup_{n \ge 0} \mathsf{E}[\|\xi(n)\|^2] < \infty.$$

Proof. To prove (i) note first that under (A2) and the Lipschitz condition on h there exists $C < \infty$ such that for all $n \ge 1$,

$$\mathsf{E}\big[\|X(n)\|^2 \mid \mathcal{F}_{n-1}\big] \le \big(1 + Ca(n-1)\big)\|X(n-1)\|^2 + Ca(n-1), \quad n \ge 0. \tag{4.2}$$

It then follows that for any $j \geq 0$ and any $m(j) < n < m(j+1)$,

$$\mathsf{E}\big[\big\|\widetilde{X}(n)\big\|^2 \mid \mathcal{F}_{n-1}\big] \leq \big(1 + Ca(n-1)\big)\big\|\widetilde{X}(n-1)\big\|^2 + Ca(n-1),$$

so that by Lemma 4.3 (ii), for all such n,

$$\begin{aligned}\mathsf{E}\big[\big\|\widetilde{X}(n+1)\big\|^2\big] &\leq \exp\big(C(T+1)\big)\big(2\mathsf{E}\big[\big\|\widetilde{X}(m(j))\big\|^2\big] + C(T+1)\big)\\ &\leq \exp\big(C(T+1)\big)\big(2 + C(T+1)\big).\end{aligned}$$

Claim (i) follows, and claim (ii) follows similarly. We then obtain claim (iii) from the definition of $\phi(\,\cdot\,)$. From (i), (ii), and (A2), we have $\sup_n \mathsf{E}[\|\widetilde{M}(n)\|^2] < \infty$. Using this and the square summability of $\{a(n)\}$ assumed in (TS), the bound (iv) immediately follows. □

LEMMA 4.6. *Suppose* $\mathsf{E}[\|X(0)\|^2] < \infty$. *Under* (A1), (A2), *and* (TS), *with probability one,*

(i) $\|\phi(t) - \widehat{\phi}(t)\| \to 0$ *as* $t \to \infty$,

(ii) $\sup_{t\geq 0}\|\phi(t)\| < \infty$.

Proof. Express $\widehat{\phi}(\,\cdot\,)$ as follows: For $m(j) \leq n < m(j+1)$,

$$\begin{aligned}\widehat{\phi}(t(n+1)-) &= \widehat{\phi}(T(j)) + \sum_{i=m(j)}^{n}\int_{t(i)}^{t(i+1)} h_{r(j)}\big(\widehat{\phi}(s)\big)ds\\ &= \widehat{\phi}(T(j)) + \epsilon_1(j) + \sum_{i=m(j)}^{n} a(i)h_{r(j)}\big(\widehat{\phi}(t(i))\big),\end{aligned} \tag{4.3}$$

where $\epsilon_1(j) = O(\sum_{i=m(j)}^{m(j+1)} a(i)^2) \to 0$ as $j \to \infty$. The "$-$" covers the case where $t(n+1) = t(m(j+1)) = T(j+1)$.

We also have by definition

$$\phi\big(t(n+1)-\big) = \phi\big(T(j)\big) + \sum_{i=m(j)}^{n} a(i)\big[h_{r(j)}\big(\phi\big(t(i)\big)\big) + \widetilde{M}(i+1)\big]. \tag{4.4}$$

For $m(j) \leq n \leq m(j+1)$, let $\varepsilon(n) = \|\phi(t(n)-) - \widehat{\phi}(t(n)-)\|$. Combining (4.3), (4.4), and the Lipschitz continuity of h, we have

$$\varepsilon(n+1) \leq \varepsilon\big(m(j)\big) + \epsilon_1(j) + \|\xi(n+1) - \xi(m(j))\| + C\sum_{i=m(j)}^{n} a(i)\varepsilon(i),$$

where $C < \infty$ is a suitable constant. Since $\varepsilon(m(j)) = 0$, we can use Lemma 4.3 (i) to obtain

$$\varepsilon(n) \leq \exp\big(C(T+1)\big)\big(\epsilon_1(j) + \epsilon_2(j)\big), \qquad m(j) \leq n \leq m(j+1),$$

where $\epsilon_2(j) = \max_{m(j)\leq n\leq m(j+1)} \|\xi(n+1) - \xi(m(j))\|$. By (iv) of Lemma 4.5 and the martingale convergence theorem [18, p. 62], $\{\xi(n)\}$ converges a.s.; thus $\epsilon_2(j) \to 0$ a.s., as $j \to \infty$. Since $\epsilon_1(j) \to 0$ as well,

$$\sup_{m(j)\leq n\leq m(j+1)} \big\|\phi(t(n)-) - \widehat{\phi}(t(n)-)\big\| = \sup_{m(j)\leq n\leq m(j+1)} \varepsilon(n) \to 0$$

as $j \to \infty$, which implies the first claim.

Result (ii) then follows from Lemma 4.2 and the triangle inequality. □

LEMMA 4.7. *Under* (A1), (A2), *and* (BS), *there exists a constant* $C_2 < \infty$ *such that for all* $j \geq 0$,

(i) $\sup_{j\geq 0, T(j)\leq t\leq T(j+1)} \mathsf{E}[\|\phi(t) - \widehat{\phi}(t)\|^2 \mid \mathcal{F}_{n(j)}] \leq C_2\overline{\alpha}$,

(ii) $\sup_{j\geq 0, T(j)\leq t\leq T(j+1)} \mathsf{E}[\|\phi(t)\|^2 \mid \mathcal{F}_{n(j)}] \leq C_2$.

Proof. Mimic the proof of Lemma 4.6 to obtain

$$\varepsilon(n+1) \leq \sum_{i=m(j)}^{n} Ca(i)\varepsilon(i) + \epsilon_0(j), \qquad m(j) \leq n < m(j+1),$$

where $\varepsilon(n) = \mathsf{E}[\|\phi(t(n)-) - \widehat{\phi}(t(n)-)\|^2 \mid \mathcal{F}_{m(j)}]^{1/2}$ for $m(j) \leq n \leq m(j+1)$, and the error term has the upper bound

$$|\epsilon_0(j)| = O(\overline{\alpha}),$$

where the bound is deterministic. By Lemma 4.3 (i) we obtain the bound,

$$\varepsilon(n) \leq \exp\big(C(T+1)\big)\epsilon_0(j), \qquad m(j) \leq n \leq m(j+1),$$

which proves (i). We, therefore, obtain (ii) using Lemma 4.2, (i), and the triangle inequality. □

Proof of Theorem 2.1. (i) By a simple conditioning argument, we may take $X(0)$ to be deterministic without any loss of generality. In particular, $\mathsf{E}[\|X(0)\|^2] < \infty$ trivially. By Lemma 4.6 (ii), it now suffices to prove that $\sup_n \|X(m(n))\| < \infty$ a.s. Fix a sample point outside the zero probability set where Lemma 4.6 fails. Pick $T > 0$ as above and $R > 0$ such that for every solution $x(\,\cdot\,)$ of the ODE (1.4) with $\|x(0)\| \leq 1$ and $r \geq R$, we have $\|x(t)\| \leq \frac{1}{4}$ for $t \in [T, T+1]$. This is possible by Lemma 4.4.

Hence by Lemma 4.6 (i) we can find an $j_0 \geq 1$ such that whenever $j \geq j_0$ and $\|X(m(j))\| \geq R$,

$$\frac{\|X(m(j+1))\|}{\|X(m(j))\|} = \phi\big(T(j+1) - \big) \leq \frac{1}{2}. \tag{4.5}$$

This implies that $\{X(m(j)) : j \geq 0\}$ is a.s. bounded, and the claim follows.

(ii) For $m(j) < n \leq m(j+1)$,

(4.6)

$$\begin{aligned}
\mathsf{E}\big[\big\|X(n)\big\|^2 \mid \mathcal{F}_{m(j)}\big]^{1/2} &= \mathsf{E}\big[\big\|\phi\big(t(n)-\big)\big\|^2 \mid \mathcal{F}_{m(j)}\big]^{1/2}\big(\big\|X(m(j))\big\| \vee 1\big) \\
&\leq \mathsf{E}\big[\big\|\phi\big(t(n)-\big) - \widehat{\phi}\big(t(n)-\big)\big\|^2 \mid \mathcal{F}_{m(j)}\big]^{1/2}\big(\big\|X\big(m(j)\big)\big\| \vee 1\big) \\
&\quad + \mathsf{E}\big[\big\|\widehat{\phi}(t(n)-)\big\|^2 \mid \mathcal{F}_{m(j)}\big]^{1/2}\big(\|X\big(m(j)\big)\| \vee 1\big).
\end{aligned}$$

Let $0 < \eta < \frac{1}{2}$, and let $\alpha^* = \eta/(2C_2)$, for C_2 as in Lemma 4.7. We then obtain for $\overline{\alpha} \leq \alpha^*$,

$$\begin{aligned}
\mathsf{E}\big[\big\|X(n)\big\|^2 \mid \mathcal{F}_{m(j)}\big]^{1/2} &\leq (\eta/2)\big(\big\|X\big(m(j)\big)\big\| \vee 1\big) \\
&\quad + \mathsf{E}\big[\big\|\widehat{\phi}\big(t(n)-\big)\big\|^2 \mid \mathcal{F}_{m(j)}\big]^{1/2}\big(\big\|X\big(m(j)\big)\big\| \vee 1\big).
\end{aligned} \tag{4.7}$$

Choose $R, T > 0$ such that for any solution $x(\cdot)$ of the ODE (1.4), $\|x(t)\| < \eta/2$ for $t \in [T, T+1]$, whenever $\|x(0)\| < 1$ and $r \geq R$. When $\|X(m(j))\| \geq R$, we then obtain

$$\mathsf{E}\big[\big\|X\big(m(j+1)\big)\big\|^2 \mid \mathcal{F}_{m(j)}\big]^{1/2} \leq \eta \big\|X\big(m(j)\big)\big\|, \tag{4.8}$$

while by Lemma 4.7 (ii) there exists a constant C such that the left-hand side (l.h.s.) of the inequality above is bounded by C a.s. when $\|X(m(j))\| \leq R$. Thus,

$$\mathsf{E}\big[\big\|X\big(m(j+1)\big)\big\|^2\big] \leq 2\eta^2 \mathsf{E}\big[\big\|X\big(m(j)\big)\big\|^2\big] + 2C^2.$$

This establishes boundedness of $\mathsf{E}[\|X(m(j+1))\|^2]$, and the proof then follows from (4.7) and Lemma 4.2. □

4.2. Convergence for (BS). LEMMA 4.8. *Suppose that* (A1), (A2), *and* (BS) *hold and that* $\overline{\alpha} \leq \alpha^*$. *Then for some constant* $C_3 < \infty$,

$$\sup_{t \geq 0} \mathsf{E}\big[\big\|\widehat{\psi}(t) - \psi(t)\big\|^2\big] \leq C_3 \overline{\alpha}.$$

Proof. By (A2) and Theorem 2.1 (ii),

$$\sup_n \mathsf{E}\big[\big\|X(n)\big\|^2\big] < \infty, \quad \sup_n \mathsf{E}\big[\big\|M(n)\big\|^2\big] < \infty.$$

The claim then follows from familiar arguments using the Bellman Gronwall lemma exactly as in the proof of Lemma 4.6. □

Proof of Theorem 2.3. (i) We apply Theorem 2.1 which allows us to choose an $R > 0$ such that

$$\sup_n \mathsf{P}\big(\|X(n)\| > R\big) < \overline{\alpha}.$$

Let $B(c)$ denote the ball centered at x^* of radius $c > 0$ and let $0 < \mu < \epsilon/2$ be such that if a solution $x(\cdot)$ of (1.2) satisfies $x(0) \in B(\mu)$, then $x(t) \in B(\epsilon/2)$ for $t \geq 0$. Pick $T > 0$ such that if a solution $x(\cdot)$ of (1.2) satisfies $\|x(0)\| \leq R$, then $x(t) \in B(\mu/2)$ for $t \in [T, T+1]$. Then for all $j \geq 0$,

$$\begin{aligned}
\mathsf{P}\big(e\big(m(j+1)\big) \geq \mu\big) &= \mathsf{P}\Big(e\big(m(j+1)\big) \geq \mu, \big\|X\big(m(j)\big)\big\| > R\Big) \\
&\quad + \mathsf{P}\Big(e\big(m(j+1)\big) \geq \mu, \|X(m(j))\| \leq R\Big) \\
&\leq \overline{\alpha} + \mathsf{P}\Big(\psi\big(T(j+1)\big) \notin B(\mu), \widehat{\psi}\big(T(j+1)\big) \in B(\mu/2)\Big) \\
&\leq \overline{\alpha} + \mathsf{P}\Big(\|\psi\big(T(j+1)\big) - \widehat{\psi}\big(T(j+1)\big)\| > \mu/2\Big) \\
&\leq O(\overline{\alpha})
\end{aligned}$$

by Lemma 4.8. Then for $m(j) \leq n < m(j+1)$,

$$\begin{aligned}
\mathsf{P}\big(e(n) \geq \epsilon\big) &= \mathsf{P}\big(e(n) \geq \epsilon, e\big(m(j)\big) \geq \mu\big) \\
&\quad + \mathsf{P}\big(e(n) \geq \epsilon, e\big(m(j)\big) \leq \mu\big) \\
&\leq O(\overline{\alpha}) + \mathsf{P}\big(\psi\big(t(n)\big) \notin B(\epsilon), \widehat{\psi}\big(t(n)\big) \in B\big(\epsilon/2\big)\big) \\
&\leq O(\overline{\alpha}) + \mathsf{P}\big(\|\psi\big(t(n)\big) - \widehat{\psi}\big(t(n)\big)\| > \epsilon/2\big) \\
&\leq O(\overline{\alpha}).
\end{aligned}$$

Since the bound on the r.h.s. is uniform in n, the claim follows.

(ii) We first establish the bound with $n = m(j+1)$, $j \to \infty$. We have for any j,

$$\begin{aligned}\mathsf{E}\big[e\big(m(j+1)\big)^2\big]^{1/2} &\le \mathsf{E}\big[\big\|\psi\big(T(j+1)-\,\big) - \widehat{\psi}\big(T(j+1)-\,\big)\big\|^2\big]^{1/2} \\ &\quad + \mathsf{E}\big[\big\|\widehat{\psi}\big(T(j+1)-\,\big) - x^*\big\|^2\big]^{1/2}. \end{aligned} \tag{4.9}$$

By exponential stability there exist $C < \infty$, $\delta > 0$ such that for all $j \ge 0$,

$$\begin{aligned}\big\|\widehat{\psi}\big(T(j+1)-\,\big) - x^*\big\| &\le C\exp\big(-\delta\big[T(j+1)-T(j)\big]\big)\big\|\widehat{\psi}\big(T(j)\big) - x^*\big\| \\ &\le C\exp(-\delta T)\big\|\widehat{\psi}\big(T(j)\big) - x^*\big\|.\end{aligned}$$

Choose T so large that $C\exp(-\delta T) \le \frac{1}{2}$ so that

$$\begin{aligned}\mathsf{E}\big[\big\|\widehat{\psi}\big(T(j+1)-\,\big) - x^*\big\|^2\big]^{1/2} &\le \frac{1}{2}\mathsf{E}\big[\big\|\widehat{\psi}\big(T(j)\big) - x^*\big\|^2\big]^{1/2} \\ &\le \frac{1}{2}\mathsf{E}\big[e\big(m(j)\big)^2\big]^{1/2} + \frac{1}{2}\mathsf{E}\big[\big\|\psi\big(T(j)\big) - \widehat{\psi}\big(T(j)\big)\big\|^2\big]^{1/2}. \end{aligned} \tag{4.10}$$

Combining (4.9) and (4.10) with Lemma 4.8 gives

$$\mathsf{E}\big[e\big(m(j+1)\big)^2\big]^{1/2} \le \frac{1}{2}\mathsf{E}\big[e\big(m(j)\big)^2\big]^{1/2} + 2\sqrt{C_3\overline{\alpha}},$$

which shows that

$$\limsup_{j\to\infty} \mathsf{E}\big[e\big(m(j)\big)^2\big] \le 16C_3\overline{\alpha}.$$

The result follows from this and Lemma 4.7 (ii). □

Proof of Theorem 2.5. The details of the proof, though pedestrian in the light of the foregoing and [6], are quite lengthy, not to mention the considerable overhead of additional notation, and are therefore omitted. We briefly sketch below a single point of departure in the proof.

In Lemma 4.6 we compare two functions $\phi(\,\cdot\,)$ and $\widehat{\phi}(\,\cdot\,)$ on the interval $[T(j), T(j+1)]$. The former in turn involved the iterates $\widetilde{X}(n)$ for $m(j) \le n < m(j+1)$ or, equivalently, $X(n)$ for $m(j) \le n < m(j+1)$. Here $X(n+1)$ was computed in terms of $X(n)$ and the "noise" $M(n+1)$. In the asynchronous case, however, the evaluation of $X_j(n+1)$ can involve $X_j(n)$ for $n - \overline{\tau} \le m \le n$, $j \ne i$. Therefore the argument leading to Lemma 4.6 calls for a slight modification. While computing $X(n), m(j) \le n < m(j+1)$, we plug into the iteration as and when required $\widetilde{X}_i(m) = X_i(m)/r(j)$. Note, however, that if the same $X_i(m)$ also features in the computation of $X_k(l)$ for $m(q) \le \ell < m(q+1)$, say, with $q \ne j$, then $\widetilde{X}_i(m)$ should be redefined there as $X_i(m)/r(q)$. Thus the definition of $\widetilde{X}_i(m)$ now becomes context-dependent.

With this minor change, the proofs of [6] can be easily combined with the arguments used in the proofs of Theorems 2.1 and 2.2 to draw the desired conclusions. □

4.3. The Markov model. The bounds that we obtain for the Markov model (2.4) are based upon the theory of ψ-irreducible Markov chains.

A subset $S \subset \mathbb{R}^d$ is called *petite* if there exists a probability measure ν on $\mathbb{R}^d$ and $\delta > 0$ such that the resolvent kernel K satisfies

$$K(x, A) := \sum_{k=0}^{\infty} 2^{-k-1} P^k(x, A) \ge \delta\nu(A), \quad x \in S,$$

for any measurable $A \subset \mathbb{R}^d$. Under assumptions (2.6) and (2.7) we show below that every compact subset of $\mathbb{R}^d$ is petite, so that $\mathbf{\Phi}$ is a ψ-irreducible T-*chain*. We refer the reader to [16] for further terminology and notation.

LEMMA 4.9. *Suppose that* (A1), (A2), (2.6), *and* (2.7) *hold and that* $\alpha \le \alpha^*$. *Then all compact subsets of* $\mathbb{R}^d$ *are petite for the Markov chain* $\boldsymbol{X}$, *and hence the chain is* ψ*-irreducible.*

Proof. The conclusions of the theorem will be satisfied if we can find a function s which is bounded from below on compact sets and a probability ν such that the resolvent kernel K satisfies the bound

$$K(x, A) \ge s(x)\nu(A)$$

for every $x \in \mathbb{R}^d$ and any measurable subset $A \subset \mathbb{R}^d$. This bound is written succinctly as $K \ge s \otimes \nu$.

The first step of the proof is to apply the implicit function theorem together with (2.6) and (2.7) to obtain a bound of the form

$$P^d(x, A) = \mathsf{P}(X(d) \in A \mid X(0) = x) \ge \epsilon\nu(A), \quad x \in O,$$

where O is an open set containing $x^*, \epsilon > 0$, and ν is the uniform distribution on O. The set O can be chosen independent of α, but the constant ϵ may depend on α. For details on this construction, see Chapter 7 of [16].

To complete the proof it is enough to show that $K(x, O) > 0$. To see this, suppose that $\alpha \le \alpha^*$ and that $W(n) = w^*$ for all n. Then the foregoing stability analysis shows that $X(n) \in O$ for all n sufficiently large. Since w^* is in the support of the marginal distribution of $\{W(n)\}$, it then follows that $K(x, O) > 0$.

From these two bounds, we then have

$$K(x, A) \ge 2^{-d} \int K(x, dy) P^d(y, A) \ge 2^{-d} \epsilon K(x, O)\nu(A).$$

This is of the form $K \ge s \otimes \nu$ with s lower semicontinuous and positive everywhere. The function s is therefore bounded from below on compact sets, which proves the claim. ∎

The previous lemma together with Theorem 2.1 allows us to establish a strong form of ergodicity for the model.

LEMMA 4.10. *Suppose that* (A1), (A2), (2.6), *and* (2.7) *hold and that* $\alpha \le \alpha^*$.

(i) *There exists a function* $V_\alpha : \mathbb{R}^d \to [1, \infty)$ *and constants* $b, L < \infty$ *and* $\epsilon_0 > 0$ *independent of* α *such that*

$$PV_\alpha(x) \le \exp(-\epsilon_0\alpha)V_\alpha(x) + bI_C(x),$$

where $C = \{x : \|x\| \le L\}$. *While the function* V_α *will depend upon* α, *it is uniformly bounded as follows,*

$$\gamma^{-1}(\|x\|^2 + 1) \le V_\alpha(x) \le \gamma(\|x\|^2 + 1),$$

where $\gamma \ge 1$ *does not depend upon* α.

(ii) *The chain is* V*-uniformly ergodic, with* $V(x) = \|x\|^2 + 1$.

Proof. Using (4.8) we may construct T and L independent of $\alpha \le \alpha^*$ such that

$$\mathsf{E}\big[\big\|X(k_0)\big\|^2 + 1 \mid X(0) = x\big] \le (1/2)(\|x\|^2 + 1), \quad \|x\| \ge L,$$

where $k_0 = [T/\alpha] + 1$. We now set

$$V_\alpha(x) = \alpha \sum_{k=0}^{k_0-1} \mathsf{E}\left[\left\|X(k)\right\|^2 + 1 \mid X(0) = x\right] 2^{k/k_0}.$$

From the previous bound, it follows directly that the desired drift inequality holds with $\epsilon_0 = \log(2)/T$. Lipschitz continuity of the model gives the bounds on V_α. This proves (i).

The V-uniform ergodicity then follows from Lemma 4.9 and Theorem 16.0.1 of [16]. □

We note that for small α and large x, the Lyapunov function V_α approximates V_∞ plus a constant, where

$$V_\infty(x) = \int_0^T \left(\|x(s)\|^2 + 1\right) 2^{s/T} ds; \quad x(0) = x,$$

and $x(\,\cdot\,)$ is a solution to (1.5). If this ODE is asymptotically stable then the function V_∞ is in fact a Lyapunov function for (1.5), provided $T > 0$ is chosen sufficiently large.

In [17] a bound is obtained on the rate of convergence ρ given in (2.5) for a chain satisfying the drift condition

$$PV_\alpha(x) \le \lambda V(x) + bI_C(x).$$

The bound depends on the "petiteness" of the set C and the constants $b < \infty$ and $\lambda < 1$. The bound on ρ obtained in [17] also tends to unity with vanishing α since in the preceding lemma we have $\lambda = \exp(-\epsilon_0\alpha) \to 1$ as $\alpha \to 0$. From the structure of the algorithm this is not surprising, but this underlines the fact that care must be taken in the choice of the stepsize α.

REFERENCES

[1] J. Abounadi, D. Bertsekas, and V. S. Borkar, *Learning algorithms for Markov decision processes with average cost*, SIAM J. Control Optim., submitted.

[2] A. G. Barto, R. S. Sutton, and C. W. Anderson, *Neuron-like elements that can solve difficult learning control problems*, IEEE Trans. Systems, Man and Cybernetics, 13 (1983), pp. 835–846.

[3] A. Benveniste, M. Métivier, and P. Priouret, *Adaptive Algorithms and Stochastic Approximations*, Springer-Verlag, Berlin, 1990.

[4] D. Bertsekas and J. Tsitsiklis, *Neuro-Dynamic Programming*, Athena Scientific, Belmont, MA, 1996.

[5] V. S. Borkar, *Stochastic approximation with two time scales*, Systems Control Lett., 29 (1997), pp. 291–294.

[6] V. S. Borkar, *Asynchronous stochastic approximation*, SIAM J. Control Optim., 36 (1998), pp. 840–851.

[7] V. S. Borkar, *Recursive self-tuning control of finite Markov chains*, Appl. Math., 24 (1996), pp. 169–188.

[8] V. S. Borkar and K. Soumyanath, *An analog scheme for fixed-point computation, part* I*: Theory*, IEEE Trans. Circuits Systems I Fund. Theory Appl., 44 (1997), pp. 351–355.

[9] J. G. Dai, *On positive Harris recurrence for multiclass queueing networks: A unified approach via fluid limit models*, Ann. Appl. Probab., 5 (1995), pp. 49–77.

[10] J. G. Dai and S. P. Meyn, *Stability and convergence of moments for multiclass queueing networks via fluid limit models*, IEEE Trans. Automat. Control, 40 (1995), pp. 1889–1904.

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

[12] T. Jaakola, M. I. Jordan, and S. P. Singh, *On the convergence of stochastic iterative dynamic programming algorithms*, Neural Computation, 6 (1994), pp. 1185–1201.
[13] V. R. Konda and V. S. Borkar, *Actor-critic–type learning algorithms for Markov decision processes*, SIAM J. Control Optim., 38 (1999), pp. 94–123.
[14] H. J. Kushner and G. G. Yin, *Stochastic Approximation Algorithms and Applications*, Springer-Verlag, New York, 1997.
[15] V. A. Malyshev and M. V. Men'sikov, *Ergodicity, continuity and analyticity of countable Markov chains*, Trans. Moscow Math. Soc., 1 (1982), pp. 1–48.
[16] S. P. Meyn and R. L. Tweedie, *Markov Chains and Stochastic Stability*, Springer-Verlag, London, 1993.
[17] S. P. Meyn and R. L. Tweedie, *Computable bounds for geometric convergence rates of Markov chains*, Ann. Appl. Probab., 4 (1994), pp. 981–1011.
[18] J. Neveu, *Discrete Parameter Martingales*, North Holland, Amsterdam, 1975.
[19] T. Sargent, *Bounded Rationality in Macroeconomics*, Clarendon Press, Oxford, 1993.
[20] J. Tsitsiklis, *Asynchronous stochastic approximation and q-learning*, Mach. Learning, 16 (1994), pp. 195–202.
[21] C. J. C. H. Watkins and P. Dayan, *Q-learning*, Mach. Learning, 8 (1992), pp. 279–292.

SIAM J. CONTROL OPTIM.
Vol. 38, No. 2, pp. 470–481

UNIQUENESS OF LOWER SEMICONTINUOUS VISCOSITY SOLUTIONS FOR THE MINIMUM TIME PROBLEM*

OLIVIER ALVAREZ†, SHIGEAKI KOIKE‡, AND ISAO NAKAYAMA‡

Abstract. We obtain the uniqueness of lower semicontinuous (LSC) viscosity solutions of the transformed minimum time problem assuming that they converge to zero on a "reachable" part of the target in appropriate directions. We present a counter-example which shows that the uniqueness does not hold without this convergence assumption.

It was shown by Soravia that the uniqueness of LSC viscosity solutions having a "subsolution property" on the target holds. In order to verify this subsolution property, we show that the dynamic programming principle (DPP) holds inside for any LSC viscosity solutions.

In order to obtain the DPP, we prepare appropriate approximate PDEs derived through Barles' inf-convolution and its variant.

Key words. semicontinuous viscosity solutions, dynamic programming principle, minimum time problem

AMS subject classifications. 49L25, 49L20, 35F30

PII. S0363012997317190

1. Introduction.

In this manuscript, we discuss the minimum time problem of deterministic optimal control, which has been studied via viscosity solution approach by many authors. As the first result, we refer to Bardi [Ba]. See also [EJ], [BF], [BS], and [BSo], who also treated the minimum time problem of differential games.

In those works, they characterized the value function of the minimum time problem to reach a given target as the unique viscosity solution of a first-order PDE. However, they treated the case only when the resulting value functions are continuous since those uniqueness results imply the continuity of solutions. We note that there often appear discontinuous value functions for practical minimum time problems.

The breakthrough to treat semicontinuous solutions for first-order PDEs was done by Barron and Jensen [BJ1]. Indeed, they introduced a new definition of semicontinuous viscosity solutions for Cauchy problems with convex Hamiltonians, which arise when we deal with optimal control problems. Under their setting, it was shown in [BJ2] that the semicontinuous value function is the unique solution of the associated PDE. We note that if we restrict ourselves to treat continuous viscosity solutions, then their definition is equivalent to that of the standard one.

Afterward, Barles [B1] discussed semicontinuous solutions for stationary problems utilizing "Barles"-convolution. With this idea, Soravia [S1] studied the Dirichlet-type problems. More precisely, he imposed a "subsolution" property on the boundary of the target, under which the uniqueness of LSC viscosity solutions for the (transformed) minimum time problem was obtained. See also [K] and [BL] for related topics. Recently, Cârjă, Mignanego, and Pieri in [CMP] (see also [C]) studied LSC viscosity

*Received by the editors February 21, 1997; accepted for publication (in revised form) June 8, 1999; published electronically January 19, 2000.

http://www.siam.org/journals/sicon/38-2/31719.html

†UPRESA 60-85, Site Colbert, Université de Rouen, 76821 Mont-Saint-Aignan Cedex, France (alvarez@univ-rouen.fr).

‡Department of Mathematics, Saitama University, 255 Shimo-Okubo, Urawa, Saitama 338-8570 Japan (skoike@rimath.saitama-u.ac.jp). The second author was supported by Grant-in-Aid for Scientific Research 09640242 and 09440067, the Ministry of Education, Science and Culture in Japan.

solutions of the minimum time problem assuming that they converge to the Dirichlet data from inside.

For the viscosity solution theory of first-order Hamilton–Jacobi equations, we refer to a new book by Bardi and Capuzzo Dolcetta [BC].

On the other hand, in nonsmooth analysis, LSC solutions have been studied in optimal control theory. For the first result, we refer to Frankowska [F]. More recently, Wolenski and Zhuang [WZ] have proved the uniqueness of LSC solutions of the minimum time problem assuming the subsolution property on the target as in [S1], which the value function satisfies. We note that their definition of solutions is slightly different from that of viscosity solutions. It is worth mentioning that in [WZ], to show the uniqueness, they compared the other LSC solution (if it exists) with the value function by the so-called invariance theory while, in the literature of the viscosity solution theory, we have shown it via comparison principle for a boundary value problem of PDEs.

Our aim here is to obtain a uniqueness result without assuming the subsolution property on the target. In fact, we will derive such a property from the definition of solutions under some continuity assumption on a "reachable" part of the boundary. Then, we will be able to apply Soravia's argument in [S1] to get the uniqueness.

Moreover, we will mention that our continuity condition is equivalent to Soravia's one. In fact, to show that the Soravia's condition implies the continuity condition, we give a direct proof, although we can prove it using the uniqueness of solutions.

In an example, we will see that this continuity assumption is necessary to obtain the uniqueness result.

Here, we shall recall the original minimum time problem. Consider the state equation associated with controls

$$\alpha \in \mathcal{A} \equiv \{\alpha : [0,\infty) \to A \text{ measurable}\},$$

where A is a compact set in $\mathbf{R}^m$ (for some $m \in \mathbf{N}$). For $x \in \mathbf{R}^n$,

$$\begin{cases} \dfrac{dX}{dt}(t) = g(X(t), \alpha(t)) \quad \text{for } t > 0, \\ X(0) = x, \end{cases} \tag{1.1}$$

where $g : \mathbf{R}^n \times A \to \mathbf{R}^n$ is a given function and $x \in \mathbf{R}^n$ is fixed.

We shall denote, under appropriate hypotheses, by $X(\cdot; x, \alpha)$ the (unique) solution of (1.1). We will also denote by $X(\cdot; x, \xi(\cdot))$ the unique solution for a vector field $\xi \in W^{1,\infty}(\mathbf{R}^n; \mathbf{R}^n)$:

$$\begin{cases} \dfrac{dX}{dt}(t) = \xi(X(t)) \quad \text{for } t > 0, \\ X(0) = x. \end{cases}$$

For simplicity, we shall suppose that

$$\mathcal{T} \subset \mathbf{R}^n \text{ is compact.} \tag{$A0$}$$

With these notations, we recall the value function of the minimum time problem:

$$V(x) = \inf_{\alpha \in \mathcal{A}} T_x^\alpha,$$

where $T_x^\alpha = \inf\{t \geq 0 \mid X(t; x, \alpha) \in \mathcal{T}\}$.

Since $V(x)$ might be infinity in a subregion of $\Omega \equiv \mathbf{R}^n \backslash \mathcal{T}$, we will have to study the free boundary problem:

$$\max_{a\in A}\{-\langle g(x,a), DV(x)\rangle - 1\} = 0 \quad \text{in } \mathcal{R} \equiv \{x \in \Omega \mid V(x) < \infty\}. \tag{1.2}$$

Since we cannot expect that $\mathcal{R}$ is open in general, as we will see, we meet some difficulty if we treat (1.2) directly. Therefore, in this paper, following the previous works, we shall consider the transformed value function by Kruzkov transformation:

$$u(x) = \inf_{\alpha\in\mathcal{A}} \left(1 - e^{-T_x^\alpha}\right).$$

Then, we can expect u to be a solution of

$$u(x) + \max_{a\in A}\{-\langle g(x,a), Du(x)\rangle\} = 1 \quad \text{in } \Omega. \tag{1.3}$$

Thus, once we verify that u is the unique solution of (1.3), we will be able to derive the reachable set by $\mathcal{R} = \{x \in \Omega \mid u(x) < 1\}$.

This paper is organized as follows. Section 2 is devoted to our definition of the minimum time problem and the DPP which implies the subsolution property. We present our uniqueness result and examples in section 3. Also, we discuss the equivalence of boundary conditions in section 3. In the final section, we prove the DPP in section 2.

2. DPP. Our hypothesis on the regularity of given functions is as follows:

$$(A1) \qquad g \in C\big(\mathbf{R}^n \times A; \mathbf{R}^n\big) \qquad \text{and} \qquad \sup_{a\in A} \|g(\cdot,a)\|_{W^{1,\infty}(\mathbf{R}^n;\mathbf{R}^n)} < \infty.$$

For later convenience, we shall consider the following general first-order PDE in a set $\Sigma \subset \mathbf{R}^n$:

$$u(x) + \max_{a\in A}\{-\langle g(x,a), Du(x)\rangle - f(x,a)\} = 0 \quad \text{in } \Sigma, \tag{2.1}$$

where $f : \mathbf{R}^n \times A \to \mathbf{R}$ is a given continuous function.

For simplicity, we shall use the notation

$$H(x,r,p) \equiv r + \max_{a\in A}\{-\langle g(x,a), p\rangle - f(x,a)\}.$$

We will suppose the following regularity on given functions in (2.1):

$$(A1') \qquad \begin{cases} g \in C\big(\mathbf{R}^n \times A; \mathbf{R}^n\big),\ f \in C\big(\mathbf{R}^n \times A; \mathbf{R}\big), \text{ and} \\ \sup\limits_{a\in A} \big\{\|g(\cdot,a)\|_{W^{1,\infty}(\mathbf{R}^n;\mathbf{R}^n)} + \|f(\cdot,a)\|_{W^{1,\infty}(\mathbf{R}^n;\mathbf{R})}\big\} < \infty. \end{cases}$$

Following [BJ1] (also [B1]), we present our definition of solutions of (2.1).

DEFINITION. *For a function $u : \Sigma \to \mathbf{R}$, we call it a subsolution (resp., supersolution) of* (2.1) *if u is LSC in Σ, and*

$$H(x,u(x),p) \le (\textit{resp.},\ \ge)\ 0 \quad \textit{for } x \in \Sigma \textit{ and } p \in D^-u(x),$$

where $D^-u(x)$ denotes the standard subdifferential of u at $x \in \Sigma$.

$$D^-u(x) = \{p \in \mathbf{R}^n \mid u(y) \ge u(x) + \langle p, y-x\rangle + o(|y-x|) \textit{ as } y \to x\}.$$

For a function $u : \Sigma \to \mathbf{R}$, we also call it a solution of (2.1) *if u is both a sub- and a supersolution of* (2.1)

We characterize the set of reachable controls in the following way. For $x \in \partial\mathcal{T}$,

$$A(x) \equiv \left\{ a \in A \;\middle|\; \begin{array}{c} \text{There exists } t > 0 \text{ such that} \\ X(s; x, -g(\cdot, a)) \in \Omega \text{ for } s \in (0, t) \end{array} \right\}.$$

We shall derive the "subsolution" properties on $\partial\mathcal{T}$ for solutions through the following propositions.

LEMMA 2.1. *Assume that* (A1) *holds. Let u be a solution of* (1.3). *Assume also that $u = 0$ on $\mathcal{T}$. Then, for $x \in \partial\mathcal{T}$ and $p \in D^- u(x)$, we have*

$$-\langle g(x, a), p\rangle \le 0 \quad \textit{provided } a \in A \backslash A(x).$$

Proof. Choose $\phi \in C^1$ such that $u - \phi$ attains its minimum over $\mathbf{R}^n$ at $x \in \partial\mathcal{T}$, $u(x) = \phi(x) = 0$, and $D\phi(x) = p$. Set $X(\cdot) \equiv X(\cdot; x, -g(\cdot, a))$.

Since $a \in A\backslash A(x)$, there exists $\{t_k > 0\}_{k=1}^{\infty}$, such that $\lim_{k\to\infty} t_k = 0$ and $X(t_k) \in \mathcal{T}$ $(k = 1, 2, \dots)$. Hence,

$$\phi(X(t_k)) - \phi(x) \le u(X(t_k)) = 0.$$

Therefore, dividing t_k and then sending $k \to \infty$, we conclude the assertion. □

For simplicity, we shall suppose that

(A0′) $\qquad \Sigma$ is open and $\partial\Sigma$ is compact.

For $\delta > 0$, we define an open subset

$$\Sigma_\delta \equiv \{x \in \Sigma \mid \operatorname{dist}(x, \partial\Sigma) > \delta\}.$$

Also, for an open subset $O \subset \Sigma$ and $x \in O$, we use the notation

$$\tau_O^{x,\alpha} = \inf\{t \ge 0 \mid X(t; x, \alpha) \notin O\}.$$

We present the DPP for (2.1), whose proof will be given in the final section since it is rather complicated.

THEOREM 2.2 (cf. [L]). *Assume that* (A0′) *and* (A1′) *hold. Let $u : \overline{\Sigma} \to \mathbf{R}$ be a bounded solution of*

$$H(x, u, Du) = 0 \quad \textit{in } \Sigma.$$

Then, for $\delta > 0$ and $x \in \Sigma_\delta$,

$$u(x) = \inf_{\alpha\in\mathcal{A}} \left\{ \int_0^{\tau_{\Sigma_\delta}^{x,\alpha}} e^{-s} f(X(s; x, \alpha), \alpha(s)) ds + e^{-\tau_{\Sigma_\delta}^{x,\alpha}} u(X(\tau_\delta^{x,\alpha}; x, \alpha)) \right\}.$$

COROLLARY 2.3. *Assume that* (A0) *and* (A1) *hold. Fix $x \in \partial\mathcal{T}$. Let $u : \overline{\Omega} \to \mathbf{R}$ be a bounded solution of*

$$u(x) + \max_{a\in A}\{-\langle g(x, a), Du(x)\rangle\} - 1 = 0 \quad \textit{in } \Omega.$$

Assume also that $u = 0$ on $\mathcal{T}$ and, for any $x \in \partial\mathcal{T}$ and $a \in A(x)$,

$$\liminf_{s\to t} u(X(s; x_t, a)) = u(x) \textit{ holds,}$$

where $x_t \equiv X(t; x, -g(\cdot, a))$ *for* $t \geq 0$.

Then,

$$u(x) - \langle g(x,a), p\rangle \leq 1 \quad \textit{for } p \in D^- u(x).$$

Proof of Corollary 2.3. Let $x \in \partial\mathcal{T}, p \in D^- u(x)$, and $a \in A(x)$ in the hypothesis. We then set $x_t = X(t; x, -g(\cdot, a)) \in \Omega$ for small $t > 0$.

Choose $\phi \in C^1$ such that $u(x) = \phi(x)$, $u \geq \phi$ in $\mathbf{R}^n$, and $D\phi(x) = p$.

Fix small $t > 0$ and choose $\delta(t) > 0$ such that $x_t \in \Omega_\delta$ for $\delta \in (0, \delta(t))$.

By Theorem 2.2, we have

$$u(x_t) \leq 1 - e^{-\tau_{\Omega_\delta}^{x_t,a}} + e^{-\tau_{\Omega_\delta}^{x_t,a}} u(X(\tau_{\Omega_\delta}^{x_t,a}; x_t, a)),$$

where a stands for the constant control; $\alpha(\cdot) \equiv a$.

We note that $\lim_{\delta\to 0} \tau_{\Omega_\delta}^{x_t,a} = t$.

We also note that the uniqueness of solutions of (1.1) yields $X(\tau_{\Omega_\delta}^{x_t,a}; x_t, a) = x_{t-\tau_{\Omega_\delta}^{x_t,a}}$. Take the limit infimum, as $\delta \to 0$, together with these in the above to get

$$\phi(x_t) - e^{-t}\phi(x) \leq u(x_t) - e^{-t}u(x) \leq 1 - e^{-t}.$$

Dividing $t > 0$ and then, sending $t \to 0$ in the above, we conclude the proof. □

3. Main results. In order to obtain the uniqueness result, we will suppose the following continuity assumption. Letting u be an LSC function in $\overline{\Omega}$, we will suppose that, for any $x \in \mathcal{T}$ and $a \in A(x)$,

$$\liminf_{s\to t} u(X(s; x_t, a)) = 0 \quad \text{for small } t > 0, \tag{A2}$$

where $x_t = X(t; x, -g(\cdot, a))$.

Notice that we do not suppose that $A(x) \neq \emptyset$ in this hypothesis.

Our uniqueness result for (1.3) is as follows.

THEOREM 3.1. *Assume that* (A0) *and* (A1) *hold. Let* u *and* $v : \mathbf{R}^n \to \mathbf{R}$ *be bounded solutions of* (1.3) *and satisfy* (A2). *Assume also that* $u = v = 0$ *on* $\mathcal{T}$. *Then,* $u = v$ *in* $\mathbf{R}^n$.

Proof of Theorem 3.1. In view of Lemma 2.1 and Corollary 2.3, we see that

$$u(x) + \max_{a\in A}\{-\langle g(x,a), p\rangle\} \leq 1 \quad \text{for } x \in \partial\mathcal{T} \text{ and } p \in D^- u(x). \tag{3.1}$$

This property enables us to apply Soravia's result, Theorem 3.1 in [S1], to conclude the proof. □

Thanks to the above theorem, it is easy to show that the relaxed value function is the unique bounded viscosity solution of (1.3) satisfying (A2). To this end, let us introduce the unique solution $\hat{X}(\cdot; x, \mu)$ of the associated state equation:

$$\hat{X}(t) = x + \int_0^t \int_A g\big(\hat{X}(s), a\big) d\mu[s](a) ds, \tag{3.2}$$

where $s \in [0, \infty) \to \mu[s] \in M(A)$ is measurable. Here, $M(A)$ is the set of all Radon probability measures on A. We shall denote by $\hat{\mathcal{A}}$ the set of such maps μ.

The relaxed value function is as follows:

$$\hat{V}(x) = \inf_{\mu\in\hat{\mathcal{A}}} \left(1 - e^{-\hat{T}_x^\mu}\right),$$

where $\hat{T}_x^\mu = \inf\{t \geq 0 \mid \hat{X}(t;x,\mu) \in \mathcal{T}\}$.

THEOREM 3.2. *Assume that* (A0) *and* (A1) *hold. Then,* $\hat{V}$ *is the unique bounded solution of* (1.3) *satisfying* (A2).

Proof of Theorem 3.2. Following the argument in [BJ2], we see that $\hat{V}$ is LSC and satisfies (1.3) in our sense in Ω.

To check that A2 holds for $\hat{V}$, we observe that, for $x \in \partial\mathcal{T}$ and $a \in A(x)$,

$$\hat{V}(X(s;X(t;x,-g(\cdot,a)),a)) \leq 1 - e^{-(t-s)}.$$

Hence, since $\hat{V}(x) = 0$, sending $s \to t$, we obtain (A2) for $\hat{V}$. We remark here that the nonnegativity of $\hat{V}$ indeed yields

$$\lim_{s\to t} \hat{V}(X(s;X(t;x,-g(\cdot,a)),a)) = 0.$$

Therefore, Theorem 3.1 immediately implies the assertion. □

Remark. If we have an LSC solution u of (1.3) satisfying (3.1), then $u(x) = \hat{V}(x)$ in $\mathbf{R}^n$. Hence, (A2) holds true for u by Soravia's argument in [S1] since $\hat{V}$ also satisfies (3.1). Thus, through the above theorem, the condition (3.1) is equivalent to (A2). See [WZ] for the same argument.

Now, we shall show that condition (3.1) implies a bit stronger assertion than (A2).

THEOREM 3.3. *Assume that* (A0) *and* (A1) *hold. Let* $u : \overline{\Omega} \to [0,\infty)$ *be a bounded subsolution of* (1.3), *satisfying* (3.1) *and* $u = 0$ *on* $\mathcal{T}$. *Then, for each* $x \in \partial\mathcal{T}$, $a \in A(x)$, *and small* $t > 0$, *we have*

$$\lim_{s\to t} u(X(s;X(t;x,-g(\cdot,a)),a)) = 0.$$

Proof of Theorem 3.3. Fix $x \in \partial\mathcal{T}$ and $a \in A(x)$. As usual, we may suppose $x = 0$ and $g(0,a) = -e_n$, where $e_n = (0,\dots,0,1)$. Furthermore, we may suppose that $g(\cdot,a) = -e_n$ near the origin. Indeed, setting $v(y) = v(y_1,\dots,y_n) = u(X(y_n;(y_1,\dots,y_{n-1},0),-g(\cdot,a)))$, we have

$$v(y) + \frac{\partial v}{\partial y_n}(y) = u(X) - \langle g(X,a), Du(X)\rangle.$$

Define $Q_\eta^h = \{x = (x',x_n) \mid -1 < x_n < h, |x'| < \eta\}$ for small $h,\eta \in (0,1)$, and $\phi(x',x_n) = 2(x_n - |x'|^2/\eta^2)$. Since $\min_{\overline{Q_\eta^h}}(u-\phi) \leq (u-\phi)(0) = 0$, the minimum point $\hat{x} \in \overline{Q_\eta^h}$ can be attained at $\hat{x} = (\hat{x}',h)$. Indeed, otherwise, we have four possibilities:

(1) In the case when $\hat{x}_n = -1$, we immediately see $(u-\phi)(\hat{x}) \geq 2 > (u-\phi)(0)$.

(2) In the case when $|\hat{x}'| = \eta$ holds, we also have $(u-\phi)(\hat{x}) > 0 = (u-\phi)(0)$.

(3) In the case when $\hat{x} = (\hat{x}',\hat{x}_n) \in Q_\eta^h \backslash \overline{\Omega}$, there is $\epsilon > 0$ such that $(u-\phi)(\hat{x} + \epsilon e_n) < (u-\phi)(\hat{x})$.

In the above three cases, we get a contradiction to the choice of the minimum point $\hat{x}$. The remaining case is as follows:

(4) In the case when $\hat{x} \in \overline{\Omega} \cap Q_\eta^h$, the definition of solutions yields

$$1 \geq u(\hat{x}) + \frac{\partial \phi}{\partial x_n}(\hat{x}) \geq 2,$$

which is a contradiction.

Therefore, taking $\eta \to 0$ along a subsequence if necessary, by the lower semicontinuity of u, we find $u(0,\dots,0,h) \leq 2h$, which concludes the assertion. □

Remark. Recently, P. Soravia kindly let us know that we can easily obtain the above assertion using the optimality principle in [S2], [S3].

The following example is due to Soravia.

EXAMPLE 3.4. *For $\mathcal{T} \equiv [-1, 1]$, consider the PDE*

$$u + \left|\frac{\partial u}{\partial x}\right| = 1 \quad \text{in } \Omega \equiv \mathbf{R}\backslash\mathcal{T}. \tag{3.3}$$

It is easy to show that the unique (continuous) solution is given by

$$V(x) = \begin{cases} 1 - e^{-|x|+1} & \text{for } |x| \geq 1, \\ 0 & \text{for } |x| < 1. \end{cases}$$

On the other hand, we observe that the following function satisfies (3.3) *in* Ω:

$$\hat{V}(x) = \begin{cases} 1 - e^{x+1} & \text{for } x < -1, \\ 0 & \text{for } |x| \leq 1, \\ 1 & \text{for } x > 1. \end{cases}$$

Notice that $\hat{V}$ does not satisfy $(A2)$ *at $x = 1$. Thus, this example indicates that it is necessary for the uniqueness result to suppose* $(A2)$.

We also note that, by Theorem 3.1, *$\hat{V}$ is the unique LSC solution of*

$$u - \frac{\partial u}{\partial x} = 1 \quad \text{in } \Omega.$$

We next give an example, in which the reachable set is not open and the discontinuity appears in Ω.

EXAMPLE 3.5. *For $\mathcal{T} \equiv \{x = (0, x_2) \in \mathbf{R}^2 \mid 0 \leq x_2 \leq 1\}$,*

$$u + \max\left\{\left|\frac{\partial u}{\partial x_1}\right|, -a(x_2)\frac{\partial u}{\partial x_2}\right\} = 1 \quad \text{in } \Omega \equiv \mathbf{R}^2\backslash\mathcal{T}, \tag{3.4}$$

where

$$a(x_2) = \begin{cases} 1 & \text{for } x_2 \geq -1, \\ x_2 + 2 & \text{for } x_2 \in (-2, -1), \\ 0 & \text{for } x_2 \leq -2. \end{cases}$$

We easily verify that the reachable set $\mathcal{R} = \{x \in \Omega \mid V(x) < 1\}$ is given by

$$\{(x_1, x_2) \in \Omega \mid -2 < x_2 \leq 1\}.$$

Moreover, it is not hard to calculate the value function:

$$V(x_1, x_2) = \begin{cases} 1 & \text{for } x_2 > 1 \text{ or } x_2 \leq -2, \\ 1 - e^{-|x_1|} & \text{for } x_2 \in (0, 1], \\ 1 - e^{-|x_1|+x_2} & \text{for } x_2 \in (-1, 0], \\ 1 - (x_2 + 2)e^{-|x_1|-1} & \text{for } x_2 \in (-2, -1]. \end{cases}$$

Notice that the discontinuity of V occurs at $(x_1, 1) \in \Omega$.

4. Proof of Theorem 2.2. The basic idea of our proof was obtained by Lions in [L] for second-order PDEs. We also refer to [EI] and [BSo]. But, in their argument, we need some regularity of solutions. Hence, we will adapt some approximation techniques.

Let u be a solution of (2.1). We shall extend u (with the same notation) to the whole space by setting $u(x) = \infty$ for $x \notin \overline{\Sigma}$.

We fix any $T > 0$.

We first approximate u by locally Lipschitz continuous functions. For $\epsilon > 0$ and $(x,t) \in \mathbf{R}^n \times (0,T)$, we define

$$u_\epsilon(x,t) = \inf_{y \in \mathbf{R}^n} \left(u(y) + e^{-\mu t} \frac{|x-y|^2}{\epsilon^2} \right)$$

and

$$u^\epsilon(x,t) = \inf_{y \in \mathbf{R}^n} \left(u(y) + e^{\mu t} \frac{|x-y|^2}{\epsilon^2} \right).$$

Here, we fix $\mu > 2 + 2\max_{a \in A} \|Dg(\cdot,a)\|_\infty$. Notice that the first one is Barles' convolution but the second one has an opposite sign of the power on the exponential.

It is immediate to see that $u_\epsilon \le u^\epsilon$ in $\mathbf{R}^n \times [0,T]$.

We can easily show the properties:

$$\text{(4.1)} \qquad \begin{cases} |x-y| \le \epsilon e^{\mu t/2} (2\|u\|_\infty)^{1/2} & \text{if } u_\epsilon(x,t) = u(y) + e^{-\mu t} \dfrac{|x-y|^2}{\epsilon^2}, \\ |x-y| \le \epsilon e^{-\mu t/2} (2\|u\|_\infty)^{1/2} & \text{if } u^\epsilon(x,t) = u(y) + e^{\mu t} \dfrac{|x-y|^2}{\epsilon^2}. \end{cases}$$

In view of these facts, we define the constant $\hat{c} = \hat{c}(T) \equiv e^{\mu T/2} (2\|u\|_\infty)^{1/2}$.

We claim that the following properties hold. For some C_1 and $C_1' > 0$ independent of ϵ and $\mu > 0$,

$$\text{(4.2)} \qquad \begin{cases} 0 \ge u_\epsilon(x,t) + q + \max\limits_{a \in A} \{ -\langle g(x,a), p \rangle - f(x,a) \} - C_1 \epsilon^2 e^{\mu t} \\ \qquad \text{for } (x,t) \in \Sigma_{\hat{c}\epsilon} \times (0,T) \text{ and } (p,q) \in D^+ u_\epsilon(x,t) \end{cases}$$

and

$$\text{(4.3)} \qquad \begin{cases} 0 \le u^\epsilon(x,t) + q + \max\limits_{a \in A} \{ -\langle g(x,a), p \rangle - f(x,a) \} + C_1' \epsilon^2 e^{-\mu t} \\ \qquad \text{for } (x,t) \in \Sigma_{\hat{c}\epsilon} \times (0,T) \text{ and } (p,q) \in D^- u^\epsilon(x,t). \end{cases}$$

Here, $D^+ u_\epsilon(x,t) = -D^-(-u_\epsilon)(x,t)$.

We note that (4.3) holds in a larger set than $\Sigma_{\hat{c}\epsilon}$ but this is sufficient to conclude the proof.

Although it is not hard to show (4.2) and (4.3) by the argument in [B1] together with (4.1), we give a brief proof for the reader's convenience.

Since (4.3) can be obtained easily by remarking the sign of the power on e, we shall only show (4.2). See also our proof for (4.2) below.

Let us recall the Barron–Jensen lemma, which will be needed also for checking the sign of q in (4.2).

LEMMA 4.1 (See [BJ1] or [K]). *Fix $(x,t) \in \mathbf{R}^n \times (0,T)$ and $(p,q) \in D^+ u_\epsilon(x,t)$. For any $\alpha > 0$, there exist $(x_k^\alpha, t_k^\alpha) \in \mathbf{R}^n \times (0,T)$, $(p_k^\alpha, q_k^\alpha) \in D^- u_\epsilon(x_k^\alpha, t_k^\alpha)$ for $k \in$*

$\{1, 2, \dots, n(\alpha)\}$ *(with some* $n(\alpha) \in \mathbf{N}$*),* $(x^\alpha, t^\alpha) \in \mathbf{R}^n \times (0, T)$*,* $C_0 > 0$*, and* $\{\theta_k^\alpha \in [0,1]\}_{k=1}^{n(\alpha)}$ *such that*

$$
\begin{cases}
\text{(i)} & \displaystyle\lim_{\alpha\to 0} \frac{|x_k^\alpha - x^\alpha|}{\alpha} = 0, \\
\text{(ii)} & \displaystyle\lim_{\alpha\to 0} \big(x^\alpha, t_k^\alpha\big) = (x,t) \quad \big(\forall k = 1, \dots, n(\alpha)\big), \\
\text{(iii)} & \alpha |p_k^\alpha| \le C_0 \quad \big(\forall \alpha > 0,\ k = 1, \dots, n(\alpha)\big), \\
\text{(iv)} & \displaystyle\sum_{k=1}^{n(\alpha)} \theta_k^\alpha = 1 \quad (\forall \alpha > 0), \\
\text{(v)} & \displaystyle\lim_{\alpha\to 0} \sum_{k=1}^{n(\alpha)} \theta_k^\alpha (p_k^\alpha, q_k^\alpha) = (p,q).
\end{cases} \tag{4.4}
$$

For $(x,t) \in \Sigma_{\hat{c}\epsilon} \times (0,T)$ and $(p,q) \in D^+ u_\epsilon(x,t)$ in (4.2), we shall choose (x_k^α, t_k^α), etc., in Lemma 4.1.

Since we may suppose $x_k^\alpha \in \Sigma_{\hat{c}\epsilon}$ for small $\alpha > 0$, in view of (4.1), we can choose $y_k^\alpha \in \Sigma$ such that

$$
u_\epsilon(x_k^\alpha, t_k^\alpha) = u(y_k^\alpha) + e^{-\mu t_k^\alpha} \frac{|x_k^\alpha - y_k^\alpha|^2}{\epsilon^2}. \tag{4.5}
$$

Since $p_k^\alpha \in D^- u(y_k^\alpha)$, the definition yields

$$
0 = u(y_k^\alpha) + \max_{a\in A}\{-\langle g(y_k^\alpha, a), p_k^\alpha\rangle - f(y_k^\alpha, a)\}.
$$

Noting $p_k^\alpha = e^{-\mu t_k^\alpha} \frac{2(x_k^\alpha - y_k^\alpha)}{\epsilon^2}$, we calculate in the following way:

$$
\begin{aligned}
0 &\ge u(y_k^\alpha) + \max_{a\in A}\{-\langle g(x_k^\alpha, a), p_k^\alpha\rangle - f(y_k^\alpha, a)\} \\
&\quad - 2 \max_{a\in A} \|Dg(\cdot, a)\|_\infty e^{-\mu t_k^\alpha} \frac{|x_k^\alpha - y_k^\alpha|^2}{\epsilon^2} \\
&\ge u(y_k^\alpha) + \max_{a\in A}\{-\langle g(x^\alpha, a), p_k^\alpha\rangle - f(x_k^\alpha, a)\} \\
&\quad - \max_{a\in A} \|Dg(\cdot, a)\|_\infty \left(2e^{-\mu t_k^\alpha} \frac{|x_k^\alpha - y_k^\alpha|^2}{\epsilon^2} + |x_k^\alpha - x^\alpha||p_k^\alpha|\right) \\
&\quad - \max_{a\in A} \|Df(\cdot, a)\|_\infty |x_k^\alpha - y_k^\alpha|.
\end{aligned}
$$

Since we may also suppose $\mu e^{-\mu t_k^\alpha} \frac{|x_k^\alpha - y_k^\alpha|^2}{\epsilon^2} + q_k^\alpha = 0$ for small $\alpha > 0$, by (4.5) and (iii) of (4.4), we can find $C_1 > 0$ such that

$$
\begin{aligned}
0 &\ge u_\epsilon(x_k^\alpha, t_k^\alpha) + q_k^\alpha + \max_{a\in A}\{-\langle g(x^\alpha, a), p_k^\alpha\rangle - f(x_k^\alpha, a)\} \\
&\quad - C_0 \max_{a\in A} \|Dg(\cdot, a)\|_\infty \frac{|x_k^\alpha - x^\alpha|}{\alpha} - C_1 \epsilon^2 e^{\mu t_k^\alpha} \\
&\quad + \left\{\mu - 2 \max_{a\in A} \|Dg(\cdot, a)\|_\infty - 2\right\} e^{-\mu t_k^\alpha} \frac{|x_k^\alpha - y_k^\alpha|^2}{\epsilon^2}.
\end{aligned}
$$

From the choice of $\mu > 0$, we see that the last term on the right-hand side of the above is nonnegative.

Taking the convex combination with $\{\theta_k^\alpha\}_{k=1}^{n(\alpha)}$ and then sending $\alpha \to 0$ with (i), (ii), (iv), and (v) of (4.4) in the above, we have

$$0 \geq u_\epsilon(x,t) + q + \max_{a\in A}\{-\langle g(x,a),p\rangle - f(x,a)\} - C_1\epsilon^2 e^{\mu t}.$$

Now, for $\delta > 0$, we choose $\eta_\delta \in C^\infty(\mathbf{R}^n)$ such that

$$0 \leq \eta_\delta \leq 1 \quad \text{in } \mathbf{R}^n, \qquad \eta_\delta = 1 \quad \text{in } \overline{\Sigma_\delta}, \qquad \text{and} \qquad \eta_\delta = 0 \quad \text{in } \left(\Sigma_{\delta/2}\right)^c.$$

We set the functions

$$\begin{cases} g_\delta(x,a) = \eta_\delta(x) g(x,a), \\ f_{\epsilon,\delta}(x,t,a) = \eta_\delta(x)(f(x,a) + C_1\epsilon^2 e^{\mu t}) + (1-\eta_\delta(x))u_\epsilon(x,t), \\ f^{\epsilon,\delta}(x,t,a) = \eta_\delta(x)(f(x,a) - C_1'\epsilon^2 e^{-\mu t}) + (1-\eta_\delta(x))u^\epsilon(x,t). \end{cases}$$

We then consider the problems: For $(x,t,p,q) \in \mathbf{R}^n \times (0,T) \times \mathbf{R}^n \times \mathbf{R}$,

$$u + u_t + H_{\epsilon,\delta}(x,t,Du) = 0 \tag{4.6}$$

and

$$u + u_t + H^{\epsilon,\delta}(x,t,Du) = 0, \tag{4.7}$$

where

$$\begin{cases} H_{\epsilon,\delta}(x,t,p) = \max_{a\in A}\{-\langle g_\delta(x,a),p\rangle - f_{\epsilon,\delta}(x,t,a)\}, \\ H^{\epsilon,\delta}(x,t,p) = \max_{a\in A}\{-\langle g_\delta(x,a),p\rangle - f^{\epsilon,\delta}(x,t,a)\}. \end{cases}$$

In what follows, we suppose that $\delta > \frac{\epsilon}{2\hat{c}}$.

We claim that u_ϵ and u^ϵ, respectively, are the standard viscosity subsolution and supersolution of $u + u_t + H_{\epsilon,\delta} = 0$ and $u + u_t + H^{\epsilon,\delta} = 0$ in $\mathbf{R}^n \times (0,T)$. For $(x,t) \in \mathbf{R}^n \times (0,T)$,

$$u_\epsilon(x,t) + q + H_{\epsilon,\delta}(x,t,p) \leq 0 \quad \text{provided } (p,q) \in D^+u_\epsilon(x,t) \tag{4.8}$$

and

$$u^\epsilon(x,t) + q + H^{\epsilon,\delta}(x,t,p) \geq 0 \quad \text{provided } (p,q) \in D^-u^\epsilon(x,t). \tag{4.9}$$

Indeed, it is immediate to check that u_ϵ and u^ϵ, respectively, satisfy that, for $(x,t) \in \mathbf{R}^n \times (0,T)$,

$$u_\epsilon(x,t) + \eta_\delta(x)q + H_{\epsilon,\delta}(x,t,p) \leq 0 \quad \text{provided } (p,q) \in D^+u_\epsilon(x,t)$$

and

$$u^\epsilon(x,t) + \eta_\delta(x)q + H^{\epsilon,\delta}(x,t,p) \geq 0 \quad \text{provided } (p,q) \in D^-u^\epsilon(x,t).$$

Here, we have used the fact $\eta_\delta(x) = 0$ for $x \notin \Sigma_{\hat{c}\epsilon}$.

We first show (4.9).

Since $(p,q) \in D^-u^\epsilon(x,t)$, from the definition, we have $q = \mu e^{\mu t}\frac{|x-y|^2}{\epsilon^2} \geq 0$ for some $y \in \Sigma$. Hence, we conclude our claim because $\eta_\delta \geq 0$.

Thus, for (4.8), it is sufficient to show that $q \le 0$ provided $(p,q) \in D^+u_\epsilon(x,t)$. This is not straightforward unlike (4.9).

However, in view of (iv) and (v) of Lemma 4.1, q can be approximated by $\sum_{k=1}^{n(\alpha)} \theta_k^\alpha q_k^\alpha$ (as $\alpha \to 0$) for $(p_k^\alpha, q_k^\alpha) \in D^- u_\epsilon(x_k^\alpha, t_k^\alpha)$ with appropriate (x_k^α, t_k^α). Hence, we can see that $q_k^\alpha = -\mu e^{-\mu t_k^\alpha} \frac{|x_k^\alpha - y|^2}{\epsilon^2} \le 0$ for some y. Therefore, $q \le 0$.

Now, we shall give the value functions $u_{\epsilon,\delta}$ and $u^{\epsilon,\delta}$, respectively, for (4.6) and (4.7) with initial condition $u_\epsilon(\cdot,0)$ and $u^\epsilon(\cdot,0)$:

$$u_{\epsilon,\delta}(x,t) = \inf_{\alpha\in\mathcal{A}} \left\{ \int_0^t e^{-s} f_{\epsilon,\delta}(X(s;x,\alpha), t-s, \alpha(s))ds + e^{-t} u_\epsilon(X(t;x,\alpha),0) \right\}$$

and

$$u^{\epsilon,\delta}(x,t) = \inf_{\alpha\in\mathcal{A}} \left\{ \int_0^t e^{-s} f^{\epsilon,\delta}(X(s;x,\alpha), t-s, \alpha(s))ds + e^{-t} u^\epsilon\big(X(t;x,\alpha),0\big) \right\}.$$

Since $f_{\epsilon,\delta}(x,t) \le f^{\epsilon,\delta}(x,t) + \epsilon^2(C_1 e^{\mu t} + C_1' e^{-\mu t})$ and $u_\epsilon \le u^\epsilon$, there exists $C_2 > 0$ such that

$$u_{\epsilon,\delta}(x,t) \le u^{\epsilon,\delta}(x,t) + C_2\epsilon^2 e^{\mu t} \quad \text{in } \mathbf{R}^n \times [0,T]. \tag{4.10}$$

We also remark that $u_{\epsilon,\delta}$ and $u^{\epsilon,\delta}$ are bounded and continuous. Hence, the standard comparison principle yields that

$$u_\epsilon(x,t) \le u_{\epsilon,\delta}(x,t) \qquad \text{and} \qquad u^{\epsilon,\delta}(x,t) \le u^\epsilon(x,t) \quad \text{in } \mathbf{R}^n \times [0,T]. \tag{4.11}$$

Fix $x \in \Sigma$ and choose $\delta > 0$ so that $x \in \Sigma_\delta$.

Then, the DPP for $u_{\epsilon,\delta}$ at (x,T) with (4.10) and (4.11) implies that

$$\begin{aligned}
u_\epsilon(x,T) &\le \inf_{\alpha\in\mathcal{A}} \left\{ \begin{array}{c} e^{-\tau_{\Sigma_\delta}^{x,\alpha}\wedge T} u_{\epsilon,\delta}\big(X(\tau_{\Sigma_\delta}^{x,\alpha}\wedge T;x,\alpha), (T-\tau_{\Sigma_\delta}^{x,\alpha})^+\big) \\ + \displaystyle\int_0^{\tau_{\Sigma_\delta}^{x,\alpha}\wedge T} e^{-s} f\big(X(s;x,\alpha),\alpha(s)\big)ds \end{array} \right\} \\
&\le \inf_{\alpha\in\mathcal{A}} \left\{ \begin{array}{c} e^{-\tau_{\Sigma_\delta}^{x,\alpha}\wedge T} u^{\epsilon,\delta}\big(X(\tau_{\Sigma_\delta}^{x,\alpha}\wedge T;x,\alpha), (T-\tau_{\Sigma_\delta}^{x,\alpha})^+\big) \\ + \displaystyle\int_0^{\tau_{\Sigma_\delta}^{x,\alpha}\wedge T} e^{-s} f\big(X(s;x,\alpha),\alpha(s)\big)ds \end{array} \right\} + C_2\epsilon^2 e^{\mu T} \qquad (4.12) \\
&\le u^\epsilon(x,T) + C_2\epsilon^2 e^{\mu T}.
\end{aligned}$$

We note that, for each $\alpha \in \mathcal{A}$, (4.10) and (4.11) imply

$$\begin{aligned}
u(X(\tau_{\Sigma_\delta}^{x,\alpha}\wedge T;x,\alpha)) &= \lim_{\epsilon\to 0} u_{\epsilon,\delta}(X(\tau_{\Sigma_\delta}^{x,\alpha}\wedge T;x,\alpha), (T-\tau_{\Sigma_\delta}^{x,\alpha})^+) \\
&= \lim_{\epsilon\to 0} u^{\epsilon,\delta}(X(\tau_{\Sigma_\delta}^{x,\alpha}\wedge T;x,\alpha), (T-\tau_{\Sigma_\delta}^{x,\alpha})^+).
\end{aligned} \tag{4.13}$$

Therefore, sending $\epsilon \to 0$ with (4.13) in (4.12), we have

$$u(x) = \inf_{\alpha\in\mathcal{A}} \left\{ \begin{array}{c} e^{-\tau_{\Sigma_\delta}^{x,\alpha}\wedge T} u\big(X(\tau_{\Sigma_\delta}^{x,\alpha}\wedge T;x,\alpha)\big) \\ + \displaystyle\int_0^{\tau_{\Sigma_\delta}^{x,\alpha}\wedge T} e^{-s} f\big(X(s;x,\alpha),\alpha(s)\big)ds \end{array} \right\}.$$

Finally, sending $T \to \infty$, we conclude the proof. □

Acknowledgments. We wish to thank Professor E. N. Barron for informing us of the manuscript [WZ]. We also wish to thank Professor M. Bardi for letting us know an interesting example (see Example 3.4) due to P. Soravia.

We finally wish to thank the referees for their suggestions on the first draft.

REFERENCES

[Ba] M. Bardi, *A boundary value problem for the minimum time function*, SIAM J. Control Optim., 27 (1989), pp. 776–785.

[BC] M. Bardi and I. Capuzzo Dolcetta, *Optimal Control and Viscosity Solutions of Hamilton-Jacobi-Bellman Equations*, Birkhäuser, Boston, 1996.

[BF] M. Bardi and M. Falcone, *An approximation scheme for the minimum time function*, SIAM J. Control Optim., 28 (1990), pp. 950–965.

[BS] M. Bardi and V. Staicu, *The Bellman equation for time-optimal control of noncontrollable nonlinear systems*, Acta Appl. Math., 31 (1993), pp. 201–223.

[BSo] M. Bardi and P. Soravia, *Hamilton-Jacobi equations with singular boundary conditions on a free boundary and applications to differential games*, Trans. Amer. Math. Soc., 325 (1991), pp. 205–229.

[B1] G. Barles, *Discontinuous viscosity solutions of first-order Hamilton-Jacobi equations: A guided visit*, Nonlinear Anal., 20 (1993), pp. 1123–1134.

[BJ1] E. N. Barron and R. Jensen, *Semicontinuous viscosity solutions of Hamilton-Jacobi equations with convex Hamiltonians*, Comm. Partial Differential Equations, 15 (1990), pp. 1713–1742.

[BJ2] E. N. Barron and R. Jensen, *Optimal control and semicontinuous viscosity solutions*, Proc. Amer. Math. Soc., 111 (1991), pp. 397–402.

[BL] E. N. Barron and W. Liu, *Semicontinuous and Continuous Blowup and Minimal Time Functions*, manuscript.

[C] O. Cârjă, *Lower semicontinuous solutions for a class of Hamilton-Jacobi-Bellman equations*, J. Optim. Theory Appl., 89 (1996), pp. 637–657.

[CMP] O. Cârjă, F. Mignanego, and G. Pieri, *Lower semicontinuous solutions of the Bellman equations for the minimum time problem*, J. Optim. Theory Appl., 85 (1995), pp. 563–574.

[EI] L. C. Evans and H. Ishii, *Differential games and nonlinear first order PDE on bounded domains*, Manuscripta Math., 49 (1984), pp. 109–139.

[EJ] L. C. Evans and M. R. James, *The Hamilton-Jacobi-Bellman equation for time-optimal control*, SIAM J. Control Optim., 27 (1989), pp. 1477–1489.

[F] H. Frankowska, *Lower semicontinuous solutions of Hamilton-Jacobi-Bellman equations*, SIAM J. Control Optim., 31 (1993), pp. 257–272.

[K] S. Koike, *Semicontinuous viscosity solutions for Hamilton-Jacobi equations with a degenerate coefficient*, Differential Integral Equations, 10 (1997), pp. 455–472.

[L] P.-L. Lions, *Optimal control of diffusion processes and Hamilton-Jacobi-Bellman equations: Part* 2*: Viscosity solutions and uniqueness*, Comm. Partial Differential Equations, 8 (1983), pp. 1229–1276.

[S1] P. Soravia, *Discontinuous viscosity solutions to Dirichlet problems for Hamilton-Jacobi equations with convex Hamiltonians*, Comm. Partial Differential Equations, 19 (1993), pp. 1493–1514.

[S2] P. Soravia, *Optimality principles and representation formulas for viscosity solutions of Hamilton-Jacobi equations,* I*: Equations of unbounded and degenerate control problems without uniqueness*, Adv. Differential Equations, 4 (1999), pp. 275–296.

[S3] P. Soravia, *Optimality principles and representation formulas for viscosity solutions of Hamilton-Jacobi equations,* II*: Equations of control problems with state constraints*, Differential Integral Equations, 12 (1999), pp. 275–293.

[WZ] P. R. Wolenski and Y. Zhuang, *Proximal analysis and the minimal time function*, SIAM J. Control Optim., 36 (1998), pp. 1048–1072.

SIAM J. CONTROL OPTIM.
Vol. 38, No. 2, pp. 482–502

OBSERVATIONS PREPROCESSING AND QUANTIZATION FOR NONLINEAR FILTERS*

NIGEL J. NEWTON†

Abstract. Methods of *preprocessing* the observations for nonlinear filters are investigated, the aim being to reduce the computational labor involved in their implementation. An example of preprocessing is *quantization*, which involves the replacement of real-valued observation samples by discrete values. Because the resulting likelihood functions have finite support, they can be held in "look-up" tables kept in some type of rapid access memory, which renders their "real-time" evaluation trivial. Preprocessed observations of this sort carry less information than *raw* observations. This loss of information is characterized here for filters with substantially noisy observation samples by means of a functional central limit theorem. Among other things, this supplies an asymptotic, effective signal-to-noise ratio for preprocessed filters.

Methods of optimizing preprocessing operations with respect to this quantified information loss are developed. In particular, optimal quantization thresholds are found for observations that are contaminated by Gaussian noise, and it is shown that the loss of information for quite coarse quantization schemes is small; for example, the asymptotic, effective signal-to-noise ratio for a filter with one-bit quantized observations is $2/\pi$ times that for the same filter with raw observations. Simulations on two examples demonstrate the validity of the asymptotic characterization, even when the observation samples are only modestly contaminated by noise.

Key words. nonlinear filtering, quantization, approximation, weak convergence, hidden Markov models

AMS subject classifications. 93E11, 60G35, 60F17

PII. S0363012997331147

1. Introduction.

Nonlinear filtering concerns methods for progressively estimating the value of a *signal* process from the available history of a related *observations* process, where, typically, the dynamics of both processes are nonlinear and influenced by random noise. The theory of such filters, in a fairly general context, is well understood (see [2], [4], [7], [8], [11], [18], and [19]), as are aspects of their approximation, in particular those relating to continuous-time filters (see, for example, [1], [3], [5], [12], [14], and [17]), but their implementation in many areas of potential application remains difficult. This is due largely to the computational complexity of the algorithms involved, which is often orders of magnitude greater than for linear filters. One aspect of implementation is the evaluation of likelihood functions for the observations, and this can be simplified by means of *preprocessing*. An example of preprocessing, which motivated this study, is *quantization*; this involves the replacement of real-valued observation samples by discrete samples. The likelihood functions for these discrete observations can be held in "look-up" tables kept in some type of rapid access memory, which renders their "real-time" evaluation trivial. Another example of preprocessing is *dimension reduction*, where an observation sample from a high-dimensional space is replaced by one from a space of lower dimension, with obvious computational advantages. Of course, such preprocessing operations reduce the information content of the observations, making the filters that subsequently use them less accurate than those using raw observations. The primary concern of this

*Received by the editors December 5, 1997; accepted for publication (in revised form) April 19,1999; published electronically January 19, 2000.
http://www.siam.org/journals/sicon/38-2/33114.html
†Department of Electronic Systems Engineering, University of Essex, Wivenhoe Park, Colchester, CO4 3SQ, UK (njn@essex.ac.uk).

article is to quantify this loss and so enable the design of preprocessing operations that minimize it within various sets of constraints. In particular, methods for determining the optimal placement of thresholds in quantization schemes are developed.

Attention is restricted, here, to filters for Markov chain signals. Preprocessing techniques for estimators of continuous-state signals (including filters for diffusions) are considered in a companion paper [16]. All the results concern filters for which the observation samples are substantially contaminated by noise. This includes not only discrete-time filters with inherently noisy observation sequences but also continuous-time filters that have been discretized in time with a small time step. For example, suppose we wish to estimate a signal, $(X_t;\, t \in [0,\infty))$, modeled by a Markov chain with rate matrix A, from continuous-time observations modeled by

$$Z_t = \int_0^t g(X_s)ds + W_t \qquad \text{for } t \in [0,\infty), \tag{1.1}$$

where g is an $\mathbb{R}^n$-valued function and $(W_t;\, t \in [0,\infty))$ is an n-dimensional standard Brownian motion; then, modulo some technical assumptions, we can calculate the posterior probabilities that X_t occupies the various states, given the observations up to time t, by means of Wonham's filter [18],

$$d\pi_t = \left(A - \sum_{j=1}^{n} (G_j - \bar{g}_{t,j} I)\, \bar{g}_{t,j} \right) \pi_t\, dt + \sum_{j=1}^{n} (G_j - \bar{g}_{t,j} I)\, \pi_t\, dZ_{t,j}, \tag{1.2}$$

where

$$\begin{aligned} \pi_t &= \mathrm{vec}_i \left\{ P(X_t = x_i \mid Z_s;\, 0 \le s \le t) \right\}, \\ G_j &= \mathrm{diag}_i \{ g_j(x_i) \}, \\ \bar{g}_t &= \sum_i g(x_i) \pi_{t,i}, \end{aligned}$$

and I is the matrix identity. ($\mathrm{vec}_i(y_i)$ is used to indicate the column vector, whose components are the elements y_i.) Time discretizations of (1.2), with time step Δt, typically involve the sampled (and, here, normalized) observations

$$\begin{aligned} Y_k &= \Delta t^{-1} (Z_{(k+1)\Delta t} - Z_{k\Delta t}) \qquad \text{for } k = 0, 1, \ldots \\ &\approx g(X_{k\Delta t}) + \Delta t^{-1} (W_{(k+1)\Delta t} - W_{k\Delta t}), \end{aligned}$$

which become increasingly noisy as Δt decreases.

In the next section, preprocessing operations for the observation sequences of discrete-time Markov chain filters are characterized by a central limit theorem. The limit is taken over sequences of filters with worsening observation noise and slowing signal dynamics. The slowing of the signal dynamics compensates for the worsening of the noise in such a way that the performance of the filters converges to that of a continuous-time limit filter. Section 3 develops applications of this result to the time discretization of filters for continuous-time Markov chain signals. The limit filter for both discrete and continuous-time cases is characterized by a matrix-valued signal-to-noise ratio, H. In section 4, this is related to other measures of performance such as error probability and mean-square error. The results of some simulations are presented in section 5. These suggest that the approach of optimizing preprocessing operations with respect to the performance measure of the limit problem is a useful

one, in that the performance measures for the limit filter closely approximate those for the filters in the sequence, even when the observation noise is only modest. This is true over a wide range of filter performances, as demonstrated by the simulations in Example 1 of section 5.

The key feature of filters that are well approximated by the limit filter is that, as a result of a combination of observation noise level and signal switching rate, they act "long term"; i.e., they make significant use of a long history of observation samples, taking little notice of a single sample. Another asymptotic performance measure appropriate for such filters is obtained if the signal dynamics are slowed but the observation noise remains fixed. In that case, the performance of the filters improves asymptotically—in particular, the error probability converges to zero. An asymptotic analysis for the steady-state error probability for filters based on raw observations is developed in [9] and [10]. This can be adapted for filters based on preprocessed observations and used as a basis for optimizing the preprocessing operations with respect to the specific performance criterion of error probability. This approach is complementary to that of section 2 in that it provides asymptotic analysis for filters with any level of observation noise, but is useful only if the filters have small error probabilities. In the context of quantization it prescribes optimal thresholds that differ from those derived in section 4 when the level of observation noise is low. Of course, for filters with high levels of observation noise and small error probabilities the two approaches yield equivalent results.

However, it is not clear how useful this alternative asymptotic analysis is for filters with low levels of observation noise. Like the analysis developed here, it relies on the filters acting "long term," which means that, when the level of observation noise is low, it is useful only if the signal switching rate is extremely low. In Example 1 of section 5, the analysis of [10] corresponds to the limit as the parameter a goes to zero. The optimal thresholds according to the two asymptotic performance measures broadly coincide for this example for values of δ (the reciprocal of the noise standard deviation) less than 1. For values of δ greater than 1 the performance of the filter is dominated by "short-term" effects (even when $a = 0.0001$). The value of a would need to be extremely small for a "window" of noise intensities, lower than those for which the high-noise asymptotic used here applies, but higher than those for which short-term effects dominate the filter performance, to open. The asymptotic analysis in [10], applied to preprocessing, may allow the fine tuning of operations for certain examples with low observation noise intensities and very low error probabilities, but this is not investigated further here.

2. An asymptotic characterization of preprocessing operations. In this section, the information lost through preprocessing operations used in conjunction with Markov chain filters is estimated by means of a central limit theorem.

Let $\{(\Omega^\delta, \mathcal{F}^\delta, P^\delta, (X_k^\delta \in \{x_1, x_2, \dots, x_m\};\ k = 0, 1, \dots), Q^\delta, p); \delta \in (0, 1]\}$ be a family of discrete-time, time-homogeneous, Markov chains parametrized by δ, and having transition probability matrices Q^δ and common initial law p $(= \mathrm{vec}_i\{P(X_0^\delta = i)\})$. For each δ, let $(Y_k^{h,\delta} \in M;\ k = 1, 2, \dots)$ be a *preprocessed* observation sequence whose terms take values in some measurable space $(M, \mathcal{M})$ and are defined by

$$Y_k^{h,\delta} = h\left(\delta g(X_k^\delta) + \zeta_k^\delta\right) \quad \text{for } k = 1, 2, \dots, \tag{2.1}$$

where $(\zeta_k^\delta \in \mathbb{R}^n;\ k = 1, 2, \dots)$ is a sequence of independent random variables (noises), defined on $(\Omega^\delta, \mathcal{F}^\delta, P^\delta)$, that are independent of $(X_k^\delta;\ k = 0, 1, \dots)$ and have com-

mon distribution $P_\zeta(dz)$, g is an $\mathbb{R}^n$-valued function on $\{x_1, x_2, \ldots, x_m\}$, and h is a measurable (preprocessing) function from $\mathbb{R}^n$ to M. (For example, quantization.)

For $z \in \mathbb{R}^n$, let $q(dy, z)$ be the probability measure induced on $(M, \mathcal{M})$ by $h(z+\zeta)$ and suppose that:

(H1) there exists a constant $\theta > 0$ such that

$$q(dy, z) \ll q(dy, 0) \quad \text{if } \|z\| < \theta.$$

Let $r(y, z)$ be a version of the associated Radon–Nikodým derivative, i.e.,

$$r(y, z) = \frac{q(dy, z)}{q(dy, 0)} \quad \text{almost everywhere (a.e.)} - q(dy, 0).$$

For sufficiently small δ, the discrete-time nonlinear filter for estimating X_k^δ from the observations $(Y_1^{h,\delta}, Y_2^{h,\delta}, \ldots, Y_k^{h,\delta})$ can be expressed recursively as follows. (Note that, for large values of δ, it may be possible within the constraints of (H1), for the measures $q(dy, g(x_i)\delta)$ and $q(dy, g(x_j)\delta)$ to be mutually singular for one or more pairs of states (x_i, x_j). A recursive expression for the nonlinear filter would then necessarily be different from the following.) Let

$$\pi_k^{h,\delta} = \mathrm{vec}_i \left\{ P^\delta \left(X_k^\delta = x_i \mid Y_1^{h,\delta}, Y_2^{h,\delta}, \ldots, Y_k^{h,\delta} \right) \right\}; \tag{2.2}$$

then from Bayes' formula it follows that

$$\begin{aligned} \pi_0^{h,\delta} &= p, \\ \pi_{k+1}^{h,\delta} &= S\left(R(Y_{k+1}^{h,\delta}, \delta) Q^\delta \pi_k^{h,\delta} \right) \quad \text{for } k = 0, 1, \ldots, \end{aligned} \tag{2.3}$$

where

$$R(y, \delta) = \mathrm{diag}_i \{ r(y, g(x_i)\delta) \},$$

and $S : \mathbb{R}^m \to \mathbb{R}^m$ is the following normalization function:

$$S_i(u) = \begin{cases} \left(\sum_{j=1}^m |u_j| \right)^{-1} u_i & \text{if } u \neq 0 \\ 0 & \text{if } u = 0 \end{cases} \qquad \text{for } i = 1, 2, \ldots, m. \tag{2.4}$$

THEOREM 2.1. *Suppose that, in addition to* (H1),

(H2) *for almost every* y $(q(dy, 0))$, $r(y, \cdot)$ *has continuous first and second derivatives in some neighborhood of the origin (in what follows,* $r_z(y, \cdot)$ *and* $r_{zz}(y, \cdot)$ *will represent, respectively, the corresponding row-vector Jacobian and the Hessian matrix);*

(H3) *there exists an* $\epsilon > 0$ *such that*

$$\int \|r_z(y, 0)\|^{3+\epsilon} \, q(dy, 0) < \infty;$$

(H4) *there exist* $\epsilon, \theta > 0$ *such that*

$$\sup_{\|z\| < \theta} \int \|r_{zz}(y, z)\|^{3+\epsilon} \, q(dy, 0) < \infty;$$

(H5) *there exists an $m \times m$ Markov rate matrix, A, such that, for all δ,*

$$Q^\delta = I + A\delta^2 + o(\delta^2),$$

where $o(\cdot)$ (and, for later use, $O(\cdot)$) have their usual (Landau) meanings.
Let $(\Omega, \mathcal{F}, P, (X_t \in \{x_1, x_2, \ldots, x_m\};\, t \in [0,\infty)), A, p)$ be a continuous-time Markov process with rate matrix A and the same initial law, p, as (X_k^δ). Let H be the $n \times n$ matrix

$$H = \int r_z'(y,0) r_z(y,0) q(dy, 0), \tag{2.5}$$

and let $(Z_t^H;\, t \in [0,\infty))$ be the observations process

$$Z_t^H = H \int_0^t g(X_s) ds + B_t^H, \tag{2.6}$$

where g is as in (2.1) and $(B_t^H;\, t \in [0,\infty))$ is an n-dimensional Brownian motion on $(\Omega, \mathcal{F}, P)$, independent of $(X_t;\, t \in [0,\infty))$, and with covariance matrix H. Let $(\pi_t^H;\, t \in [0,\infty))$ be the nonlinear filter for (X_t) given (Z_t^H), i.e.,

$$\pi_t^H = \mathrm{vec}_i \left\{ P\left(X_t = x_i \mid Z_s^H;\, s \in [0,t] \right) \right\}. \tag{2.7}$$

Then the piecewise constant extension of $(\pi_k^{h,\delta};\, k = 0, 1, \ldots)$, $(\pi_{[\delta^{-2}t]}^{h,\delta};\, t \in [0,\infty))$, considered as a family of random variables in the Skorohod space $D_{\mathbb{R}^m}[0,\infty)$, converges weakly to $(\pi_t^H;\, t \in [0,\infty))$ as $\delta \to 0$. ($[x]$ signifies the integer part of x.)

Remark 1. If the rank, d, of H is nonzero, then $(Z_t^H;\, t \in [0,\infty))$ is equivalent to the following d-vector observations process:

$$\tilde{Z}_t^L = L \int_0^t g(X_s) ds + W_t,$$

where L is any $d \times n$ matrix such that $L'L = H$, and $(W_t;\, t \in [0,\infty))$ is a standard d-dimensional Brownian motion. Thus H can be interpreted as the (matrix-valued) signal-to-noise ratio of the limit filter (2.7), which is itself the weak limit of the discrete-time filters of (2.3) as the signal dynamics slow, (H5), and the observation noise worsens, (2.1).

Remark 2. The statement of the theorem remains true if the signal process is *controlled* in the sense that the matrix of transition probabilities, Q^δ, depends on $(\pi_k^{h,\delta})$, in which case $(X_k^\delta,\, \pi_k^{h,\delta})$ is Markov and

$$P^\delta \left(X_{k+1}^\delta = x_i \mid X_k^\delta = x_j,\, \pi_k^{h,\delta} \right) = Q_{i,j}^\delta(\pi_k^{h,\delta}).$$

Hypothesis (H5) is replaced by the following.

(H5′) There exists a Lipschitz continuous Markov rate function, $A : S_m \to \mathbb{R}^{m \times m}$ (where S_m is the simplex $\{x \in [0,1]^m : \sum_i x_i = 1\}$), such that, for all δ,

$$Q^\delta(p) = I + A(p)\delta^2 + o(\delta^2) \qquad \text{uniformly in } p \in S_m;$$

the requirement that the observation noise, $(\zeta_k^\delta;\, k = 1, 2, \ldots)$, be independent of the signal, $(X_k^\delta;\, k = 0, 1, \ldots)$, is replaced by the requirement that $(\zeta_{k+1}^\delta, \zeta_{k+2}^\delta, \ldots)$ be

independent of X_k^δ for each k, and the limit filter is that for the *controlled*, continuous-time process $(X_t,\ A(\pi_t^H))$, given the observations (2.6), where X_s and $B_t^H - B_s^H$ are independent for all $0 \le s \le t < \infty$. The steps in the proof are the same as those for the uncontrolled case but involve more complex notation.

Proof of Theorem 2.1. The *unnormalized* version of the discrete-time filters (2.3),

$$\begin{aligned}\rho_0^{h,\delta} &= p,\\ \rho_{k+1}^{h,\delta} &= R(Y_{k+1}^{h,\delta},\delta)Q^\delta \rho_k^{h,\delta} \quad \text{for } k = 0,1,\dots\end{aligned}$$

is first expanded in terms of δ. Because of (H2), for sufficiently small δ, the Radon–Nikodým derivatives, $r(y, g(x_i)\delta)$, can be expanded as follows:

$$r(y, g(x_i)\delta) = 1 + r_z(y,0)g(x_i)\delta + f(y,x_i,\delta)\delta^2,$$

where

$$f(y,x_i,\delta) = g'(x_i)\int_0^1\int_0^\beta r_{zz}(y,\alpha g(x_i)\delta)d\alpha\, d\beta\, g(x_i).$$

This, together with (H5), gives

$$\rho_{k+1}^{h,\delta} = \rho_k^{h,\delta} + F(Y_{k+1}^{h,\delta},\delta)\rho_k^{h,\delta},$$

where

$$\begin{aligned}F(y,\delta) = A\delta^2 &+ \operatorname{diag}_i\{r_z(y,0)g(x_i)\}\delta + \operatorname{diag}_i\{f(y,x_i,\delta)\}O(\delta^2)\\ &+ (I + \operatorname{diag}_i\{r_z(y,0)g(x_i)\})\, o(\delta^2).\end{aligned}$$

Now, $(\rho_k^{h,\delta})$ is a Markov process in $\mathbb{R}^m$ with (time-homogeneous) transition function $\mu(u,B)$ given by

$$\mu(u,B) = P_u\left(u + F(Y,\delta)u \in B\right),$$

where

$$Y = h(\delta g(X) + \zeta),$$

X ($\in \{x_1, x_2, \dots, x_m\}$), and ζ ($\in \mathbb{R}^n$) are independent random variables with distributions $S(Q^\delta u)$ and $P_\zeta(dz)$, respectively, defined on some probability space $(\Omega_u, \mathcal{F}_u, P_u)$, and S is the normalization function defined in (2.4). Because of (H3) and (H4), for sufficiently small δ,

$$\begin{aligned}\mathbf{E}_u\|F(Y,\delta)\|^{2+\epsilon} &= \sum_j \int \|F(y,\delta)\|^{2+\epsilon} q(dy, g(x_j)\delta)(Q^\delta S(u))_j\\ &= \sum_j \int \|F(y,\delta)\|^{2+\epsilon}\left(1 + r_z(y,0)g(x_j)\delta + f(y,x_j,\delta)\delta^2\right)\\ &\qquad\qquad \times q(dy,0)(Q^\delta S(u))_j\\ &= O(\delta^{2+\epsilon}) \qquad \text{uniformly in } u,\end{aligned}$$

and so by Markov's inequality, for any $\theta > 0$,

$$\begin{aligned}\delta^{-2}P_u\left(\|F(Y,\delta)\| > \theta\right) &\le \theta^{-(2+\epsilon)}\delta^{-2}\mathbf{E}_u\|F(Y,\delta)\|^{2+\epsilon}\\ &= O(\delta^\epsilon) \quad \text{uniformly in } u.\end{aligned} \tag{2.8}$$

Again, for sufficiently small δ,

$$\begin{aligned}\mathbf{E}_u r_z(Y,0) &= \sum_j \int r_z(y,0) q(dy, g(x_j)\delta)(Q^\delta S(u))_j \\ &= \sum_j \int r_z(y,0)\left(1 + r_z(y,0)g(x_j)\delta + f(y,x_j,\delta)\delta^2\right) q(dy,0)(Q^\delta S(u))_j \\ &= \bar{g}'(u)H\delta + O(\delta^2) \qquad \text{uniformly in } u,\end{aligned}$$

where

$$\bar{g}(u) = \sum_j g(x_j)(S(u))_j.$$

Similarly, for θ as defined in (H4),

$$\sup_{\|z\|<\theta} \mathbf{E}_u r_{zz}(Y,z) = O(\delta) \qquad \text{uniformly in } u,$$

and so

$$\begin{aligned}\mathbf{E}_u F(Y,\delta) &= A\delta^2 + \operatorname{diag}_i\{\bar{g}'(u)Hg(x_i)\}\delta^2 + o(\delta^2) \\ &= A\delta^2 + \sum_{l=1}^n (H\bar{g}(u))_l G_l \delta^2 + o(\delta^2)\end{aligned} \tag{2.9}$$

uniformly in u, where $G_l = \operatorname{diag}_i\{g_l(x_i)\}$. Also, for sufficiently small δ,

$$\begin{aligned}\mathbf{E}_u F(Y,\delta)uu'F'(Y,\delta) &= \mathbf{E}_u \operatorname{diag}_i\{r_z(Y,0)g(x_i)\}uu' \operatorname{diag}_i\{r_z(Y,0)g(x_i)\}\delta^2 + O(\delta^3) \\ &= \sum_{k,l=1}^n H_{k,l}G_k uu' G_l \delta^2 + O(\delta^3)\end{aligned} \tag{2.10}$$

uniformly on compacts.

The continuous-time filter (2.7) is Wonham's filter, which can be expressed recursively in unnormalized form as follows:

$$\begin{aligned}\rho_0^H &= p, \\ d\rho_t^H &= A\rho_t^H dt + \sum_{l=1}^n G_l \rho_t^H dZ_{t,l}^H \\ &= \left(A + \sum_{l=1}^n \left(H\bar{g}\left(\rho_t^H\right)\right)_l G_l\right)\rho_t^H dt + \sum_{l=1}^n G_l \rho_t^H d\nu_{t,l}^H, \\ \pi_t^H &= S(\rho_t^H),\end{aligned} \tag{2.11}$$

where

$$\nu_t^H := Z_t^H - \int_0^t H\bar{g}(\rho_s)ds$$

is the associated *innovations process*.

The coefficients in (2.11) fulfill Itô's Lipschitz continuity and linear growth conditions and so (2.11) has a unique strong solution. It follows that the associated

martingale problems for each initial law on $(\mathbb{R}^m, \mathcal{B}^m)$ all have unique solutions. This, and properties (2.8), (2.9), and (2.10) of the transition function, $\mu(u, B)$, allow the application of a functional central limit theorem (a simple variation of Lemma 4.2 on p. 355 of [6]) showing that, for any sequence $\delta_l \to 0$,

$$\left(\rho^{h,\delta_l}_{[\delta_l^{-2}t]}\right) \Rightarrow \left(\rho^H_t\right) \quad \text{in } D_{\mathbb{R}^m}[0,\infty).$$

The theorem now follows from the a.e. (with respect to the distribution of (ρ^H_t)) continuity of the mapping

$$\Gamma : D_{\mathbb{R}^m}[0,\infty) \to D_{\mathbb{R}^m}[0,\infty),$$

defined by

$$\Gamma_t(v) = S(v_t) \qquad \text{for all } t \in [0,\infty). \qquad \square$$

The following examples illustrate some preprocessing operations and noise distributions that fulfill (H1)–(H4), and give the corresponding signal-to-noise ratios for the limit filter.

Example 1. Lossless preprocessing. If h is 1–1, and the noise distribution, $P_\zeta(dz)$, has a density, $p(z)$, with the following properties:

(H1-1) $p(z) > 0$ for all z;

(H2-1) $p(z)$ has continuous first and second derivatives (denoted $p_z(z)$ and $p_{zz}(z)$);

(H3-1) there exists an $\epsilon > 0$ such that

$$\int \frac{\|p_z(u)\|^{3+\epsilon}}{p^{2+\epsilon}(u)} du < \infty;$$

(H4-1) there exist $\epsilon, \theta > 0$ such that

$$\sup_{\|z\|<\theta} \int \frac{\|p_{zz}(u-z)\|^{3+\epsilon}}{p^{2+\epsilon}(u)} du < \infty;$$

then $(h, P_\zeta(dz))$ fulfills (H1)–(H4), and

$$H = \int \frac{p'_z(u) p_z(u)}{p(u)} du.$$

Two specific examples of noise distributions satisfying the above are Gaussian,

$$p(z) = \frac{1}{(2\pi)^{n/2}|V|^{1/2}} \exp\left(-\tfrac{1}{2}(z-\mu)' V^{-1}(z-\mu)\right)$$

for which $H = V^{-1}$, and Cauchy (with $n = 1$),

$$p(z) = \frac{a}{\pi(1+(az)^2)}$$

for which $H = a^2$. This latter example illustrates an important difference between linear and nonlinear filtering. At first sight, it would seem reasonable to estimate a slowly varying signal given a large number of noisy observations by some sort of linear, moving-average filter. However, this would be completely fruitless in this case, where the noise distribution is "heavy tailed." The nonlinear filter does use a moving average

technique but applies it to the log-likelihood function of the observations, rather than the observations themselves. This example also illustrates that the discrete-time filters (2.3) need not themselves have signal-to-noise ratios in the mean-square sense in order for Theorem 2.1 to apply.

Of course, the preprocessing in this example is artificial in that it does not sacrifice any information, and so the signal-to-noise ratio of the limit filter, H, is that corresponding to the *raw* observations.

Example 2. Quantization. Let h be the discrete-valued function

$$h(z) = \sum_{i=1}^{N} b_i \chi_{B_i}(z),$$

where N could be infinite, the b_i are distinct elements of M, the B_i are disjoint Borel sets in $\mathbb{R}^n$ with nonzero Lebesgue measure whose union is $\mathbb{R}^n$, and χ_{B_i} is the indicator function of the set B_i. If $P_\zeta(dz)$ has a density, $p(z)$, with the following properties:

(H1-2) $p(z) > 0$ for all z;

(H2-2) $p(z)$ has continuous first and second derivatives ($p_z(z)$ and $p_{zz}(z)$);

(H3-2) there exists an $\epsilon > 0$ such that

$$\sum_i \left\| \int_{B_i} p_z(u)du \right\|^{3+\epsilon} P_\zeta(B_i)^{-(2+\epsilon)} < \infty;$$

(H4-2) there exist $\epsilon, \theta > 0$ such that

$$\sup_{\|z\|<\theta} \sum_i \left\| \int_{B_i} p_{zz}(u-z)du \right\|^{3+\epsilon} P_\zeta(B_i)^{-(2+\epsilon)} < \infty;$$

then $(h, P_\zeta(dz))$ fulfills (H1)–(H4), and

$$H = \sum_i \int_{B_i} \int_{B_i} p_z'(u) p_z(v) du\, dv\, P_\zeta(B_i)^{-1}.$$

The conditions on $P_\zeta(dz)$ can be weakened if the sets B_i are sufficiently regular. For example, there is no need for $P_\zeta(dz)$ to have a density at all interior points of the B_i.

Example 3. Rectangular quantization. Let h be the discrete-valued function

$$h(z) = \sum_{j_1=0}^{N_1} \sum_{j_2=0}^{N_2} \cdots \sum_{j_n=0}^{N_n} b_J \chi_{B_J}(z),$$

where $N_i < \infty$ for all i, $J = (j_1, j_2, \ldots, j_n)$, the b_J are distinct elements of M, and the B_J are the following n-dimensional rectangles:

$$B_J = (t_{1,j_1}, t_{1,j_1+1}) \times (t_{2,j_2}, t_{2,j_2+1}) \times \cdots \times (t_{n,j_n}, t_{n,j_n+1}).$$

Here, for each i,

$$-\infty = t_{i,0} < t_{i,1} < \cdots < t_{i,N_i} < t_{i,N_i+1} = \infty.$$

Let A be the set of boundaries

$$A = \mathbb{R}^n - \cup_J B_J.$$

If, for some $\alpha > 0$, $P_\zeta(dz)$ has a density, $p(z)$, in the neighborhood

$$N_\alpha := \{z \in \mathbb{R}^n : \|z - a\| < \alpha \quad \text{for some } a \in A\},$$

with the following properties:

(H1-3) $p(z) > 0$ for all $z \in N_\alpha$;

(H2-3) $p(z)$ has a continuous first derivative on N_α ($p_z(z)$); then (H1)–(H4) are satisfied, and

$$H = \sum_{j_1=0}^{N_1} \sum_{j_2=0}^{N_2} \cdots \sum_{j_n=0}^{N_n} \operatorname{vec}_i \{\lambda_{J,i}(t_{i,j_i}) - \lambda_{J,i}(t_{i,j_i+1})\} \times \operatorname{vec}_i' \{\lambda_{J,i}(t_{i,j_i}) - \lambda_{J,i}(t_{i,j_i+1})\} P_\zeta(B_J)^{-1},$$

where

$$\lambda_{J,i}(t) = \begin{cases} \displaystyle\int_{\bar{B}_J \cap \{z_i = t\}} p(z)\, dz_1\, dz_2 \ldots dz_{i-1}\, dz_{i+1} \ldots dz_n & \text{if } t \text{ is finite,} \\ 0 & \text{if } t \text{ is infinite,} \end{cases}$$

and $\bar{B}_J$ is the closure of B_J. In particular, if $n = 1$,

$$H = \frac{p(t_1)^2}{P_\zeta((-\infty, t_1))} + \sum_{j=1}^{N-1} \frac{(p(t_j) - p(t_{j+1}))^2}{P_\zeta((t_j, t_{j+1}))} + \frac{p(t_N)^2}{P_\zeta((t_N, \infty))}. \tag{2.12}$$

Example 4. Linear dimension reduction. If $M = \mathbb{R}^d$, $h(z) = Bz$ for some $d \times n$ matrix, B, with rank d ($< n$), and the noise distribution has a density, $p(z)$, with the following properties:

(H1-4) $\int_{\mathbb{R}^{n-d}} p(Vv - UU'z)dv > 0$ for all z;

(H2-4) $p(z)$ has continuous first and second derivatives ($p_z(z)$ and $p_{zz}(z)$);

(H3-4) there exists an $\epsilon > 0$ such that

$$\int_{\mathbb{R}^d} \frac{\left\| \int_{\mathbb{R}^{n-d}} p_z(Vv + Uu)dv \right\|^{3+\epsilon}}{\left(\int_{\mathbb{R}^{n-d}} p(Vv + Uu)dv \right)^{2+\epsilon}} du < \infty;$$

(H4-4) there exist $\epsilon, \theta > 0$ such that

$$\sup_{\|z\| < \theta} \int_{\mathbb{R}^d} \frac{\left\| \int_{\mathbb{R}^{n-d}} p_{zz}(Vv + Uu - UU'z)dv \right\|^{3+\epsilon}}{\left(\int_{\mathbb{R}^{n-d}} p(Vv + Uu)dv \right)^{2+\epsilon}} du < \infty;$$

where V and U are, respectively, $n \times (n-d)$- and $n \times d$-dimensional matrices whose columns form orthonormal bases for, respectively, the kernel of B and its orthogonal complement, then $(h, P_\zeta(dz))$ fulfills (H1)–(H4) and

$$H = UU' \int_{\mathbb{R}^d} \frac{\int_{\mathbb{R}^{n-d}} p_z'(Vv + Uu)dv \int_{\mathbb{R}^{n-d}} p_z(Vv + Uu)dv}{\int_{\mathbb{R}^{n-d}} p(Vv + Uu)dv} du\, UU'.$$

3. Continuous-time filters. Theorem 2.1 concerns discrete-time, discrete-state filters. Suppose now that $(\Omega, \mathcal{F}, P, (X_t \in \{x_1, x_2, \ldots, x_m\};\ t \in [0,\infty)), A, p)$ is a continuous-time Markov chain with rate matrix A and initial law p that we wish to estimate from observations of the continuous-time, n-vector process

$$Z_t = \int_0^t g(X_s)ds + W_t,$$

where $(W_t;\ t \in [0,\infty))$ is a standard n-dimensional Brownian motion on $(\Omega, \mathcal{F}, P)$, independent of $(X_t;\ t \in [0,\infty))$. One step in an implementation of the nonlinear filter for this problem is discretization in time. With a step size of δ^2, this yields (normalized) raw discrete-time observations of the following form:

$$\begin{aligned} Z_k^\delta &= \delta^{-1}(Z_{(k+1)\delta^2} - Z_{k\delta^2}) \\ &= \delta g(X_{k\delta^2}) + \delta^{-1}(V^\delta_{(k+1)\delta^2} - V^\delta_{k\delta^2}), \end{aligned}$$

where

$$V_t^\delta = \int_0^t g^\delta(X,s)ds + W_t$$

and

$$g^\delta(X,t) = g(X_t) - g(X_{[\delta^{-2}t]\delta^2}).$$

These observations differ from the raw observations of (2.1) in that the noise terms, $\delta^{-1}(V^\delta_{(k+1)\delta^2} - V^\delta_{k\delta^2})$, are statistically dependent on the discrete-time signal process, $(X_{k\delta^2};\ k = 0, 1, \ldots)$. However, the process $(g^\delta(X,t);\ t \in [0,\infty))$ is bounded, and so Girsanov's theorem can be used to define new probability measures on $(\Omega, \mathcal{F})$ under which the noise terms *are* independent of $(X_{k\delta^2};\ k = 0, 1, \ldots)$. (See, for example, Theorem 6.2 in [15].) In fact, if P_T^δ is defined by

$$\begin{aligned} \frac{dP_T^\delta}{dP} &= \exp\left(-\int_0^T g^\delta(X,t)'dW_t - \frac{1}{2}\int_0^T \|g^\delta(X,t)\|^2 dt\right) \\ &=: M_T^\delta, \end{aligned}$$

then P_T^δ is a probability measure under which the distribution of $(X_t;\ t \in [0,T])$ is unaltered but $(V_t^\delta;\ t \in [0,T])$ becomes a standard Brownian motion, independent of $(X_t;\ t \in [0,\infty))$. Thus Theorem 2.1 provides, under P_T^δ, a weak limit for the nonlinear filters for the discrete-time Markov chains, $(\Omega, \mathcal{F}, P_T^\delta, (X_{k\delta^2};\ k = 0, 1, \ldots), \exp(A\delta^2), p)$, given the preprocessed observations

$$Y_k^{h,\delta} = h(Z_k^\delta).$$

In fact, $(\pi^{h,\delta}_{[\delta^{-2}t]};\ t \in [0,\infty))$, defined by (2.2)–(2.4) with P^δ replaced by P_T^δ, and $Y_k^{h,\delta}$ as defined above, converges weakly in $D_{\mathbb{R}^m}[0,\infty)$ to the continuous-time filter of (2.5)–(2.7), i.e., the original continuous-time filter of this section, but with the new signal-to-noise ratio, H. What is not immediately apparent is that the same can be said for the nonlinear filter for $(X_{k\delta^2};\ k = 0, 1, \ldots)$ under the "true" measure, P, $(\tilde\pi^{h,\delta}_{[\delta^{-2}t]};\ t \in [0,\infty))$, where

$$\tilde\pi_k^{h,\delta} := \mathrm{vec}_i\left\{P\left(X_{k\delta^2} = x_i \mid Y_1^{h,\delta}, Y_2^{h,\delta}, \ldots, Y_k^{h,\delta}\right)\right\} \qquad \text{for } k = 0, 1, \ldots.$$

The following proposition addresses this issue and shows that the filters, $(\pi_k^{h,\delta};\ k = 0,1,\ldots)$, are strong approximations to the filters with respect to the true measure. This is of independent interest as the former are easier to implement than the latter. For example, in the case of n-bit quantization, the one-step update of the former filters consists of multiplication by one of 2^n precomputed matrices.

PROPOSITION 3.1.

(i) *For any $T < \infty$ and any $\theta \in (0,\infty)$,*

$$\lim_{\delta\to 0} \mathbf{E} \sup_{k\le \delta^{-2}T} \left\| \tilde{\pi}_k^{h,\delta} - \pi_k^{h,\delta} \right\|^\theta = 0.$$

(ii) *For any $T < \infty$ and any sequence $(\delta_l \in (0,1];\ l = 1,2,\ldots)$ with the property*

$$\sum_l \delta_l^{2-\epsilon} < \infty \qquad \textit{for some } \epsilon > 0, \tag{3.1}$$

$$\lim_{l\to\infty} \sup_{k\le \delta_l^{-2}T} \|\tilde{\pi}_k^{h,\delta_l} - \pi_k^{h,\delta_l}\| = 0 \qquad \textit{almost surely (a.s.).}$$

(iii) $(\tilde{\pi}^{h,\delta}_{[\delta^{-2}t]};\ t \in [0,\infty))$*, considered as a family of random variables in the Skorohod space $D_{\mathbb{R}^m}[0,\infty)$, converges weakly to the continuous-time filter of* (2.5)–(2.7).

Proof. For each k, let $\mathcal{Y}_k^{h,\delta}$ be the σ-field generated by $Y_1^{h,\delta}, Y_2^{h,\delta}, \ldots, Y_k^{h,\delta}$. From the basic properties of conditional expectation it follows that, for any $k < \infty$, any $C \in \mathcal{Y}_k^{h,\delta}$, and any $1 \le i \le m$,

$$\int_C \pi_{k,i}^{h,\delta} dP_T^\delta = \int_C \mathbf{E}\left(\chi_{\{x_i\}}(X_{k\delta^2}) M_T^\delta \mid \mathcal{Y}_k^{h,\delta}\right) \mathbf{E}_T^\delta\left((M_T^\delta)^{-1} \mid \mathcal{Y}_k^{h,\delta}\right) dP_T^\delta,$$

and so

$$\begin{aligned}\pi_{k,i}^{h,\delta} - \tilde{\pi}_{k,i}^{h,\delta} = {} & \mathbf{E}\left(\chi_{\{x_i\}}(X_{k\delta^2}) M_T^\delta \mid \mathcal{Y}_k^{h,\delta}\right) \mathbf{E}_T^\delta\left((M_T^\delta)^{-1} - 1 \mid \mathcal{Y}_k^{h,\delta}\right) \\ & + \mathbf{E}\left(\chi_{\{x_i\}}(X_{k\delta^2})(M_T^\delta - 1) \mid \mathcal{Y}_k^{h,\delta}\right) \qquad \text{a.s.},\end{aligned}$$

from which the triangle and Jensen's inequalities yield

$$\begin{aligned}\left\|\pi_k^{h,\delta} - \tilde{\pi}_k^{h,\delta}\right\| \le {} & \mathbf{E}\left(M_T^\delta \mid \mathcal{Y}_k^{h,\delta}\right) \mathbf{E}_T^\delta\left(|(M_T^\delta)^{-1} - 1| \mid \mathcal{Y}_k^{h,\delta}\right) \\ & + \mathbf{E}\left(|M_T^\delta - 1| \mid \mathcal{Y}_k^{h,\delta}\right).\end{aligned} \tag{3.2}$$

Since $g^\delta(X,t)$ is bounded,

$$\sup_{\delta\in(0,1]} \mathbf{E}(M_T^\delta)^{2j} < \infty \qquad \text{for all integers } j \text{ (positive or negative).} \tag{3.3}$$

Also, for any $\theta \in (0,\infty)$, there exists a $K < \infty$ such that

$$\begin{aligned}\sup_{t\in[0,\infty)} \mathbf{E}\|g^\delta(X,t)\|^\theta &\le K \sup_{t\in[0,\delta^2]} \sum_{i\ne j} \|g(x_j) - g(x_i)\|^\theta \left|\exp(At)_{i,j}\right| \\ &= O(\delta^2).\end{aligned}$$

Now, for some $\alpha \in (0,1)$,

$$M_T^\delta - 1 = -\left(\int_0^T g^\delta(X,t)' dW_t + \frac{1}{2} \int_0^T \|g^\delta(X,t)\|^2 dt \right) (M_T^\delta)^\alpha,$$

and so Hölder's inequality, (3.3), and a standard result on the moments of stochastic integrals (see, for example, Lemma 4.12 in [15]) show that, for any $\theta \in (0,\infty)$, there exists a $K < \infty$ such that

$$\begin{aligned} \mathbf{E}|M_T^\delta - 1|^\theta &\le K \left(\mathbf{E} \left| \int_0^T g^\delta(X,t)' dW_t + \frac{1}{2} \int_0^T \|g^\delta(X,t)\|^2 dt \right|^{\theta p} \right)^{1/p} \\ &= O(\delta^{2/p}), \end{aligned}$$

where $p = \min\{\alpha > 1 : \theta\alpha/2 \text{ is an integer}\}$. Similarly,

$$\mathbf{E}_T^\delta |(M_T^\delta)^{-1} - 1|^\theta = O(\delta^{2/p}).$$

Hölder's and Jensen's inequalities, when applied to (3.2), show that, for any $\theta \in [1,\infty)$ and any $q \in (1,\infty)$,

$$\begin{aligned} \mathbf{E} \sup_{k \le \delta^{-2} T} \left\| \pi_k^{h,\delta} - \tilde{\pi}_k^{h,\delta} \right\|^\theta &\le 2^{\theta-1} \left(\mathbf{E} \sup_k \mathbf{E} \left((M_T^\delta)^{\theta q/(q-1)} \mid \mathcal{Y}_k^{h,\delta} \right) \right)^{(q-1)/q} \\ &\quad \times \left(\mathbf{E} \sup_k \mathbf{E}_T^\delta \left(|(M_T^\delta)^{-1} - 1|^{\theta q} \mid \mathcal{Y}_k^{h,\delta} \right) \right)^{1/q} \\ &\quad + 2^{\theta-1} \left(\mathbf{E} \sup_k \mathbf{E} \left(|M_T^\delta - 1|^\theta \mid \mathcal{Y}_k^{h,\delta} \right) \right). \end{aligned} \tag{3.4}$$

Another application of Hölder's and Jensen's inequalities and Doob's submartingale inequality shows that

$$\begin{aligned} \mathbf{E} \sup_k \mathbf{E}_T^\delta \left(|(M_T^\delta)^{-1} - 1|^{\theta q} \mid \mathcal{Y}_k^{h,\delta} \right) &\le \left(\mathbf{E}_T^\delta (M_T^\delta)^{-q/(q-1)} \right)^{(q-1)/q} \\ &\quad \times \left(\mathbf{E}_T^\delta \sup_k \mathbf{E}_T^\delta \left(|(M_T^\delta)^{-1} - 1|^{\theta q^2} \mid \mathcal{Y}_k^{h,\delta} \right) \right)^{1/q} \\ &\le K \left(\mathbf{E}_T^\delta |(M_T^\delta)^{-1} - 1|^{\theta q^2} \right)^{1/q} \end{aligned}$$

for some $K < \infty$. A similar treatment of the other terms in the right-hand side of (3.4) shows that, for some $K < \infty$,

$$\begin{aligned} \mathbf{E} \sup_{k \le \delta^{-2} T} \left\| \tilde{\pi}_k^{h,\delta} - \pi_k^{h,\delta} \right\|^\theta &\le K \left(\mathbf{E}_T^\delta |(M_T^\delta)^{-1} - 1|^{\theta q^2} \right)^{1/q^2} + K \mathbf{E} |M_T^\delta - 1|^\theta \\ &= O(\delta^{2/p}), \end{aligned}$$

where p is as defined above. This proves (i). It also shows that, for any $0 < \epsilon < 1$,

$$\mathbf{E} \sup_{k \le \delta^{-2} T} \left\| \tilde{\pi}_k^{h,\delta} - \pi_k^{h,\delta} \right\|^{2-\epsilon} = O(\delta^{2-\epsilon}),$$

and this, together with (3.1) and the Borel–Cantelli lemma, proves (ii).

Now let

$$e_t^\delta = \left\| \pi^{h,\delta}_{[\delta^{-2}t]} - \tilde{\pi}^{h,\delta}_{[\delta^{-2}t]} \right\|.$$

For any $0 \le h \le t \le T - h$, if $\delta > \sqrt{2h}$,

$$\left|e_t^\delta - e_{t+h}^\delta\right| \left|e_t^\delta - e_{t-h}^\delta\right| = 0,$$

and so, from (i) and Hölder's inequality,

$$\begin{aligned}\sup_{\delta<1} \mathbf{E}\left(\left|e_t^\delta - e_{t+h}^\delta\right|^{\theta/2} \left|e_t^\delta - e_{t-h}^\delta\right|^{\theta/2}\right) &= \sup_{\delta\le\sqrt{2h}} \mathbf{E}\left(\left|e_t^\delta - e_{t+h}^\delta\right|^{\theta/2} \left|e_t^\delta - e_{t-h}^\delta\right|^{\theta/2}\right) \\ &= O(h^{1/p}),\end{aligned}$$

where p is as defined above. An application of the theorems on pages 137–139 in [6] now shows that the family $\{(\pi^{h,\delta}_{[\delta^{-2}t]} - \tilde{\pi}^{h,\delta}_{[\delta^{-2}t]}) : \delta \in (0,1]\}$ is relatively compact. This and part (i) of the proposition establish part (iii). □

Results of this type, but for a more general class of nonlinear estimation problems (including filters for diffusions), can be found in [16].

4. The rank ordering of preprocessing operations. As discussed earlier, the matrix H is the signal-to-noise ratio of the limit filter of Theorem 2.1, which we clearly want to be large in some sense. In fact, we typically want to choose h such that a cost of the following form is small:

$$\begin{aligned} C_\phi(h,\delta,k) &:= J_\phi(\pi_k^{h,\delta}) \\ &= \mathbf{E}\phi(\pi_k^{h,\delta}), \end{aligned} \tag{4.1}$$

where J_ϕ is the cost function corresponding to a continuous, convex (upward) function ϕ. A variation to this fixed-time cost criterion that is appropriate if $\pi_k^{h,\delta}$ has a steady-state distribution is a cost criterion with respect to this distribution:

$$C_\phi(h,\delta) = \lim_{k\to\infty} C_\phi(h,\delta,k).$$

For example, if we wanted to choose a value for X^δ from its range with the minimal probability of being in error, then an appropriate function, ϕ, would be

$$\phi(p) = -\max_i\{p_i\},$$

whereas, if we wanted to estimate a moment $\mathbf{E}f(X_k^\delta)$ with minimal mean-square error, then an appropriate ϕ would be

$$\phi(p) = -\left(\sum_i f(x_i)p_i\right)^2.$$

Theorem 2.1 shows that, under hypotheses (H1)–(H5), for any continuous ϕ,

$$J_\phi\left(\pi^{h,\delta}_{[\delta^{-2}t]}\right) \to J_\phi(\pi_t^H) \qquad \text{for all } t \in [0,\infty),$$

where (π_t^H) is the limit filter of the theorem, i.e., the costs for the discrete-time filters according to a given criterion converge to the cost for the limit filter according to the same criterion.

If $(\pi_k^{h,\delta})$ is an approximation to a continuous-time filter based on sampled and preprocessed observations, then it is appropriate for it to inherit the cost function of the continuous-time filter, in which case the cost for $(\pi_k^{h,\delta})$ will converge to the original cost for the continuous-time filter, but with the new signal-to-noise ratio, H.

Clearly, given a limit cost $J_\phi(\pi_t^H)$, we could, at least in principle, optimize the parameters of preprocessing operations for filters where δ is small. In practice, though, this can be difficult, and we may not have such a precise notion of cost if, for example, we need to estimate several features of the signal with no clear weighting of importance. However, there is still a partial ordering of preprocessing operations, which holds for all reasonable cost functions.

PROPOSITION 4.1. *Let $(X_t;\, t \in [0,\infty))$ be the continuous-time Markov chain of Theorem* 2.1, *and $(\pi_t^{H_1};\, t \in [0,\infty))$ and $(\pi_t^{H_2};\, t \in [0,\infty))$ the nonlinear filters for estimating it from observations $(Z_t^{H_1};\, t \in [0,\infty))$ and $(Z_t^{H_2};\, t \in [0,\infty))$, respectively, defined by* (2.6) *with matrices H_1 and H_2, and independent noises, $(B_t^{H_1};\, t \in [0,\infty))$ and $(B_t^{H_2};\, t \in [0,\infty))$. If $H_1 - H_2$ is positive semidefinite, then*

$$J_\phi(\pi_t^{H_1}) \le J_\phi(\pi_t^{H_2})$$

for all $t \in [0,\infty)$ and all cost functions of the form (4.1) *for which ϕ is continuous and convex upward.*

Proof. Set $H_3 = H_1 - H_2$, and let $(Z_t^{H_3};\, t \in [0,\infty))$ be a corresponding observations process with a noise process, $(B_t^{H_3})$, which is independent of $(B_t^{H_2})$. Examination of the nonlinear filtering equations (2.7) shows that the filter for (X_t), given the $2n$-dimensional observation process $(Z_t^{H_2}, Z_t^{H_3};\, t \in [0,\infty))$, involves only the sum

$$\begin{aligned} Z_t^{H_2} + Z_t^{H_3} &= (H_2 + H_3) \int_0^t g(X_s)ds + B_t^{H_2} + B_t^{H_3} \\ &= H_1 \int_0^t g(X_s)ds + B_t^{H_2} + B_t^{H_3}. \end{aligned}$$

Now $B_t^{H_2} + B_t^{H_3}$ is an n-dimensional Brownian motion with covariance matrix H_1, and so $(Z_t^{H_2} + Z_t^{H_3})$ is equivalent to $(Z_t^{H_1})$ for the filtering problem. Thus, by Jensen's inequality, setting $\mathcal{Z}_t^{H_2} = \sigma(Z_s^{H_2};\, s \in [0,t])$, etc.,

$$\begin{aligned} \mathbf{E}\phi\left(\pi_t^{H_1}\right) &= \mathbf{E}\phi\left(\mathrm{vec}_i\left\{P\left(X_t = x_i \mid \mathcal{Z}_t^{H_2} \vee \mathcal{Z}_t^{H_3}\right)\right\}\right) \\ &\le \mathbf{E}\mathbf{E}\left(\phi\left(\mathrm{vec}_i\left\{P\left(X_t = x_i \mid \mathcal{Z}_t^{H_2}\right)\right\}\right) \mid \mathcal{Z}_t^{H_3}\right) \\ &= \mathbf{E}\phi\left(\pi_t^{H_2}\right). \qquad \square \end{aligned}$$

The partial ordering of Proposition 4.1 becomes a complete ordering if the components of the noise process $(\zeta_i;\, i = 1, 2 \ldots, n)$ are independent, and the preprocessing function, h, consists of n separate functions of the n components of the raw observation. In particular, this is true if $n = 1$. Thus the choice of thresholds t_j in (2.12) that maximize H will be asymptotically optimal for the nonlinear filtering problem in terms of any of the cost functions of (4.1).

The optimal thresholds for zero-mean, unit-variance Gaussian noise (with a number of different degrees of quantization) are given, along with the corresponding values

TABLE 4.1
Optimal quantization thresholds for standard Gaussian noise (to three decimal places).

Number of bits	Thresholds								H
1	0								0.637
2	0	±0.982							0.883
3	0	±0.501	±1.050	±1.748					0.965
4	0	±0.258	±0.522	±0.800	±1.100	±1.437	±1.844	±2.401	0.990

for H in Table 4.1. The reduction in the signal-to-noise ratio of the limit filter arising through the use of 1-bit quantization is a factor of only $2/\pi$, and the reduction factors rapidly approach unity as the number of bits in the quantization increases. Clearly, if the noise has a nonstandard Gaussian distribution, then the thresholds should all be shifted by its mean and scaled by its standard deviation. The optimal thresholds of Table 4.1 are also pertinent to the quantization of multidimensional observations if the components of the noise are independent and Gaussian, and the components of the observations are quantized separately.

The asymptotic signal-to-noise ratio for 1-bit quantization is similar to that obtained by Kushner in [13] for a continuous-time system comprising a limiter, a gain control, and a linear filter. The input to this system is taken to be the sum of a (useful) signal and wide-band Gaussian noise. As the bandwidth of the noise increases, the output of the system converges weakly to the output that would be obtained if the limiter and gain control were removed, except that the signal-to-noise ratio is reduced by a factor that depends on the nature of the signal and the limiting behavior of the noise. If the signal is sinusoidal and the noise correlation is exponential, then the reduction factor obtained by Kushner is $2/(\pi \ln 2)$. The problem is somewhat different from that considered here, although both problems concern some form of accumulating larger and larger numbers of samples of an increasingly noise corrupted signal. Kushner's problem can be interpreted in this way by time scaling.

A generalization of the techniques discussed here includes *feedback* from the filter to the preprocessing operation, i.e., uses preprocessing operations of the following form:

$$Y_k^{h,\delta} = h\left(\delta g(X_k^\delta) + \zeta_k, \pi_{k-1}^{h,\delta}\right).$$

For example, we might use a quantization scheme with thresholds that change when $\pi_k^{h,\delta}$ moves between different regions of its range. The functions $h(z + \cdot, p)$ induce a two-parameter family of distributions on $(M, \mathcal{M})$, $q(dy, z, p)$, with corresponding Radon–Nikodým derivatives, $r(y, z, p)$. Theorem 2.1 extends to this case showing that the discrete-time filters converge to a continuous-time filter with observations process given by (2.6), but with a *time-varying*, signal-to-noise ratio of $H(\pi_t^H)$, where

$$H(p) = \int r_z'(y, 0, p) r_z(y, 0, p) q(dy, 0, p).$$

(Sufficient conditions are that (H1)–(H4) be true *uniformly in* p and that $H(p)$ be Lipschitz continuous.) In principle, we could optimize within classes of such feedback preprocessing operations according to cost functions of the form (4.1) but, once again, this is not easy, except in special cases. A simple variation of Proposition 4.1 shows that, if, for two feedback preprocessing operations h_1 and h_2, the limit signal-to-noise

ratio functions, $H_1(\cdot)$ and $H_2(\cdot)$, are such that $H_1(p) - H_2(p)$ is positive semidefinite for all p, then

$$J_\phi\left(\pi_t^{H_1(\cdot)}\right) \leq J_\phi\left(\pi_t^{H_2(\cdot)}\right)$$

for all $t \in [0, \infty)$ and all cost functions of the form (4.1). Thus, we may partially order the feedback preprocessing operations. If the components of the raw observations process are perturbed by independent noises and are preprocessed separately (in particular if the raw observations process is scalar), then the *instantaneous*, asymptotic signal-to-noise ratios, $H(p)$, of a feedback preprocessing operation will be completely ordered in the above sense. There is, therefore, nothing to be gained by the use of feedback in such cases; the optimal feedback preprocessing operation is clearly the feedback-less operation, $h(\cdot, p*)$, where $p*$ is the parameter that optimizes $H(p)$.

5. Examples. The optimal quantization techniques for observations perturbed by Gaussian noise, described above, were used with the following two examples. In both cases, the filters were based on raw observations, and 1- and 2-bit quantized observations were simulated. The quantization levels used throughout are those given in Table 4.1. An ergodic simulation technique was used involving the time averaging of variates obtained from the posterior distributions provided by the various filters. A number of the results were checked against results obtained from direct simulations. The filters were allowed to reach steady-state before data recording started, and so the cost functions concerned are relative to the invariant distributions of the filters. All simulations were run until all the sample standard deviations were less than 0.32% of their corresponding statistics. Error bars have been omitted from the graphs for the sake of clarity.

Example 5. Binary signal. In this example $(X_k^\delta \in \{-1, +1\}, k = 0, 1, \ldots)$ is a binary Markov process with the following matrix of transition probabilities,

$$Q^\delta = \begin{bmatrix} 1 - a\delta^2 & a\delta^2 \\ a\delta^2 & 1 - a\delta^2 \end{bmatrix},$$

and the raw observations are given by

$$Z_k^\delta = \delta X_k^\delta + \zeta_k, \tag{5.1}$$

where $(\zeta_k;\ k = 1, 2, \ldots)$ is an independently and identically distributed (i.i.d.) standard Gaussian sequence, independent of (X_k^δ).

Figures 5.1 and 5.2 show the steady-state error probabilities for the associated nonlinear filters based on the variously quantized observations, for values of the normalized switching rate, a, of 0.1 and 0.0001. The three disconnected points on the left-hand sides of the figures are the asymptotic error probabilities for the filters predicted by Theorem 2.1. These were calculated from the (known) steady-state distribution of the nonlinear filter for a continuous-time binary Markov process, given observations (2.6). (The values of H used for the quantized filters are those given in Table 4.1, and a value of 1 was used for the filter based on raw observations.) As the figures show, not only do the error probabilities approach the predicted limits, but they do so for fairly large values of δ. Thus, at least in this example, the use of quantization thresholds optimized with respect to limit cost functions appears to be justified.

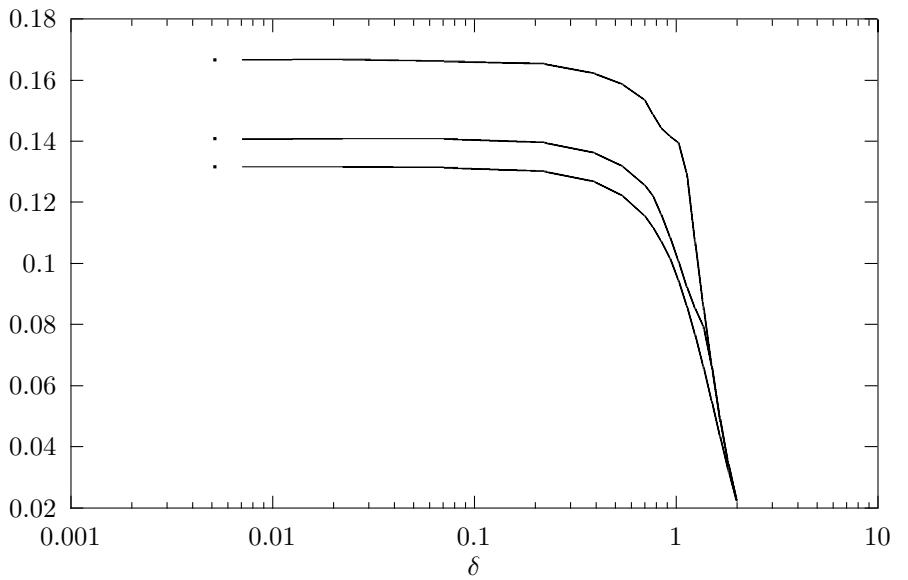


FIG. 5.1. *Filter error probabilities for Example* 1 *(*$a = 0.1$*). Bottom curve: raw observations; middle curve:* 2*-bit quantized observations; top curve:* 1*-bit quantized observations.*

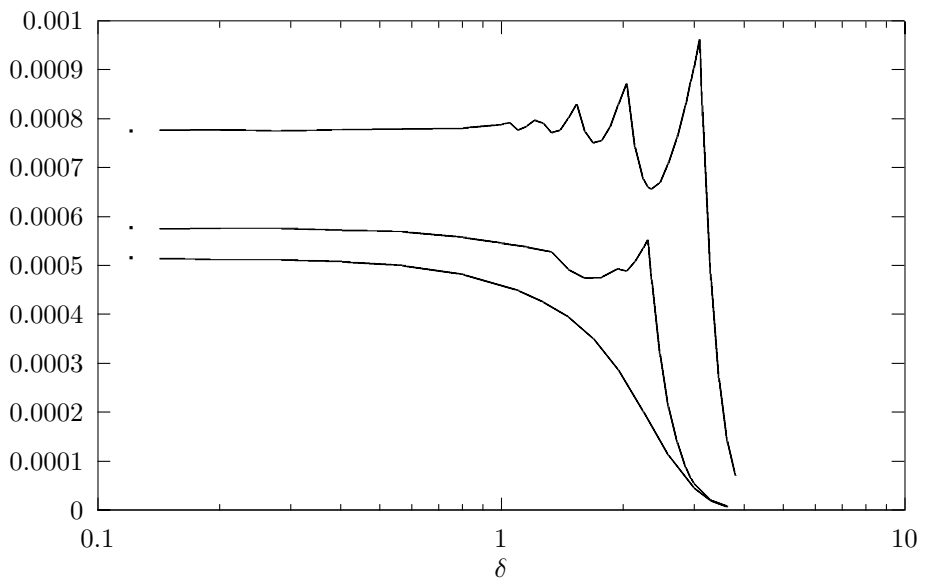


FIG. 5.2. *Filter error probabilities for Example* 1 *(*$a = 0.0001$*). Bottom curve: raw observations; middle curve:* 2*-bit quantized observations; top curve:* 1*-bit quantized observations.*

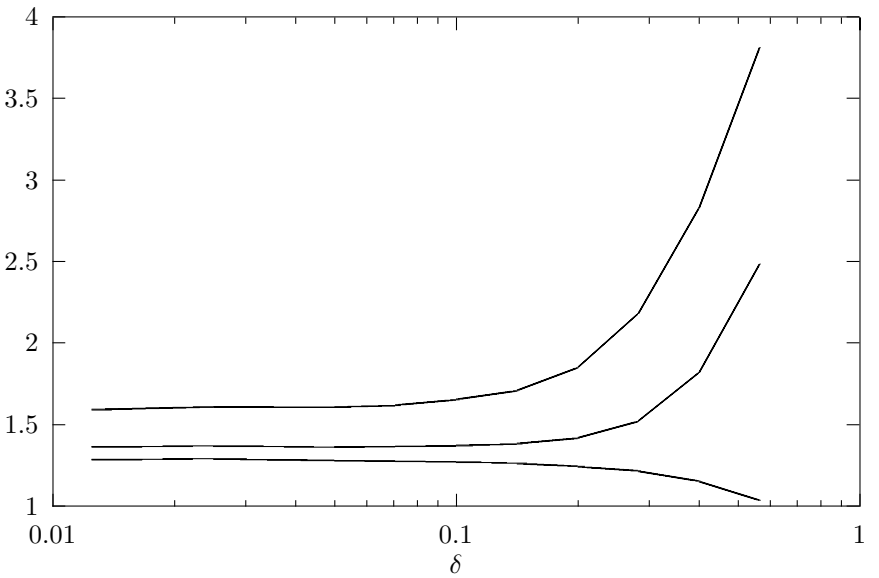


FIG. 5.3. *Mean-square filter errors for Example* 2 *(*$a = 1.0$*). Bottom curve: raw observations; middle curve:* 2*-bit quantized observations; top curve:* 1*-bit quantized observations.*

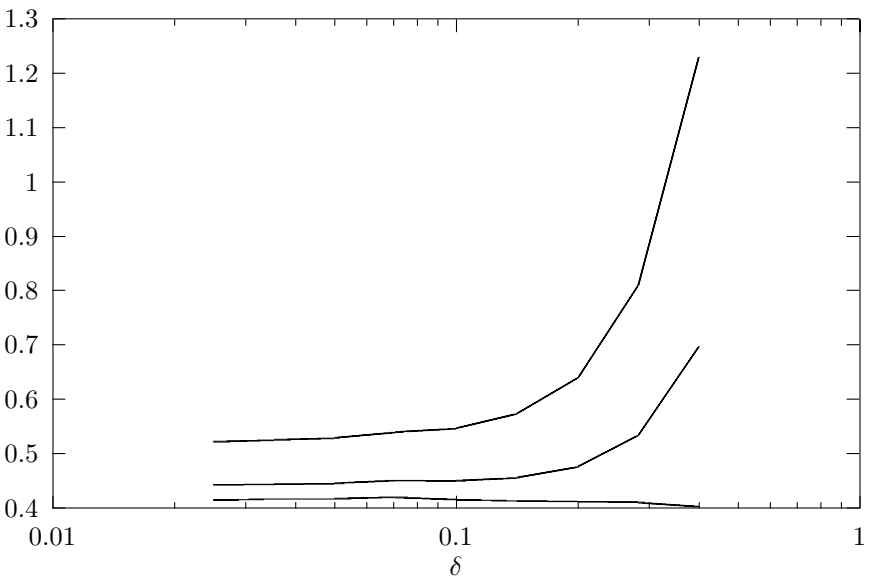


FIG. 5.4. *Mean-square filter errors for Example* 2 *(*$a = 0.1$*). Bottom curve: raw observations; middle curve:* 2*-bit quantized observations; top curve:* 1*-bit quantized observations.*

Example 6. Birth-death approximation to a diffusion. In this example the signal, $(X_k^\delta \in \{\pm 0.5, \pm 1.5, \pm 2.5 \ldots, \pm 9.5\},\ k = 0, 1, \ldots)$, is a 20-state birth-death process with transition probabilities

$$P^\delta(X_{k+1} = y \mid X_k = x) = \begin{cases} a\delta^2 & \text{if } y = x + 1 \text{ and } x \le 8.5, \\ & \text{or } y = x - 1 \text{ and } x \ge -8.5, \\ 1 - 2a\delta^2 & \text{if } -8.5 \le y = x \le 8.5, \\ 1 - a\delta^2 & \text{if } y = x = -9.5 \text{ or } 9.5, \\ 0 & \text{for all other } x, y, \end{cases}$$

and the raw observations are given by (5.1). The signal here can be interpreted as a discrete-state, weak approximation to a diffusion process with zero drift term and a diffusion coefficient that rapidly approaches zero for arguments beyond ±9 but is constant elsewhere.

Figures 5.3 and 5.4 show the mean-square error in the posterior mean of X obtained from the filters with the same three preprocessing operations as were used with Example 5, for values of a of 1.0 and 0.1. Once again, the graphs suggest that the weak limit is approached for fairly large values of δ.

REFERENCES

[1] J. F. Bennaton, *Discrete-time Galerkin approximations to the nonlinear filtering solution*, J. Math. Anal. Appl., 110 (1985), pp. 364–383.

[2] J. M. C. Clark, *The design of robust approximations to the stochastic differential equations of nonlinear filtering*, in Communication Systems and Random Process Theory, J. K. Skwirzynski, ed., NATO Advanced Study Institute Series, Sijthoff and Noordhoff, Alphen aan den Rijn, 1978, pp. 721–734.

[3] D. Crisan and T. Lyons, *Nonlinear filtering and measure-valued processes*, Probab. Theory Related Fields, 109 (1997), pp. 217–244.

[4] M. H. A. Davis, *A pathwise solution of the equations of nonlinear filtering*, Theory Probab. Appl., 27 (1983), pp. 167–175.

[5] G. B. Di Masi, M. Pratelli, and W. J. Runggaldier, *An approximation for the nonlinear filtering problem with error bound*, Stochastics, 14 (1985), pp. 247–271.

[6] S. N. Ethier and T. G. Kurtz, *Markov Processes: Characterisation and Convergence*, John Wiley, New York, 1986.

[7] M. Fujisaki, G. Kallianpur, and H. Kunita, *Stochastic differential equations for the nonlinear filtering problem*, Osaka J. Math., 9 (1972), pp. 19–40.

[8] G. Kallianpur and C. Striebel, *Estimation of stochastic systems: Arbitrary system process with additive white noise observation errors*, Ann. Math. Statist., 39 (1968), pp. 785–801.

[9] R. Khas'minskii and B. Lazareva, *On some filtration procedure for jump Markov process observed in white Gaussian noise*, Ann. Statist., 20 (1992), pp. 2153–2160.

[10] R. Khasminskii and O. Zeitouni, *Asymptotic filtering for finite state Markov chains*, Stochastic Process. Appl., 63 (1996), pp. 1–10.

[11] H. J. Kushner, *Dynamical equations for non-linear filtering*, J. Differential Equations, 3, (1967), pp. 179–190.

[12] H. J. Kushner, *A robust discrete-state approximation to the optimal nonlinear filter for a diffusion*, Stochastics, 3 (1979), pp. 75–83.

[13] H. J. Kushner, *Approximation and Weak Convergence Methods for Random Processes, With Application to Stochastic Systems Theory*, MIT Press, Cambridge, MA, 1984.

[14] F. Le Gland, *Splitting-up approximation for SPDEs and SDEs with application to nonlinear filtering*, in Stochastic Partial Differential Equations and Their Applications, Lecture Notes in Control and Inform. Sci. 176, Springer-Verlag, Berlin, 1992, pp. 177–187.

[15] R. S. Liptser and A. N. Shiryayev, *Statistics of Random Processes* 1—*General Theory*, Springer–Verlag, New York, 1977.

[16] N. J. Newton, *Observation sampling and quantisation for continuous-time estimators*, to appear in Stochastic Process. Appl.

[17] J. Picard, *Approximation of nonlinear filtering problems and order of convergence*, in Filtering and Control of Random Processes (E.N.S.T.-C.N.E.T. 1983), Lecture Notes in Control and Informat. Sci. 61, H. Korezlioglu, G. Mazziotto, and J. Szpirglas, eds., Springer–Verlag, New York, 1984, pp. 219–236.

[18] W. M. Wonham, *Some applications of stochastic differential equations to optimal nonlinear filtering*, SIAM J. Control, 2 (1964), pp. 347–369.

[19] M. Zakai, *On the optimal filtering of diffusion processes*, Z. Wahrscheinlichkeitstheorie und Verw Gebiete, 11 (1969), pp. 230–243.

SIAM J. CONTROL OPTIM.
Vol. 38, No. 2, pp. 503–537

DYNAMIC DOMAIN DECOMPOSITION IN APPROXIMATE AND EXACT BOUNDARY CONTROL IN PROBLEMS OF TRANSMISSION FOR WAVE EQUATIONS*

J. E. LAGNESE† AND G. LEUGERING‡

Abstract. This paper is concerned with dynamic domain decomposition for optimal boundary control and for approximate and exact boundary controllability of wave propagation in heterogeneous media. We consider a cost functional which penalizes the deviation of the final state of the solution of the global problem from a specified target state. For any fixed value of the penalty parameter, optimality conditions are derived for both the global optimal control problem and for local optimal control problems obtained by a domain decomposition and a saddle-point–type iteration. Convergence of the iterations to the solution of the global optimality system is established. We then pass to the limit in the iterations as the penalty parameter increases without bound and show that the limiting local iterations converge to the solution of the optimality system associated with the problem of finding the minimum norm control that drives the solution of the global problem to a specified target state.

Key words. domain decomposition, optimal control, controllability, saddle-point iteration

AMS subject classifications. 93B05, 93B40, 49N10, 65K10

PII. S0363012998333530

1. Introduction.

This paper explores optimal boundary final-value control as well as approximate and exact boundary controllability for scalar wave equations in heterogeneous media by means of domain decomposition. Indeed, we consider the situation where the material properties are piecewise constant within a material body which occupies a bounded domain $\Omega \in \mathbb{R}^d$ and hence deal with a *transmission problem*. It is assumed that the subdomains Ω_i, $i = 1, \dots, m$, within which the material properties are constant satisfy

$$\Omega_i \cap \Omega_j = \emptyset, \quad i \neq j, \quad \overline{\Omega}_i \subset \Omega, \quad i = 1, \dots, m-1, \quad \Omega_m = \Omega \setminus \bigcup_{i=1}^{m-1} \overline{\Omega}_i,$$

and that $\partial\Omega$, $\partial\Omega_i$, $i = 1, \dots, m$, are smooth. However, our results will hold for other decompositions $\{\Omega_i\}_{i=1}^m$ of Ω and for less regular boundaries if it is known that solutions of the global and local problems considered below have sufficient regularity.

The basic idea of domain decomposition in the present context is to handle the constant coefficient subdomain problems individually, with the impact on the environment of such a subdomain being modeled through some inhomogeneities, and to then assemble the controlled local processes into the global one by means of some iteration. Obviously, the communication between the local problems via transmission conditions has to be restored in the limit of the iteration. In this sense, we consider the present problem a paradigm and the procedure given in this paper as a general

*Received by the editors February 3, 1998; accepted for publication (in revised form) June 21, 1999; published electronically January 25, 2000.

http://www.siam.org/journals/sicon/38-2/33353.html

†Department of Mathematics, Georgetown University, Washington, DC 20057 (lagnese@archimedes.math.georgetown.edu). The research of this author was supported in part by NSF grant DMS-9972034.

‡Fakultät für Mathematik und Physik, Universität Bayreuth, D-95440 Bayreuth, Germany (leugering@uni-bayreuth.de). The research of this author was supported by Deutsche Forschungsgemeinschaft grants DFG Le 595/9-2 and DFG Le 595/12-1.

tool to handle more complex multilink flexible structures as discussed in Lagnese, Leugering, and Schmidt [19].

The motivation for the approach considered in this paper is twofold. On the one hand, controllability properties and optimal control laws (feedbacks, etc.) are more readily available at the local subdomain level. Consequently, there is good reason to believe that it may be possible to construct suitable control laws for the global, more complex, structure by iteration, even in situations where such control properties are not a priori known at the global level. On the other hand, and perhaps more importantly, the approach is motivated by its *numerical potential:* complex heterogeneous structures or media, after some discretization, inevitably lead to very large systems to be solved numerically. In typical applications, such as to flexible structures, structural elements of different kinds (strings, beams, membranes, plates and/or shells) are coupled through joints and interfaces. Such transmission problems are usually very difficult to handle numerically at the global level. It seems very natural, therefore, to decompose such structures into their canonical constituents (structural elements, subdomains) to discretize those and to run the computations *in parallel* with the communications between subdomains taking place once an iteration step is completed. The local problems can be treated numerically as in Glowinski and J.-L. Lions ([14], [15]).

There is extensive literature on domain decomposition methods for direct simulation of static or transient problems (see, for instance, [10], [30], [16] as examples of very recent expositions and [31] for textbook material in the context of numerical domain decomposition); comparatively little is known about such methods in the context of optimal control problems. Let us mention the work by Benamou [2], [3], [4], [5], [6], [7] and Benamou and Després [8], where elliptic, parabolic, and hyperbolic problems with constant coefficients in Ω are considered together with a cost functional which involves the entire state over space and time (in addition to the control). In these papers the authors use an extension of P. L. Lions's method [27], originally obtained for elliptic problems. Regarding transmission problems, the principle of extension of domain decomposition in optimal control of heterogeneous materials was already mentioned in [8] in connection with the Hemholtz equation. Let us also mention the work of Bamberger, Glowinski, and Tran [1], where a domain decomposition technique is introduced for the computation of the acoustic wave equation in which the bulk modulus and density fields are allowed to be discontinuous across an interface.

In fact, the nonoverlapping Schwarz alternating method in [27] has been shown by Glowinski and LeTallec [13] to be equivalent to an augmented Lagrangian saddle-point iteration. This analogy has been used by Bounaim [9] to extend the domain decomposition to elliptic optimal control problems. The idea pursued there is to add the cost function to the augmented Lagrangian interpretation of the underlying domain decomposition. The corresponding saddle-point problem, however, is solved by a gradient method rather than on the basis of the Uzawa-type iteration obtained after elimination of the Lagrange multipliers and artificial interface parameters. Very recently, J. L. Lions and Pironneau [25] (based on [26]) have considered a Lagrangian approach for both overlapping and nonoverlapping domain decompositions for elliptic and parabolic problems with a possible extension to hyperbolic problems. Their approach is similar to Bounaim's [9] method with the difference that they consider the original Lagrange functional to match the continuity at the artificial interfaces and use conjugate gradients. The augmented Lagrangian point of view was also used in [22] in order to devise a domain decomposition for mechanical networks of one-dimensional elements, where *multiple* nodes naturally appear. In that context the basic Lagrangian

approach, pursued in [25], could be utilized as well. Also in that work, the numerical implementation was discussed and simulations were presented. It should be mentioned that in all of the works [2], [3], [4], [5], [6], [7], [8], [22], the proof of convergence of the decomposition utilizes ideas inspired by Després [11].

To our best knowledge, this paper is the first to apply a dynamic domain decomposition method to state constrained optimal control problems. (For domain decomposition in optimal control of elliptic equations with control constraints, see [4].) Generalizations to more complex multiply connected structures as well as the numerical implementation of the algorithms given in this paper are presently under investigation. The only numerical study of dynamic domain decomposition in optimal control problems in dimension greater than one that is known to us is the preprint by Benamou [6] (see also [7]), where a scalar constant coefficient wave equation on a square is considered. In that work a *complete* decomposition down to the finite element level is used. Quadrilateral elements, a leap-frog scheme in time, and numerical integration leading to lumped masses are employed. In addition, various relaxations are used to speed up convergence.

Let us now formulate more precisely the problems to be studied. Set

$$\Gamma_{ij} = \partial\Omega_i \cap \partial\Omega_j = \Gamma_{ji}, \quad i \neq j, \quad \Gamma = \partial\Omega.$$

It is assumed that Γ_{ij} is either empty or has a nonempty interior. The exterior unit normal vector to Ω_i is denoted by ν_i. Introduce the Hilbert spaces

$$H = L^2(\Omega) = \prod_{i=1}^{m} L^2(\Omega_i),$$

$$\begin{aligned} V &= \{(\phi_1, \dots, \phi_m) : \phi_i \in H^1(\Omega_i),\ \phi_i = \phi_j \text{ on } \Gamma_{ij}\} \\ &= \{(\phi_1, \dots, \phi_m) : \phi_i = \phi|_{\Omega_i},\ \phi \in H^1(\Omega)\} \end{aligned}$$

with respective norms

$$\|(\phi_1, \dots, \phi_m)\| = \left(\sum_{i=1}^{m} \int_{\Omega_i} |\phi_i|^2 dx \right)^{1/2},$$

$$\|(\phi_1, \dots, \phi_m)\|_V = \left(\sum_{i=1}^{m} \int_{\Omega_i} (a_i |\nabla \phi_i|^2 + |\phi_i|^2) dx \right)^{1/2},$$

where $a_i > 0$. Let V' denote the dual space of V with respect to H, $(\cdot, \cdot)_{V'}$ the scalar product in the $V' - V$ duality, and A the Riesz isomorphism of V onto V'.

Let $T > 0$ and set

$$Q_i = \Omega_i \times (0, T), \quad \Sigma_{ij} = \Gamma_{ij} \times (0, T), \quad \Sigma = \Gamma \times (0, T).$$

Let $f \in L^2(\Sigma)$. We consider the following transmission problem: for $i = 1, \dots, m$,

$$\begin{aligned} y_{i,tt} - a_i \Delta y_i &= 0 \quad \text{in } Q_i, \\ a_m \frac{\partial y_m}{\partial \nu_m} &= f \quad \text{on } \Sigma, \\ y_i = y_j, \quad a_i \frac{\partial y_i}{\partial \nu_i} + a_j \frac{\partial y_j}{\partial \nu_j} &= 0 \quad \text{on } \Sigma_{ij}, \\ y_i(0) = y_{i,t}(0) &= 0 \quad \text{in } \Omega_i. \end{aligned} \tag{1.1}$$

For each $f \in \mathcal{U} := L^2(\Sigma)$, (1.1) has a unique solution $\mathbf{Y} := (y_1, \dots, y_m)$ with

$$\mathbf{Y} \in C([0,T]; H) \cap C^1([0,T]; V'),$$

and the map $f \mapsto (\mathbf{Y}, \mathbf{Y}_{,t}) : \mathcal{U} \mapsto C([0,T]; H \times V')$ is continuous. The solution of (1.1) may be interpreted in the sense of transposition (cf. Komornik [17, Theorem 4.7]):

$$(\mathbf{Y}_{,t}(t), \Phi(t))_{V'} - (\mathbf{Y}(t), \Phi_{,t}(t)) = \int_0^t \int_\Gamma f \phi_m \, d\Gamma \, dt, \quad \forall (\Phi^0, \Phi^1) \in V \times H,\ 0 \le t \le T, \tag{1.2}$$

where $\Phi := (\phi_1, \dots, \phi_m)$ is the solution of the problem

$$\begin{aligned} \phi_{i,tt} - a_i \Delta \phi_i &= 0 \quad \text{in } Q_i,\ i = 1, \dots, m, \\ a_m \frac{\partial \phi_m}{\partial \nu_m} &= 0 \quad \text{on } \Sigma, \\ \phi_i = \phi_j, \quad a_i \frac{\partial \phi_i}{\partial \nu_i} + a_j \frac{\partial \phi_j}{\partial \nu_j} &= 0 \quad \text{on } \Sigma_{ij}, \\ \Phi(0) = \Phi^0 \in V, \quad \Phi_{,t}(0) &= \Phi^1 \in H. \end{aligned} \tag{1.3}$$

One has $(\Phi, \Phi_{,t}) \in C([0,T]; V \times H)$ and

$$\|(\Phi, \Phi_{,t})\|_{L^\infty(0,T;V\times H)} \le C \|(\Phi^0, \Phi^1)\|_{V \times H}.$$

It is a consequence of Holmgren's theorem that the system (1.1) is *approximately controllable* for T sufficiently large, that is,

$$\mathcal{R}_T := \{(\mathbf{Y}(T), \mathbf{Y}_{,t}(T)) :\ f \in \mathcal{U}\} \quad \text{is dense in } H \times V'.$$

(This remains true for controls which vanish outside $\mathcal{O} \times (0,T)$, where $\mathcal{O}$ is a fixed, nonempty open set in Γ.) Further, (1.1) is known to be *exactly controllable* to $V \times H$, that is, $V \times H \subset \mathcal{R}_T$, provided certain conditions on the geometries of the regions Ω_i and the elastic parameters a_i are satisfied. For example, exact controllability holds if all of the a_i are the same. Exact controllability also holds if (i) $\Omega_1 = \omega_1$, $\Omega_{i+1} = \omega_{i+1} \backslash \overline{\omega}_i$ for $i = 1, \dots, m-1$, where ω_i are open sets such that $\overline{\omega_i} \subset \omega_{i+1}$, $i = 1, \dots, m-1$, $\omega_m = \Omega$; (ii) there is a point $x_0 \in \Omega_1$ such that $(x - x_0) \cdot \nu_i \ge 0$, $\forall x \in \partial\omega_i$, for $i = 1, \dots, m$; and (iii) $a_i \ge a_{i+1}$ for $i = 1, \dots, m-1$ (cf. [24, Chap. VI] and [18]).

Let $(\mathbf{Y}_T, \dot{\mathbf{Y}}_T) \in H \times V'$, and consider the optimal control problem

$$J(f) = \frac{1}{2} \int_\Sigma |f|^2 d\Sigma + \frac{k}{2} (\|\mathbf{Y}(T) - \mathbf{Y}_T\|^2 + \|\mathbf{Y}_{,t}(T) - \dot{\mathbf{Y}}_T\|^2_{V'}) \to \inf \tag{1.4}$$

subject to (1.1), where the infimum is taken over $\mathcal{U}$. Standard theory gives the existence of a unique minimizer f characterized by

$$0 = \int_\Sigma f g \, d\Sigma + k(\mathbf{U}(T), \mathbf{Y}(T) - \mathbf{Y}_T) + k(\mathbf{U}_{,t}(T), A^{-1}(\mathbf{Y}_{,t}(T) - \dot{\mathbf{Y}}_T))_{V'} \quad \forall g \in \mathcal{U},$$

where $\mathbf{U} = \mathbf{U}_g \in C([0,T];H) \cap C^1([0,T];V')$ is the solution of (1.1) with f replaced by g. Introduce the adjoint state $\mathbf{P} = (p_1, \dots, p_m)$ as the solution of the system

$$\begin{aligned} p_{i,tt} - a_i \Delta p_i &= 0 \quad \text{in } Q_i,\ i = 1, \dots, m, \\ a_m \frac{\partial p_m}{\partial \nu_m} &= 0 \quad \text{on } \Sigma, \\ p_i = p_j, \quad a_i \frac{\partial p_i}{\partial \nu_i} + a_j \frac{\partial p_j}{\partial \nu_j} &= 0 \quad \text{on } \Sigma_{ij}, \end{aligned} \tag{1.5}$$

$$\mathbf{P}(T) = kA^{-1}(\mathbf{Y}_{,t}(T) - \dot{\mathbf{Y}}_T), \quad \mathbf{P}_{,t}(T) = -k(\mathbf{Y}(T) - \mathbf{Y}_T). \tag{1.6}$$

For any $f \in \mathcal{U}$, (1.5), (1.6) has a unique solution with $(\mathbf{P}, \mathbf{P}_{,t}) \in C([0,T];V \times H)$. From (1.2) we have

$$0 = (\mathbf{U}_{,t}(T), \mathbf{P}(T))_{V'} - (\mathbf{U}(T), \mathbf{P}_{,t}(T)) - \int_\Sigma g p_m \, d\Sigma \quad \forall g \in \mathcal{U},$$

from which it follows that the optimal control is given by

$$f = -p_m|_\Sigma. \tag{1.7}$$

The optimality system is therefore (1.1), (1.5)–(1.7).

Remark 1.1. The solution of (1.1), (1.5)–(1.7) satisfies $(\mathbf{Y}, \mathbf{Y}_{,t}) \in C([0,T];V \times H)$. In fact, the optimal control $f = -p_m|_\Sigma \in C([0,T];H^{1/2}(\Gamma))$ so that the conclusion follows by localization and a result of Miyatake [28], taking into account the above assumptions on the regions Ω_i. It then follows that if $(\mathbf{Y}_T, \dot{\mathbf{Y}}_T) \in V \times H$, we have $(\mathbf{P}(T), \mathbf{P}_{,t}(T)) \in D(A) \times V$, where

$$\begin{aligned} D(A) = \{\Phi \in V \,|\, A\Phi \in H\} = \Big\{ &(\phi_1, \dots, \phi_m) \,|\, \phi_i \in H^2(\Omega_i), \\ &\phi_i = \phi_j,\ a_i \frac{\partial \phi_i}{\partial \nu_i} + a_j \frac{\partial \phi_j}{\partial \nu_j} = 0 \text{ on } \Gamma_{ij},\ \ a_m \frac{\partial \phi_m}{\partial \nu_m} = 0 \text{ on } \Gamma \Big\}. \end{aligned}$$

As a consequence, $p_m|_\Sigma \in C^1([0,T];H^{1/2}(\Gamma))$, so that $(\mathbf{Y}_{,t}, \mathbf{Y}_{,tt}) \in C([0,T];V \times H)$, and therefore $y_i \in C([0,T], H^2(\Omega_i))$, $i = 1, \dots, m$. In particular, the transmission conditions hold in the sense of traces on Γ_{ij}.

Remark 1.2. The first condition in (1.6) is the elliptic transmission problem

$$\begin{aligned} -a_i \Delta p_i(T) + p_i(T) &= k(y_{i,t}(T) - \dot{y}_{iT}) \quad \text{in } \Omega_i, \\ a_m \frac{\partial p_m(T)}{\partial \nu_m} &= 0 \quad \text{on } \Gamma, \\ p_i(T) = p_j(T), \quad a_i \frac{\partial p_i(T)}{\partial \nu_i} + a_j \frac{\partial p_j(T)}{\partial \nu_j} &= 0 \quad \text{on } \Gamma_{ij}. \end{aligned} \tag{1.8}$$

In (1.8),

$$y_{i,t}(T) - \dot{y}_{iT} := A_i[A^{-1}(\mathbf{Y}_{,t}(T) - \dot{\mathbf{Y}}_T)|_{\Omega_i}] \in (H^1(\Omega_i))',$$

where A_i is the Riesz isomorphism of $H^1(\Omega_i)$ onto its dual with the norm on $H^1(\Omega_i)$ given by

$$\|\phi\|_{H^1(\Omega_i)} := \left(\int_{\Omega_i} (a_i |\nabla \phi|^2 + |\phi|^2) dx \right)^{1/2}.$$

This is because

$$A(\phi_1, \ldots, \phi_m) = (A_1\phi_1, \ldots, A_m\phi_m), \ \ \forall (\phi_1, \ldots, \phi_m) \in V.$$

If $\mathbf{Y}(\cdot\,; k)$ denotes the solution of (1.1) corresponding to the optimal control (1.7), and if T is sufficiently large, it is easy to see that

$$(\mathbf{Y}(T; k), \mathbf{Y}_t(T; k)) \to (\mathbf{Y}_T, \dot{\mathbf{Y}}_T) \quad \text{in } H \times V' \text{ as } k \to \infty.$$

Therefore, a control that steers the solution to a neighborhood of $(\mathbf{Y}_T, \dot{\mathbf{Y}}_T)$ at time T may be obtained via the optimality system (1.1), (1.5)–(1.7) with k sufficiently large. On the other hand, if (1.1) is exactly controllable to $(\mathbf{Y}_T, \dot{\mathbf{Y}}_T) \in V \times H$ at time T, the minimum $L^2(\Sigma)$ norm control that steers the solution of (1.1) to $(\mathbf{Y}_T, \dot{\mathbf{Y}}_T)$ at time T is the solution of the optimal control problem

$$\inf_{f \in \mathcal{U}} \int_\Sigma |f|^2 d\Sigma$$

subject to (1.1) and the state constraint

$$\mathbf{Y}(T) = \mathbf{Y}_T, \quad \mathbf{Y}_{,t}(T) = \dot{\mathbf{Y}}_T.$$

It is well known that the optimality system for this problem is given by (1.1), (1.5),

$$\mathbf{P}(T) = \mathbf{P}_T, \quad \mathbf{P}_{,t}(T) = \dot{\mathbf{P}}_T, \tag{1.9}$$

where $(\mathbf{P}_T, \dot{\mathbf{P}}_T) \in H \times V'$ is the solution of the equation

$$((\mathbf{P}_T, \dot{\mathbf{P}}_T), (-\dot{\mathbf{Y}}_T, \mathbf{Y}_T))_{H \times V'} = \int_\Sigma |p_m|^2 d\Sigma, \tag{1.10}$$

with the optimal control again given by (1.7).

The main object of this paper is to present domain decomposition methods (DDMs) for both the approximate controllability and exact controllability problems. Each DDM is based on an *under* of the nonoverlapping Schwarz alternating algorithm and the introduction of skew-symmetric, Robin, iterative transmission conditions between the subdomains Ω_i that couple the direct and adjoint states in the optimality systems associated with the approximate, resp., the exact, controllability problem. The use of this type of decomposition, without relaxation, in problems of optimal control was first proposed by Benamou [2], [3], [4], [5], [6], [7]. The introduction of relaxation in conjunction with the nonoverlapping Schwarz alternating algorithm has been previously investigated in [11] for direct approximation and in [8] in conjunction with some optimal control problems related to the Helmholtz equation. The transmission conditions (3.3) below may be viewed as a generalization of those in [8] to optimal control problems involving penalization of the final state. Let us comment that the convergence of our DDMs *without relaxation* remains an open question.

For both the approximate and exact controllability problems, the corresponding DDM is a sequence of boundary value problems on the region Q_i. In the case of approximate controllability, these problems, denoted by $\{D_i^n(k)\}_{n=1}^\infty$, depend on the penalty parameter k. We denote the DDM for the exact controllability problem by $\{D_i^n(\infty)\}_{n=1}^\infty$. Let us denote by $y_i^n(\cdot\,; k)$, $p_i^n(\cdot\,; k)$ the solution of $D_i^n(k)$; by $y_i^n(\cdot\,; \infty)$,

$p_i^n(\cdot\,;\infty)$ the solution of $D_i^n(\infty)$; by $\{y_i(\cdot\,;k), p_i(\cdot\,;k)\}_{i=1}^m$ the solution of the optimality system (1.1), (1.5)–(1.7); and by $\{y_i(\cdot\,;\infty), p_i(\cdot\,;\infty)\}_{i=1}^m$ the solution of (1.1), (1.5), (1.7), (1.9), and (1.10). We wish to determine under what conditions the following diagram is valid in appropriate norms:

$$\begin{array}{ccc} (y_i^n(\cdot\,;k), p_i^n(\cdot\,;k)) & \xrightarrow{\;n\to\infty\;} & (y_i(\cdot\,;k), p_i(\cdot\,;k)) \\ {\scriptstyle k\to\infty}\big\downarrow & & {\scriptstyle k\to\infty}\big\downarrow \\ (y_i^n(\cdot\,;\infty), p_i^n(\cdot\,;\infty)) & \xrightarrow{\;n\to\infty\;} & (y_i(\cdot\,;\infty), p_i(\cdot\,;\infty)) \end{array} \tag{1.11}$$

Assuming that $(\mathbf{Y}_T, \dot{\mathbf{Y}}_T) \in V \times H$, it will be proved that the top limit is always true and that the remaining ones are also valid if (1.1) is exactly controllable to $V \times H$.

As is to be expected from the works of Benamou and of Leugering cited above, for fixed n, the boundary value problem $D_i^n(k)$ is in fact the optimality system for a certain local optimal control problem on the region Q_i. (The same is true for $D_i^n(\infty)$.) The local optimal control problem associated with $D_i^n(k)$ is introduced in the next section. In section 3 the convergence $(y_i^n(\cdot\,;k), p_i^n(\cdot\,;k)) \to (y_i(\cdot\,;k), p_i(\cdot\,;k))$ as $n \to \infty$ is established. In section 4 it is proved that $(y_i(\cdot\,;k), p_i(\cdot\,k)) \to (y_i(\cdot\,;\infty), p_i(\cdot\,\infty))$ as $k \to \infty$, while in section 5 it is shown that $(y_i^n(\cdot\,;k), p_i^n(\cdot\,;k)) \to (y_i^n(\cdot\,;\infty), p_i^n(\cdot\,;\infty))$ as $k \to \infty$. Finally, in the last section it is established that $(y_i^n(\cdot\,;\infty), p_i^n(\cdot\,;\infty)) \to (y_i(\cdot\,;\infty), p_i(\cdot\,\infty))$ as $n \to \infty$.

2. The local optimal control problems. For $i = 1, \dots, m$ we set

$$\gamma_i = \bigcup_{j:\Gamma_{ij}\neq\emptyset} \Gamma_{ij}, \quad \Gamma_i = \begin{cases} \emptyset, & i \neq m, \\ \Gamma, & i = m, \end{cases} \quad S_i = \gamma_i \times (0,T), \quad \Sigma_i = \Gamma_i \times (0,T).$$

Then $\partial\Omega_i = \gamma_i \cup \Gamma_i$. We also set $H_i = L^2(\Omega_i)$, $V_i = H^1(\Omega_i)$ endowed with the norm

$$\|\phi\|_{V_i} = \left(\int_{\Omega_i} (a_i |\nabla\phi|^2 + |\phi|^2) dx + \alpha \int_{\gamma_i} |\phi|^2 d\Gamma \right)^{1/2}, \quad \alpha > 0,$$

and denote by V_i' the dual space of V_i with respect to H_i. Of course, V_i' and $(H^1(\Omega_i))'$ are the same as sets and have equivalent norms, but the Riesz isomorphism $\mathcal{A}_i$ of V_i onto V_i' is not the same as the Riesz isomorphism A_i defined above. (However, they are connected through the bounded invertible mapping B_i on V_i defined by $(\phi,\psi)_{H^1(\Omega_i)} = (B_i\phi, \psi)_{V_i}\ \forall \phi, \psi \in V_i$.)

Set $y_{iT} = \mathbf{Y}_T|_{\Omega_i} \in H_i$, $\dot{y}_{iT} = A_i(A^{-1}\dot{\mathbf{Y}}_T|_{\Omega_i}) \in V_i'$ (see Remark 1.2). Suppose that λ_i, μ_i are given in $L^2(S_i)$ and τ_i is given in $L^2(\gamma_i)$. We define $\sigma_i \in V_i'$ by

$$(\sigma_i, \phi)_{V_i'} = \int_{\gamma_i} \tau_i \phi \, d\Gamma \quad \forall \phi \in V_i,$$

and consider the optimal control problems

$$\begin{aligned} J_i(f_1, f_2) = {} & \frac{1}{2} \int_{\Sigma_i} |f_2|^2 d\Sigma + \frac{1}{2\beta} \int_{S_i} (|f_1|^2 + |\beta z_i + \mu_i|^2) d\Sigma \\ & + \frac{k}{2} (\|z_i(T) - y_{iT}\|_{H_i}^2 + \|z_{i,t}(T) - \dot{y}_{iT} + k^{-1}\sigma_i\|_{V_i'}^2) \to \inf, \end{aligned}$$

where $\beta > 0$, subject to

$$
\begin{aligned}
z_{i,tt} - a_i \Delta z_i &= 0 \quad \text{in } Q_i, \\
a_i \frac{\partial z_i}{\partial \nu_i} &= f_1 + \lambda_i \quad \text{on } S_i, \\
a_i \frac{\partial z_i}{\partial \nu_i} &= f_2 \quad \text{on } \Sigma_i, \\
z_i(0) = z_{i,t}(0) &= 0 \quad \text{in } \Omega_i,
\end{aligned} \tag{2.1}
$$

where the infimum is taken over $f_1 \in L^2(S_i)$ and $f_2 \in L^2(\Sigma_i)$. This optimal control problem is well posed since the solution of (2.1) satisfies

$$z_i|_{S_i} \in L^2(S_i), \quad z_i|_{\Sigma_i} \in L^2(\Sigma_i),$$

(see Lasiecka–Triggiani [20]). In fact, from [20] we have $(z_i, z_{i,t}) \in C([0,T]; H^\alpha(\Omega_i) \times H^{\alpha-1}(\Omega_i))$, $z_i|_{\Sigma_{ij}} \in H^{2\alpha-1}(\Sigma_{ij})$, $z_i|_{\Sigma_i} \in H^{2\alpha-1}(\Sigma_i)$, where $\alpha = 3/5 - \varepsilon \ \forall \varepsilon > 0$, and

$$
\begin{aligned}
&\|z_i\|_{H^{2\alpha-1}(\Sigma_i)} + \sum_{j:\Gamma_{ij} \neq \emptyset} \|z_i\|_{H^{2\alpha-1}(\Sigma_{ij})} + \|(z_i, z_{i,t})\|_{L^\infty(0,T;H^\alpha(\Omega_i)\times H^{\alpha-1}(\Omega_i))} \\
&\quad \le C\big(\|f_1 + \lambda_i\|_{L^2(S_i)} + \|f_2\|_{L^2(\Sigma_i)}\big).
\end{aligned}
$$

From Tataru [33] we have the even stronger regularity result

$$z_i \in H^{2/3}(Q_i), \quad z_i|_{\Sigma_{ij}} \in H^{1/3}(\Sigma_{ij}), \quad z_i|_{\Sigma_i} \in H^{1/3}(\Sigma_i).$$

The unique optimal solution f_1, f_2 of the above problem is characterized by the variational equation

$$
\begin{aligned}
0 = &\int_{\Sigma_i} f_2 g_2 \, d\Sigma + \int_{S_i} [\beta^{-1} f_1 g_1 + (\beta z_i + \mu_i) u_i] d\Sigma \\
&+ k[(u_i(T), z_i(T) - y_{iT})_{H_i} + (u_{i,t}(T), \mathcal{A}_i^{-1}(z_{i,t}(T) - \dot{y}_{iT} + k^{-1}\sigma_i))_{V_i'}] \\
&\quad \forall g_1 \in L^2(S_i),\ g_2 \in L^2(\Sigma_i),
\end{aligned}
$$

where $\mathcal{A}_i$ is the Riesz isomorphism of V_i onto V_i' and where u_i is the solution of

$$
\begin{aligned}
u_{i,tt} - a_i \Delta u_i &= 0 \quad \text{in } Q_i, \\
a_i \frac{\partial u_i}{\partial \nu_i} &= g_1 \quad \text{on } S_i, \\
a_i \frac{\partial u_i}{\partial \nu_i} &= g_2 \quad \text{on } \Sigma_i, \\
u_i(0) = u_{i,t}(0) &= 0 \quad \text{in } \Omega_i.
\end{aligned}
$$

Introduce the adjoint state as the solution of

$$
\begin{aligned}
q_{i,tt} - a_i \Delta q_i &= 0 \quad \text{in } Q_i, \\
a_i \frac{\partial q_i}{\partial \nu_i} &= \beta z_i + \mu_i \quad \text{on } S_i, \\
a_i \frac{\partial q_i}{\partial \nu_i} &= 0 \quad \text{on } \Sigma_i, \\
q_i(T) = k\mathcal{A}_i^{-1}(z_{i,t}(T) - \dot{y}_{iT} + k^{-1}\sigma_i) &\in V_i, \\
q_{i,t}(T) = -k(z_i(T) - y_{iT}) &\in H_i.
\end{aligned} \tag{2.2}
$$

The solutions z_i, u_i, q_i all possess the same regularity and after an integration by parts (which can be justified by virtue of the regularity of u_i and q_i) we obtain

$$0 = (u_{i,t}(T), q_i(T))_{V_i'} - (u_i(T), q_{i,t}(T))_{H_i} - \int_{\Sigma_i} q_i g_2 d\Sigma - \int_{S_i} [q_i g_1 - (\beta z_i + \mu_i) u_i] d\Sigma, \quad \forall g_1 \in L^2(S_i),\ g_2 \in L^2(\Sigma_i). \tag{2.3}$$

It follows that

$$f_1 = -\beta q_i|_{S_i}, \quad f_2 = -q_i|_{\Sigma_i}.$$

The optimality system is therefore (2.2) and

$$\begin{aligned} z_{i,tt} - a_i \Delta z_i &= 0 \quad \text{in } Q_i, \\ a_i \frac{\partial z_i}{\partial \nu_i} + \beta q_i &= \lambda_i \quad \text{on } S_i, \\ a_i \frac{\partial z_i}{\partial \nu_i} + q_i &= 0 \quad \text{on } \Sigma_i, \\ z_i(0) = z_{i,t}(0) &= 0 \quad \text{in } \Omega_i. \end{aligned} \tag{2.4}$$

Remark 2.1. Note that $q_i(T)$ is the solution of the elliptic boundary value problem

$$\begin{aligned} -a_i \Delta q_i(T) + q_i(T) &= k(z_{i,t}(T) - \dot{y}_{iT}) \quad \text{in } \Omega_i, \\ a_i \frac{\partial q_i(T)}{\partial \nu_i} + \alpha q_i(T) &= \tau_i \quad \text{on } \gamma_i, \\ a_i \frac{\partial q_i(T)}{\partial \nu_i} &= 0 \quad \text{on } \Gamma_i. \end{aligned}$$

For purposes of studying convergence of the domain decomposition in the next section, we introduce a space $\mathcal{X}$ and a bounded linear operator on $\mathcal{X}$ as follows:

$$\mathcal{X} := \prod_{i=1}^{m} L^2(S_i) \times L^2(S_i) \times L^2(\gamma_i)$$

with norm $\|\cdot\|_{\mathcal{X}}$ given by

$$\|X\|_{\mathcal{X}}^2 = \sum_{i=1}^{m} \left\{ \frac{1}{\beta} \int_{S_i} (|\lambda_i|^2 + |\mu_i|^2) d\Sigma + \frac{1}{\alpha k} \int_{\gamma_i} |\tau_i|^2 d\Gamma \right\},$$

where $X = \{(\lambda_i, \mu_i, \tau_i) : i = 1, \dots, m\}$. Now let $X \in \mathcal{X}$ and define a linear mapping $\mathcal{T} : \mathcal{X} \mapsto \mathcal{X}$ as follows: let (z_i, q_i) be the solutions of the local optimality systems (2.2), (2.4) corresponding to X, where $y_{iT} = \dot{y}_{iT} = 0$, $i = 1, \dots, m$ (this is done for purposes of the proof of Theorem 3.1 below). Set

$$(\mathcal{T}X)_{ij} = \left(\left(-a_j \frac{\partial z_j}{\partial \nu_j} + \beta q_j \right) |_{\Sigma_{ij}}, \left(-a_j \frac{\partial q_j}{\partial \nu_j} - \beta z_j \right) |_{\Sigma_{ij}}, \left(-a_j \frac{\partial q_j(T)}{\partial \nu_j} + \alpha q_j(T) \right) |_{\gamma_{ij}} \right),$$

$$(\mathcal{T}X)_i = \{(\mathcal{T}X)_{ij} : j : \Gamma_{ij} \neq \emptyset\}, \quad \mathcal{T}X = \{(\mathcal{T}X)_i : i = 1, \dots, m\}. \tag{2.5}$$

Note that X is a fixed point of $\mathcal{T}$ if and only if $\{(z_i, q_i) : i = 1, \dots, m\}$ is a solution of the global optimality system (1.1), (1.5)–(1.7). It follows that $\mathcal{T}$ has a unique fixed point. Since the solution of the optimal control problem (1.1), (1.4) corresponding to $\mathbf{Y}_T = \dot{\mathbf{Y}}_T = 0$ is clearly $f = 0$, it follows that this fixed point is $X = 0$. The following result shows that $\mathcal{T}$ is nonexpansive.

LEMMA 2.2. *For any* $X \in \mathcal{X}$,

$$\|\mathcal{T}X\|_{\mathcal{X}}^2 = \|X\|_{\mathcal{X}}^2 - \frac{4}{k}\sum_{i=1}^{m}\big[\|q_i(T)\|_{H^1(\Omega_i)}^2 + \|q_{i,t}(T)\|_{L^2(\Omega_i)}^2\big] - 4\int_\Sigma |q_m|^2 d\Sigma.$$

Proof. One has

$$\begin{aligned}
\|X\|_{\mathcal{X}}^2 &= \sum_{i=1}^{m}\bigg\{\frac{1}{\beta}\int_{S_i}\bigg[\bigg|a_i\frac{\partial z_i}{\partial \nu_i} + \beta q_i\bigg|^2 + \bigg|a_i\frac{\partial q_i}{\partial \nu_i} - \beta z_i\bigg|^2\bigg]\,d\Sigma \\
&\qquad + \frac{1}{\alpha k}\int_{\gamma_i}\bigg|a_i\frac{\partial q_i(T)}{\partial \nu_i} + \alpha q_i(T)\bigg|^2 d\Gamma\bigg\} \\
&= \sum_{i=1}^{m}\frac{1}{\beta}\bigg\{\int_{S_i}\bigg[\bigg|a_i\frac{\partial z_i}{\partial \nu_i}\bigg|^2 + \bigg|a_i\frac{\partial q_i}{\partial \nu_i}\bigg|^2 + \beta^2(|z_i|^2 + |q_i|^2) \\
&\qquad + 2\beta\bigg(q_i a_i\frac{\partial z_i}{\partial \nu_i} - z_i a_i\frac{\partial q_i}{\partial \nu_i}\bigg)\bigg]\,d\Sigma \\
&\qquad + \frac{1}{\alpha k}\int_{\gamma_i}\bigg[\bigg|a_i\frac{\partial q_i(T)}{\partial \nu_i}\bigg|^2 + \alpha^2|q_i(T)|^2 + 2\alpha q_i(T)a_i\frac{\partial q_i(T)}{\partial \nu_i}\bigg]d\Gamma\bigg\},
\end{aligned}$$

$$\begin{aligned}
\|\mathcal{T}X\|_{\mathcal{X}}^2 = \sum_{i=1}^{m}\sum_{j:\Gamma_{ij}\neq\emptyset}\bigg\{&\frac{1}{\beta}\int_{\Sigma_{ij}}\bigg[\bigg|-a_j\frac{\partial z_j}{\partial \nu_j} + \beta q_j\bigg|^2 + \bigg|a_j\frac{\partial q_j}{\partial \nu_j} + \beta z_j\bigg|^2\bigg]\,d\Sigma \\
&+ \frac{1}{\alpha k}\int_{\Gamma_{ij}}\bigg|-a_j\frac{\partial q_j(T)}{\partial \nu_j} + \alpha q_j(T)\bigg|^2 d\Gamma\bigg\}.
\end{aligned}$$

Since $\sum_{i=1}^m\sum_{j:\Gamma_{ij}\neq\emptyset} = \sum_{j=1}^m\sum_{i:\Gamma_{ji}\neq\emptyset}$, the last equation may be written

$$\begin{aligned}
\|\mathcal{T}X\|_{\mathcal{X}}^2 = \sum_{j=1}^{m}\bigg\{&\frac{1}{\beta}\int_{\Sigma_j}\bigg[\bigg|-a_j\frac{\partial z_j}{\partial \nu_j} + \beta q_j\bigg|^2 + \bigg|a_j\frac{\partial q_j}{\partial \nu_j} + \beta z_j\bigg|^2\bigg]\,d\Sigma \\
&+ \frac{1}{\alpha k}\int_{\gamma_j}\bigg|-a_j\frac{\partial q_j(T)}{\partial \nu_j} + \alpha q_j(T)\bigg|^2 d\Gamma\bigg\}.
\end{aligned}$$

Therefore

$$\begin{aligned}
(2.6)\quad \|\mathcal{T}X\|_{\mathcal{X}}^2 - \|X\|_{\mathcal{X}}^2 = -4\sum_{i=1}^{m}\bigg\{&\int_{S_i}\bigg(q_i a_i\frac{\partial z_i}{\partial \nu_i} - z_i a_i\frac{\partial q_i}{\partial \nu_i}\bigg)d\Sigma \\
&+\frac{1}{k}\int_{\gamma_i} q_i(T)a_i\frac{\partial q_i(T)}{\partial \nu_i}\bigg\}d\Gamma.
\end{aligned}$$

From (2.3) we have

$$\begin{aligned}
(2.7)\qquad 0 = (z_{i,t}(T), q_i(T))_{V_i'} - (z_i(T), q_{i,t}(T))_{H_i} - \int_{\Sigma_i} q_i a_i\frac{\partial z_i}{\partial \nu_i}\,d\Sigma \\
- \int_{S_i}\bigg(q_i a_i\frac{\partial z_i}{\partial \nu_i} - z_i a_i\frac{\partial q_i}{\partial \nu_i}\bigg)d\Sigma.
\end{aligned}$$

From (2.2),

$$(z_{i,t}(T), q_i(T))_{V_i'} - (z_i(T), q_{i,t}(T))_{H_i} = \frac{1}{k}\big(\|q_i(T)\|^2_{H^1(\Omega_i)} + \|q_{i,t}(T)\|^2_{L^2(\Omega_i)}\big) - \frac{1}{k}\int_{\gamma_i} q_i(T) a_i \frac{\partial q_i(T)}{\partial \nu_i}\, d\Gamma. \tag{2.8}$$

Lemma 2.2 follows from (2.6)–(2.8). □

Lemma 2.2. can be regarded as an extension to optimality systems of the isometry Lemma 2.4 of Després [12]. Later on we shall use the following result of Opial [29, Theorem 3].

PROPOSITION 2.3. *Let $\mathcal{C}$ be a closed convex set in a uniformly convex Banach space $\mathcal{X}$ having a weakly continuous duality mapping, and let $\mathcal{T} : \mathcal{C} \mapsto \mathcal{C}$ be a nonexpansive mapping with at least one fixed point. Set $\mathcal{T}_\epsilon = \epsilon I + (1-\epsilon)\mathcal{T}$. Then, for any $X \in \mathcal{C}$ and any $\epsilon \in (0,1)$, the sequence of successive approximations $\{\mathcal{T}_\epsilon^n X\}$ is weakly convergent to a fixed point of $\mathcal{T}$.*

3. Domain decomposition. For $i = 1, \dots, m$ and for $n = 0, 1, \dots$, consider the problem

$$\begin{aligned} y_{i,tt}^{n+1} - a_i \Delta y_i^{n+1} &= 0 \quad \text{in } Q_i, \\ a_i \frac{\partial y_i^{n+1}}{\partial \nu_i} + \beta p_i^{n+1} &= \lambda_{ij}^n \quad \text{on } \Sigma_{ij}, \\ a_i \frac{\partial y_i^{n+1}}{\partial \nu_i} + p_i^{n+1} &= 0 \quad \text{on } \Sigma_i, \\ y_i^{n+1}(0) = y_{i,t}^{n+1}(0) &= 0, \end{aligned} \tag{3.1}$$

$$\begin{aligned} p_{i,tt}^{n+1} - a_i \Delta p_i^{n+1} &= 0 \quad \text{in } Q_i, \\ a_i \frac{\partial p_i^{n+1}}{\partial \nu_i} - \beta y_i^{n+1} &= \mu_{ij}^n \quad \text{on } \Sigma_{ij}, \\ a_i \frac{\partial p_i^{n+1}}{\partial \nu_i} &= 0 \quad \text{on } \Sigma_i, \\ p_i^{n+1}(T) = k\mathcal{A}_i^{-1}(y_{i,t}^{n+1}(T) - \dot{y}_{iT} + k^{-1}\sigma_i^n), \\ p_{i,t}^{n+1}(T) &= -k(y_i^{n+1}(T) - y_{iT}), \end{aligned} \tag{3.2}$$

where λ_{ij}^0, $\mu_{ij}^0 \in L^2(\Sigma_{ij})$ and, for $n = 1, 2, \dots$ and fixed $\epsilon \in [0, 1)$,

$$\begin{aligned} \lambda_{ij}^n &= (1-\epsilon)\left(-a_j \frac{\partial y_j^n}{\partial \nu_j} + \beta p_j^n\right)|_{\Sigma_{ij}} + \epsilon\left(a_i \frac{\partial y_i^n}{\partial \nu_i} + \beta p_i^n\right)|_{\Sigma_{ij}}, \\ \mu_{ij}^n &= (1-\epsilon)\left(-a_j \frac{\partial p_j^n}{\partial \nu_j} - \beta y_j^n\right)|_{\Sigma_{ij}} + \epsilon\left(a_i \frac{\partial p_i^n}{\partial \nu_i} - \beta y_i^n\right)|_{\Sigma_{ij}}, \end{aligned} \tag{3.3}$$

and where $\sigma_i^n \in V_i'$ is defined by

$$(\sigma_i^n, \phi)_{V_i'} = \sum_{j:\Gamma_{ij} \neq \emptyset} \int_{\Gamma_{ij}} \tau_{ij}^n \phi \, d\Gamma \quad \forall \phi \in V_i, \tag{3.4}$$

with $\tau_{ij}^0 \in L^2(\Gamma_{ij})$ arbitrary and for $n = 1, 2, \ldots,$

$$\tau_{ij}^n = (1-\epsilon)\left(-a_j \frac{\partial p_j^n(T)}{\partial \nu_j} + \alpha p_j^n(T)\right)|_{\Gamma_{ij}} + \epsilon\left(a_i \frac{\partial p_i^n(T)}{\partial \nu_i} + \alpha p_i^n(T)\right)|_{\Gamma_{ij}}. \tag{3.5}$$

Note that the iteration on $p_i^{n+1}(T)$ is a relaxation of the nonoverlapping Schwarz alternating method applied to the global elliptic transmission problem (1.8) for the final state $p(T)$. Written as an elliptic boundary value problem, it is

$$\begin{aligned} -a_i \Delta p_i^{n+1}(T) + p_i^{n+1}(T) &= k(y_{i,t}^{n+1}(T) - \dot{y}_{iT}) \quad \text{in } \Omega_i, \\ a_i \frac{\partial p_i^{n+1}(T)}{\partial \nu_i} + \alpha p_i^{n+1}(T) &= \tau_{ij}^n \quad \text{on } \Gamma_{ij}, \\ a_i \frac{\partial p_i^{n+1}(T)}{\partial \nu_i} &= 0 \quad \text{on } \Gamma_i \end{aligned} \tag{3.6}$$

(see Remark 2.1). Note also that (3.1), (3.2) is the optimality system for the local optimal control problem on Q_i considered in the previous section with λ_i, μ_i and τ_i replaced by $(\lambda_{ij}^n)_{j:\Gamma_{ij}\neq\emptyset}$, $(\mu_{ij}^n)_{j:\Gamma_{ij}\neq\emptyset}$, and $(\tau_{ij}^n)_{j:\Gamma_{ij}\neq\emptyset}$, resp. The parameter ϵ is a relaxation parameter; when $\epsilon = 0$ the iterations at the interfaces Σ_{ij} are exactly those introduced by Benamou [2], [3], [4], [5], [6], [7].

If λ_{ij}^n, $\mu_{ij}^n \in L^2(\Sigma_{ij})$, $\tau_{ij}^n \in L^2(\Gamma_{ij})$, $i = 1, \ldots, m$, $j : \Gamma_{ij} \neq \emptyset$, then because of the regularity possessed by the solutions y_i^{n+1}, p_i^{n+1}, $i = 1, \ldots, m$, we have

$$\lambda_{ij}^{n+1} \in L^2(\Sigma_{ij}), \quad \mu_{ij}^{n+1} \in L^2(\Sigma_{ij}), \quad \tau_{ij}^{n+1} \in L^2(\Gamma_{ij}).$$

If, therefore, $\lambda_{ij}^0 \in L^2(\Sigma_{ij})$, $\mu_{ij}^0 \in L^2(\Sigma_{ij})$, $\tau_{ij}^0 \in L^2(\Gamma_{ij})$, then (3.1), (3.2) is well set for each $n = 0, 1, \ldots$.

Consider now the global optimality system (1.1), (1.5)–(1.7). If we set

$$\begin{aligned} \lambda_{ij} &= (1-\epsilon)\left(-a_j \frac{\partial y_j}{\partial \nu_j} + \beta p_j\right)|_{\Sigma_{ij}} + \epsilon\left(a_i \frac{\partial y_i}{\partial \nu_i} + \beta p_i\right)|_{\Sigma_{ij}}, \\ \mu_{ij} &= (1-\epsilon)\left(-a_j \frac{\partial p_j}{\partial \nu_j} - \beta y_j\right)|_{\Sigma_{ij}} + \epsilon\left(a_i \frac{\partial p_i}{\partial \nu_i} - \beta y_i\right)|_{\Sigma_{ij}}, \\ \tau_{ij} &= (1-\epsilon)\left(-a_j \frac{\partial p_j(T)}{\partial \nu_j} + \alpha p_j(T)\right)|_{\Gamma_{ij}} + \epsilon\left(a_i \frac{\partial p_i(T)}{\partial \nu_i} + \alpha p_i(T)\right)|_{\Gamma_{ij}}, \\ (\sigma_i, \phi)_{V_i'} &= \sum_{j:\Gamma_{ij}\neq\emptyset} \int_{\Gamma_{ij}} \tau_{ij}\phi \, d\Gamma \quad \forall \phi \in V_i, \end{aligned}$$

then it is seen that y_i, p_i is formally a solution of (3.1), (3.2) with $\lambda_{ij}^n, \mu_{ij}^n, \tau_{ij}^n$ replaced by $\lambda_{ij}, \mu_{ij}, \tau_{ij}$, resp., and y_i, p_i is an actual solution if it is known that

$$\lambda_{ij} \in L^2(\Sigma_{ij}), \quad \mu_{ij} \in L^2(\Sigma_{ij}), \quad \tau_{ij} \in L^2(\Gamma_{ij}).$$

This amounts to requiring that

$$\frac{\partial y_j}{\partial \nu_j}|_{\Sigma_{ij}} \in L^2(\Sigma_{ij}), \quad \frac{\partial p_j}{\partial \nu_j}|_{\Sigma_{ij}} \in L^2(\Sigma_{ij}), \quad \frac{\partial p_j(T)}{\partial \nu_j}|_{\Gamma_{ij}} \in L^2(\Gamma_{ij}), \tag{3.7}$$

which will hold if it is assumed that $(\mathbf{Y}_T, \dot{\mathbf{Y}}_T) \in V \times H$ (see Remark 1.1.).

We will prove the following convergence result.

THEOREM 3.1. *Assume that* $(\mathbf{Y}_T, \dot{\mathbf{Y}}_T) \in V \times H$ *and* $\epsilon \in (0,1)$. *Let* $\mathbf{Y} = (y_1, \dots, y_m)$, $\mathbf{P} = (p_1, \dots, p_m)$ *be the solution of the global optimality system* (1.1), (1.5)–(1.7), *and* y_i^n, p_i^n *be the solutions of the local optimality systems* (3.1), (3.2), $i = 1, \dots, m$. *Set* $\mathbf{Y}^n = (y_1^n, \dots, y_m^n)$, $\mathbf{P}^n = (p_1^n, \dots, p_m^n)$. *Then as* $n \to \infty$,

$$\mathbf{Y}^n \to \mathbf{Y}, \quad \mathbf{P}^n \to \mathbf{P} \quad \text{in } C\Big([0,T]; \prod_{i=1}^m L^2(\Omega_i)\Big),$$
$$\mathbf{Y}^n_{,t} \to \hat{\mathbf{Y}}, \quad \mathbf{P}^n_{,t} \to \hat{\mathbf{P}} \quad \text{in } C\Big([0,T]; \prod_{i=1}^m (H^1(\Omega_i))'\Big),$$

where $\hat{\mathbf{Y}}(t)|_V = \mathbf{Y}_{,t}(t)$, $\hat{\mathbf{P}}(t)|_V = \mathbf{P}_{,t}(t)$, $0 \le t \le T$,

$$p_m^n|_\Sigma \to p_m|_\Sigma \quad \text{strongly in } L^2(\Sigma),$$
$$y_i^n|_{S_i} \to y_i|_{S_i}, \quad p_i^n|_{S_i} \to p_i|_{S_i} \quad \text{strongly in } L^2(S_i),$$
$$\frac{\partial y_i^n}{\partial \nu_i}|_{S_i} \to \frac{\partial y_i}{\partial \nu_i}|_{S_i}, \quad \frac{\partial p_i^n}{\partial \nu_i}|_{S_i} \to \frac{\partial p_i}{\partial \nu_i}|_{S_i} \quad \text{weakly in } L^2(S_i),$$
$$\mathbf{P}^n(T) \to \mathbf{P}(T) \quad \text{strongly in } \prod_{i=1}^m V_i,$$
$$\mathbf{P}^n_{,t}(T) \to \mathbf{P}_{,t}(T) \quad \text{strongly in } H,$$
$$\frac{\partial p_i^n(T)}{\partial \nu_i}|_{\gamma_i} \to \frac{\partial p_i(T)}{\partial \nu_i}|_{\gamma_i} \quad \text{weakly in } L^2(\gamma_i).$$

Remark 3.2. The proof of convergence for the case without relaxation ($\epsilon = 0$) is an open problem.

Proof. For $n = 1, 2, \dots$, set

$$\tilde{y}_i^n = y_i^n - y_i, \quad \tilde{p}_i^n = p_i^n - p_i,$$
$$\tilde{\lambda}_{ij}^n = \lambda_{ij}^n - \lambda_{ij} = (1-\epsilon)\left(-a_j \frac{\partial \tilde{y}_j^n}{\partial \nu_j} + \beta \tilde{p}_j^n\right)|_{\Sigma_{ij}} + \epsilon\left(a_i \frac{\partial \tilde{y}_i^n}{\partial \nu_i} + \beta \tilde{p}_i^n\right)|_{\Sigma_{ij}},$$
$$\tilde{\mu}_{ij}^n = \mu_{ij}^n - \mu_{ij} = (1-\epsilon)\left(-a_j \frac{\partial \tilde{p}_j^n}{\partial \nu_j} - \beta \tilde{y}_j^n\right)|_{\Sigma_{ij}} + \epsilon\left(a_i \frac{\partial \tilde{p}_i^n}{\partial \nu_i} - \beta \tilde{y}_i^n\right)|_{\Sigma_{ij}},$$
$$\tilde{\tau}_{ij}^n = \tau_{ij}^n - \tau_{ij} = (1-\epsilon)\left(-a_j \frac{\partial \tilde{p}_j^n(T)}{\partial \nu_j} + \alpha \tilde{p}_j^n(T)\right)|_{\Gamma_{ij}} + \epsilon\left(a_i \frac{\partial \tilde{p}_i^n(T)}{\partial \nu_i} + \alpha \tilde{p}_i^n(T)\right)|_{\Gamma_{ij}},$$
$$\tilde{\sigma}_i^n = \sigma_i^n - \sigma_i.$$

The pair $(\tilde{y}_i^n, \tilde{p}_i^n)$ satisfies

$$\begin{aligned}
&\tilde{y}_{i,tt}^{n+1} - a_i \Delta \tilde{y}_i^{n+1} = 0, \quad \tilde{p}_{i,tt}^{n+1} - a_i \Delta \tilde{p}_i^{n+1} = 0 \quad \text{in } Q_i,\\
&a_i \frac{\partial \tilde{y}_i^{n+1}}{\partial \nu_i} + \beta \tilde{p}_i^{n+1} = \tilde{\lambda}_{ij}^n, \quad a_i \frac{\partial \tilde{p}_i^{n+1}}{\partial \nu_i} - \beta \tilde{y}_i^{n+1} = \tilde{\mu}_{ij}^n \quad \text{on } \Sigma_{ij},\\
&a_i \frac{\partial \tilde{y}_i^{n+1}}{\partial \nu_i} + \tilde{p}_i^{n+1} = 0, \quad a_i \frac{\partial \tilde{p}_i^{n+1}}{\partial \nu_i} = 0 \quad \text{on } \Sigma_i,\\
&\tilde{y}_i^{n+1}(0) = \tilde{y}_{i,t}^{n+1}(0) = 0,\\
&\tilde{p}_i^{n+1}(T) = k\mathcal{A}_i^{-1}(\tilde{y}_{i,t}^{n+1}(T) + k^{-1}\tilde{\sigma}_i^n), \quad \tilde{p}_{i,t}^{n+1}(T) = -k\tilde{y}_i^{n+1}(T).
\end{aligned} \tag{3.8}$$

It is important to note that the iterations on the interfaces Σ_{ij} may be expressed in terms of the mapping $\mathcal{T}$ defined in the last section as follows. Set

$$X_i^n := \left(\left(a_i \frac{\partial \tilde{y}_i^n}{\partial \nu_i} + \beta \tilde{p}_i^n \right) |_{\Sigma_i}, \left(a_i \frac{\partial \tilde{p}_i^n}{\partial \nu_i} - \beta \tilde{y}_i^n \right) |_{\Sigma_i}, \left(a_i \frac{\partial \tilde{p}_i^n(T)}{\partial \nu_i} + \alpha \tilde{p}_i^n(T) \right) |_{\gamma_i} \right),$$
$$X^n = \{X_i^n\}_{i=1}^m.$$

Then the interface conditions may be expressed as the relaxed fixed point iteration

$$X^{n+1} = (1-\epsilon)\mathcal{T}X^n + \epsilon X^n. \tag{3.9}$$

It follows from Lemma 2.2 and Proposition 2.3 that if $0 < \epsilon < 1$, X^n converges weakly in $\mathcal{X}$ to a fixed point X of $\mathcal{T}$. This fixed point must be $X = 0$ as was noted above (immediately preceding the statement of Lemma 2.2). Thus $X^n \to 0$ and $\mathcal{T}X^n \to 0$ weakly in $\mathcal{X}$, from which follows that

$$\tilde{y}_i^n|_{S_i} \to 0, \quad \tilde{p}_i^n|_{S_i} \to 0, \quad \frac{\partial \tilde{y}_i^n}{\partial \nu_i}|_{S_i} \to 0, \quad \frac{\partial \tilde{p}_i^n}{\partial \nu_i}|_{S_i} \to 0$$

weakly in $L^2(S_i)$, and

$$\tilde{p}_i^n(T)|_{\gamma_i} \to 0, \quad \frac{\partial \tilde{p}_i^n(T)}{\partial \nu_i}|_{\gamma_i} \to 0$$

weakly in $L^2(\gamma_i)$.

From (3.9) and Lemma 2.2 we have

$$\begin{aligned} \|X^{n+1}\|_{\mathcal{X}}^2 &= (1-\epsilon)^2 \|\mathcal{T}X^n\|_{\mathcal{X}}^2 + \epsilon^2 \|X^n\|_{\mathcal{X}}^2 + 2\epsilon(1-\epsilon)(X^n, \mathcal{T}X^n)_{\mathcal{X}} \\ &= ((1-\epsilon)^2 + \epsilon^2)\|X^n\|_{\mathcal{X}}^2 - \frac{4(1-\epsilon)^2}{k} \sum_{i=1}^m \mathcal{E}_i(\tilde{p}_i^n(T)) \\ &\quad - 4(1-\epsilon)^2 \int_\Sigma |\tilde{p}_m^n|^2 d\Sigma + 2\epsilon(1-\epsilon)(X^n, \mathcal{T}X^n)_{\mathcal{X}}, \end{aligned} \tag{3.10}$$

where

$$\mathcal{E}_i(\phi) = \|\phi\|_{H^1(\Omega_i)}^2 + \|\phi_{,t}\|_{L^2(\Omega_i)}^2.$$

By the same calculation as in the proof of Lemma 2.2, we obtain

$$\begin{aligned} \|X^n\|_{\mathcal{X}}^2 = \sum_{i=1}^m \Bigg\{ &\frac{1}{\alpha k} \int_{\gamma_i} \left(\alpha^2 |\tilde{p}_i^n(T)|^2 + \left| a_i \frac{\partial \tilde{p}_i^n(T)}{\partial \nu_i} \right|^2 \right) d\Gamma \\ &+ \frac{1}{\beta} \int_{S_i} \left(\beta^2 (|\tilde{y}_i^n|^2 + |\tilde{p}_i^n|^2) + \left| a_i \frac{\partial \tilde{y}_i^n}{\partial \nu_i} \right|^2 + \left| a_i \frac{\partial \tilde{p}_i^n}{\partial \nu_i} \right|^2 \right) d\Sigma \Bigg\} \\ &+ \frac{2}{k} \sum_{i=1}^m \mathcal{E}_i(\tilde{p}_i^n(T)) + 2 \int_\Sigma |\tilde{p}_m^n|^2 d\Sigma. \end{aligned} \tag{3.11}$$

It follows from (3.10), (3.11) that

$$\begin{aligned} \|X^{n+1}\|_{\mathcal{X}}^2 = ((1-\epsilon)^2 + \epsilon^2)E^n - \frac{2}{k}(1-2\epsilon) \sum_{i=1}^m \mathcal{E}_i(\tilde{p}_i^n(T)) \\ - 2(1-2\epsilon) \int_\Sigma |\tilde{p}_m^n|^2 d\Sigma + 2\epsilon(1-\epsilon)(X^n, \mathcal{T}X^n)_{\mathcal{X}}, \end{aligned} \tag{3.12}$$

where

$$E^n := \sum_{i=1}^{m} \left\{ \frac{1}{\alpha k} \int_{\gamma_i} \left(\alpha^2 |\tilde{p}_i^n(T)|^2 + \left| a_i \frac{\partial \tilde{p}_i^n(T)}{\partial \nu_i} \right|^2 \right) d\Gamma \right.$$
$$\left. + \frac{1}{\beta} \int_{S_i} \left(\beta^2 (|\tilde{y}_i^n|^2 + |\tilde{p}_i^n|^2) + \left| a_i \frac{\partial \tilde{y}_i^n}{\partial \nu_i} \right|^2 + \left| a_i \frac{\partial \tilde{p}_i^n}{\partial \nu_i} \right|^2 \right) d\Sigma \right\}.$$

From (3.12) and (3.11) with n replaced by $n+1$ we obtain

$$E^{n+1} = ((1-\epsilon)^2 + \epsilon^2) E^n - \frac{2}{k} \sum_{i=1}^{m} [(1-2\epsilon)\mathcal{E}_i(\tilde{p}_i^n(T)) + \mathcal{E}_i(\tilde{p}_i^{n+1}(T))] \tag{3.13}$$
$$- 2 \int_{\Sigma} [(1-2\epsilon)|\tilde{p}_m^n|^2 + |\tilde{p}_m^{n+1}|^2] d\Sigma + 2\epsilon(1-\epsilon)(X^n, \mathcal{T}X^n)_{\mathcal{X}}.$$

We have

$$(X^n, \mathcal{T}X^n)_{\mathcal{X}} = \sum_{i=1}^{m} \sum_{j:\Gamma_{ij} \neq \emptyset} \left\{ \frac{1}{\beta} \int_{\Sigma_{ij}} \left[\left(a_i \frac{\partial \tilde{y}_i^n}{\partial \nu_i} + \beta \tilde{p}_i^n \right) \left(-a_j \frac{\partial \tilde{y}_j^n}{\partial \nu_j} + \beta \tilde{p}_j^n \right) \right.\right.$$
$$\left. - \left(a_i \frac{\partial \tilde{p}_i^n}{\partial \nu_i} - \beta \tilde{y}_i^n \right) \left(a_j \frac{\partial \tilde{p}_j^n}{\partial \nu_j} + \beta \tilde{y}_j^n \right) \right] d\Sigma$$
$$\left. + \frac{1}{\alpha k} \int_{\Gamma_{ij}} \left(a_i \frac{\partial \tilde{p}_i^n(T)}{\partial \nu_i} + \alpha \tilde{p}_i^n(T) \right) \left(-a_j \frac{\partial \tilde{p}_j^n(T)}{\partial \nu_i} + \alpha \tilde{p}_j^n(T) \right) d\Gamma \right\}$$
$$= \sum_{i=1}^{m} \sum_{j:\Gamma_{ij} \neq \emptyset} \left\{ \frac{1}{\beta} \int_{\Sigma_{ij}} \left[-a_i \frac{\partial \tilde{y}_i^n}{\partial \nu_i} a_j \frac{\partial \tilde{y}_j^n}{\partial \nu_j} - a_i \frac{\partial \tilde{p}_i^n}{\partial \nu_i} a_j \frac{\partial \tilde{p}_j^n}{\partial \nu_j} \right.\right.$$
$$+ \beta^2 (\tilde{p}_i^n \tilde{p}_j^n + \tilde{y}_i^n \tilde{y}_j^n) + \beta \left(\tilde{p}_j^n a_i \frac{\partial \tilde{y}_i^n}{\partial \nu_i} - \tilde{p}_i^n a_j \frac{\partial \tilde{y}_j^n}{\partial \nu_j} \right)$$
$$\left. -\beta \left(\tilde{y}_j^n a_i \frac{\partial \tilde{p}_i^n}{\partial \nu_i} - \tilde{y}_i^n a_j \frac{\partial \tilde{p}_j^n}{\partial \nu_j} \right) \right] d\Sigma$$
$$+ \frac{1}{\alpha k} \int_{\Gamma_{ij}} \left[-a_i \frac{\partial \tilde{p}_i^n(T)}{\partial \nu_i} a_j \frac{\partial \tilde{p}_j^n(T)}{\partial \nu_j} + \alpha^2 \tilde{p}_i^n(T) \tilde{p}_j^n(T) \right.$$
$$\left. \left. + \alpha \left(\tilde{p}_j^n(T) a_i \frac{\partial \tilde{p}_i^n(T)}{\partial \nu_i} - \tilde{p}_i^n(T) a_j \frac{\partial \tilde{p}_j^n(T)}{\partial \nu_j} \right) \right] d\Gamma \right\}.$$

One has

$$\sum_{i=1}^{m} \sum_{j:\Gamma_{ij} \neq \emptyset} \int_{\Sigma_{ij}} \left(\tilde{p}_j^n a_i \frac{\partial \tilde{y}_i^n}{\partial \nu_i} - \tilde{p}_i^n a_j \frac{\partial \tilde{y}_j^n}{\partial \nu_j} \right) d\Sigma = 0,$$

and similarly for the other integrals with mixed terms. Therefore

$$(X^n, \mathcal{T}X^n)_{\mathcal{X}} = \sum_{i=1}^{m} \sum_{j:\Gamma_{ij} \neq \emptyset} \left\{ \frac{1}{\beta} \int_{\Sigma_{ij}} \left[- a_i \frac{\partial \tilde{y}_i^n}{\partial \nu_i} a_j \frac{\partial \tilde{y}_j^n}{\partial \nu_j} - a_i \frac{\partial \tilde{p}_i^n}{\partial \nu_i} a_j \frac{\partial \tilde{p}_j^n}{\partial \nu_j} \right.\right.$$
$$\left. + \beta^2 (\tilde{p}_i^n \tilde{p}_j^n + \tilde{y}_i^n \tilde{y}_j^n) \right] d\Sigma$$
$$\left. + \frac{1}{\alpha k} \int_{\Gamma_{ij}} \left[-a_i \frac{\partial \tilde{p}_i^n(T)}{\partial \nu_i} a_j \frac{\partial \tilde{p}_j^n(T)}{\partial \nu_j} + \alpha^2 \tilde{p}_i^n(T) \tilde{p}_j^n(T) \right] d\Gamma \right\}.$$

Since

$$\sum_{i=1}^{m}\int_{S_i}|f_i|^2d\Sigma=\sum_{i=1}^{m-1}\sum_{\substack{j:\Gamma_{ij}\neq\emptyset\\ j>i}}\int_{\Sigma_{ij}}(|f_i|^2+|f_j|^2)\,d\Sigma,$$

$$\sum_{i=1}^{m}\sum_{j:\Gamma_{ij}\neq\emptyset}\int_{\Sigma_{ij}}f_if_j\,d\Sigma=2\sum_{i=1}^{m-1}\sum_{\substack{j:\Gamma_{ij}\neq\emptyset\\ j>i}}\int_{\Sigma_{ij}}f_if_j\,d\Sigma,$$

we have

(3.14)

$$\begin{aligned}
&((1-\epsilon)^2+\epsilon^2)E^n+2\epsilon(1-\epsilon)(X^n,\mathcal{T}X^n)_{\mathcal{X}}\\
&=\sum_{i=1}^{m}\sum_{\substack{j:\Gamma_{ij}\neq\emptyset\\ j>i}}\left\{\frac{1}{\beta}\int_{\Sigma_{ij}}\left[((1-\epsilon)^2+\epsilon^2)\left(\left|a_i\frac{\partial\tilde y_i^n}{\partial\nu_i}\right|^2+\left|a_j\frac{\partial\tilde y_j^n}{\partial\nu_j}\right|^2\right.\right.\right.\\
&\quad+\left.\left|a_i\frac{\partial\tilde p_i^n}{\partial\nu_i}\right|^2+\left|a_j\frac{\partial\tilde p_j^n}{\partial\nu_j}\right|^2\right)-4\epsilon(1-\epsilon)\left(a_i\frac{\partial\tilde y_i^n}{\partial\nu_i}a_j\frac{\partial\tilde y_j^n}{\partial\nu_j}+a_i\frac{\partial\tilde p_i^n}{\partial\nu_i}a_j\frac{\partial\tilde p_j^n}{\partial\nu_j}\right)\\
&\quad\left.+\beta^2((1-\epsilon)^2+\epsilon^2)(|\tilde y_i^n|^2+|\tilde y_j^n|^2+|\tilde p_i^n|^2+|\tilde p_j^n|^2)+4\beta^2\epsilon(1-\epsilon)(\tilde y_i^n\tilde y_j^n+\tilde p_i^n\tilde p_j^n)\right]d\Sigma\\
&\quad+\frac{1}{\alpha k}\int_{\Gamma_{ij}}\left[((1-\epsilon)^2+\epsilon^2)\left(\left|a_i\frac{\partial\tilde p_i^n(T)}{\partial\nu_i}\right|^2+\left|a_j\frac{\partial\tilde p_j^n(T)}{\partial\nu_j}\right|^2\right.\right.\\
&\quad\left.+\alpha^2(|\tilde p_i^n(T)|^2+|\tilde p_j^n(T)|^2)\right)\\
&\quad\left.-4\epsilon(1-\epsilon)a_i\frac{\partial\tilde p_i^n(T)}{\partial\nu_i}a_j\frac{\partial\tilde p_j^n(T)}{\partial\nu_j}+4\alpha^2\epsilon(1-\epsilon)\tilde p_i^n(T)\tilde p_j^n(T)\right]d\Gamma\Bigg\}\\
&\le[(1-\epsilon)^2+\epsilon^2+2\epsilon(1-\epsilon)]E^n=E^n.
\end{aligned}$$

Substitution of (3.14) into (3.13) yields

$$\begin{aligned}
E^{n+1}\le E^n-\frac{2}{k}\sum_{i=1}^{m}[\mathcal{E}_i(\tilde p_i^{n+1}(T))+(1-2\epsilon)\mathcal{E}_i(\tilde p_i^n(T))]\\
-2\int_\Sigma(|\tilde p_m^{n+1}|^2+(1-2\epsilon)|\tilde p_m^n|^2)\,d\Sigma.
\end{aligned}$$

By iteration we obtain

$$E^{n+1}\le E^1-2\sum_{\ell=1}^{n+1}\left[\frac{1}{k}\sum_{i=1}^{m}c_\ell(\epsilon)\mathcal{E}_i(\tilde p_i^\ell(T))-\int_\Sigma c_\ell(\epsilon)|\tilde p_m^\ell|^2d\Sigma\right],\tag{3.15}$$

where

$$c_1(\epsilon)=1-2\epsilon,\quad c_{n+1}(\epsilon)=1,\quad c_\ell(\epsilon)=2(1-\epsilon),\ \ell=2,\dots,n.$$

It follows from (3.15) that for $0\le\epsilon<1$,

$$\sum_{\ell=1}^{\infty}\sum_{i=1}^{m}\mathcal{E}_i(\tilde p_i^\ell(T))<\infty,\qquad\sum_{\ell=1}^{\infty}\int_\Sigma|\tilde p_m^\ell|^2d\Sigma<\infty$$

so that

$$\tilde{p}_m^n|_\Sigma \to 0 \quad \text{strongly in } L^2(\Sigma), \;\; \mathcal{E}_i(\tilde{p}_i^n(T)) \to 0, \quad i = 1, \dots, m,$$

from the second of which follows that

$$\begin{aligned} \tilde{p}_i^n(T) &\to 0 \quad \text{in } V_i, \\ \tilde{p}_{i,t}^n(T) &\to 0 \quad \text{in } L^2(\Omega_i), \quad i = 1, \dots, m. \end{aligned} \tag{3.16}$$

The boundedness of $\{E^n\}_{n=1}^\infty$, together with (3.16) and the regularity results of Lasiecka–Triggiani [20, Theorems 2.0, 2.1] mentioned above, allow us to conclude that

$$\begin{aligned} &\{\tilde{y}_i^n\}_{n=1}^\infty, \quad \{\tilde{p}_i^n\}_{n=1}^\infty \quad \text{are bounded in } C([0,T]; H^{3/5-\varepsilon}(\Omega_i)), \\ &\{\tilde{y}_{i,t}^n\}_{n=1}^\infty, \quad \{\tilde{p}_{i,t}^n\}_{n=1}^\infty \quad \text{are bounded in } C([0,T]; (H^{2/5+\varepsilon}(\Omega_i))'), \\ &\{\tilde{y}_{i,tt}^n\}_{n=1}^\infty, \quad \{\tilde{p}_{i,tt}^n\}_{n=1}^\infty \quad \text{are bounded in } L^2(0,T; (H^{7/5+\varepsilon}(\Omega_i))'). \end{aligned} \tag{3.17}$$

In particular, $\tilde{y}_i^n|_{S_i}$, $\tilde{p}_i^n|_{S_i}$ are bounded in $L^\infty(0,T; H^{1/10-\varepsilon}(\gamma_i))$. Since they converge weakly to zero in $L^2(S_i)$ when $0 < \epsilon < 1$, a compactness result of Simon [32, Corollary 5] shows that the convergence is *strong* in that space (in fact, it is strong in $L^\infty(0,T; L^2(\gamma_i))$). The same result also implies that on a subsequence we have

$$\begin{aligned} \tilde{y}_i^n \to \tilde{y}_i, \quad \tilde{p}_i^n \to \tilde{p}_i \quad &\text{in } C([0,T]; L^2(\Omega_i)), \\ \tilde{y}_{i,t}^n \to \tilde{y}_{i,t}, \quad \tilde{p}_{i,t}^n \to \tilde{p}_{i,t} \quad &\text{in } C([0,T]; (H^1(\Omega_i))'), \end{aligned}$$

where $\tilde{y}_i$, $\tilde{p}_i$ is the solution of

$$\begin{aligned} \tilde{y}_{i,tt} - a_i \Delta \tilde{y}_i = 0, \quad \tilde{p}_{i,tt} - a_i \Delta \tilde{p}_i = 0 \quad &\text{in } Q_i, \\ a_i \frac{\partial \tilde{y}_i}{\partial \nu_i} = 0, \quad a_i \frac{\partial \tilde{p}_i}{\partial \nu_i} = 0 \quad &\text{on } S_i, \\ a_i \frac{\partial \tilde{y}_i}{\partial \nu_i} = 0, \quad a_i \frac{\partial \tilde{p}_i}{\partial \nu_i} = 0 \quad &\text{on } \Sigma_i, \\ \tilde{y}_i(0) = \tilde{y}_{i,t}(0) = \tilde{p}_i(T) = \tilde{p}_{i,t}(T) = 0. & \end{aligned}$$

It follows that $\tilde{y}_i = \tilde{p}_i = 0$ in Q_i, $i = 1, \dots, m$. □

4. The limit of the global optimality system. In this section we study the limiting behavior as $k \to \infty$ of the solution of the global optimality system (1.1), (1.5)–(1.7) under the assumption that the global problem (1.1) is exactly controllable to $V \times H$ for T sufficiently large. As is to be expected, the limit of the global optimality system is the optimality system for the state constrained optimal control problem

$$J(f) = \frac{1}{2} \int_\Sigma |f|^2 d\Sigma \to \inf \tag{4.1}$$

subject to (1.1) and

$$\mathbf{Y}(T) = \mathbf{Y}_T, \quad \mathbf{Y}_{,t}(T) = \dot{\mathbf{Y}}_T. \tag{4.2}$$

We denote by $J_k(f)$ the cost functional in (1.4) and by $\mathbf{Y}(\cdot\,; k)$, $\mathbf{P}(\cdot\,; k)$ the solution of the optimality system (1.1), (1.5)–(1.7). Let Φ be the solution of (1.3), where $(\Phi^0, \Phi^1) \in V \times H$. The assumption that $V \times H$ lies in the range of the control-to-state

mapping $f \mapsto (\mathbf{Y}(T), \mathbf{Y}_{,t}(T)) : L^2(\Sigma) \mapsto H \times V'$ for T sufficiently large is equivalent to the assumption that (1.3) is *continuously observable* from Σ for T sufficiently large in the following sense: there exists $T_0 > 0$ such that

$$\|\Phi^0\|_H^2 + \|\Phi^1\|_{V'}^2 \le C_T \int_\Sigma |\phi_m|^2 d\Sigma, \quad T > T_0, \tag{4.3}$$

for some constant C_T independent of (Φ^0, Φ^1).

We shall prove the following result.

THEOREM 4.1. *Suppose that* (4.3) *holds,* $T > T_0$, *and let* $(\mathbf{Y}_T, \dot{\mathbf{Y}}_T) \in V \times H$. *Then as* $k \to \infty$ *the solution of the optimality system* (1.1), (1.5)–(1.7) *satisfies*

$$\begin{gathered}
\mathbf{Y}(\cdot\,; k) \to \mathbf{Y}(\cdot), \quad \mathbf{P}(\cdot\,; k) \to \mathbf{P}(\cdot) \quad \text{strongly in } C([0,T]; H), \\
\mathbf{Y}_{,t}(\cdot\,; k) \to \mathbf{Y}_{,t}(\cdot), \quad \mathbf{P}_{,t}(\cdot\,; k) \to \mathbf{P}_{,t}(\cdot) \quad \text{strongly in } C([0,T]; V'), \\
p_m(\cdot\,; k)|_\Sigma \to p_m(\cdot)|_\Sigma \quad \text{strongly in } L^2(\Sigma),
\end{gathered}$$

where $\mathbf{Y} = (y_1, \dots, y_m)$, $\mathbf{P} = (p_1, \dots, p_m)$ *satisfy*

$$\begin{gathered}
y_{i,tt} - a_i \Delta y_i = 0, \quad p_{i,tt} - a_i \Delta p_i = 0 \quad \text{in } Q_i, \\
y_i = y_j, \quad a_i \frac{\partial y_i}{\partial \nu_i} + a_j \frac{\partial y_j}{\partial \nu_j} = 0 \quad \text{on } \Sigma_{ij}, \\
p_i = p_j, \quad a_i \frac{\partial p_i}{\partial \nu_i} + a_j \frac{\partial p_j}{\partial \nu_j} = 0 \quad \text{on } \Sigma_{ij}, \\
a_m \frac{\partial y_m}{\partial \nu_m} = -p_m, \quad a_m \frac{\partial p_m}{\partial \nu_m} = 0 \quad \text{on } \Sigma, \\
\mathbf{Y}(0) = \mathbf{Y}_{,t}(0) = 0, \quad \mathbf{P}(T) = \mathbf{P}_T, \quad \mathbf{P}_{,t}(T) = \dot{\mathbf{P}}_T, \\
\mathbf{Y}(T) = \mathbf{Y}_T, \quad \mathbf{Y}_{,t}(T) = \dot{\mathbf{Y}}_T,
\end{gathered} \tag{4.4}$$

and where $(\mathbf{P}_T, \dot{\mathbf{P}}_T) \in H \times V'$ *is the unique solution of*

$$((\mathbf{P}_T, \dot{\mathbf{P}}_T), (-\dot{\mathbf{Y}}_T, \mathbf{Y}_T))_{H \times V'} = \int_\Sigma |p_m|^2 d\Sigma. \tag{4.5}$$

The system (4.4) *is the optimality system for the problem* (4.1) *subject to* (1.1) *and* (4.2).

Proof. Let $T > T_0$ and $f_k(\cdot) = -p_m(\cdot\,; k)|_\Sigma$ be the optimal control for the optimal control problem (1.1), (1.4). Let f_0 be any $L^2(\Sigma)$ control such that the solution of (1.1) satisfies (4.2). Then

$$J_k(f_k) \le J_k(f_0) = \frac{1}{2} \|f_0\|_{L^2(\Sigma)}; \tag{4.6}$$

hence

$$\begin{gathered}
\{f_k\} \quad \text{is bounded in } L^2(\Sigma), \\
\sqrt{k}(\mathbf{Y}(T; k) - \mathbf{Y}_T) \quad \text{is bounded in } H, \\
\sqrt{k}(\mathbf{Y}_{,t}(T; k) - \dot{\mathbf{Y}}_T) \quad \text{is bounded in } V'.
\end{gathered}$$

Therefore

$$\begin{aligned}
&\lim_{k \to \infty} \mathbf{Y}(T; k) = \mathbf{Y}_T \quad \text{strongly in } H, \\
&\lim_{k \to \infty} \mathbf{Y}_{,t}(T; k) = \dot{\mathbf{Y}}_T \quad \text{strongly in } V',
\end{aligned} \tag{4.7}$$

and, as $k \to \infty$ through an appropriate subnet of $k > 0$,

$$f_k \to f_\infty \quad \text{weakly in } L^2(\Sigma)$$

for some $f_\infty \in L^2(\Sigma)$ that satisfies

$$\|f_\infty\|_{L^2(\Sigma)} \le \liminf \|f_k\|_{L^2(\Sigma)}. \tag{4.8}$$

The solution $\mathbf{Y}(\cdot\,;k)$ satisfies

$$\|(\mathbf{Y}(\cdot,;k), \mathbf{Y}_{,t}(\cdot\,;k))\|_{C([0,T];H\times V')} \le C\|f_k\|_{L^2(\Sigma)}, \tag{4.9}$$

where C is independent of k, and therefore we may extract a subnet of $k > 0$ on which as $k \to \infty$,

$$\begin{aligned} \mathbf{Y}(\cdot\,;k) &\to \mathbf{Y}(\cdot) \quad \text{weakly}^* \text{ in } L^\infty(0,T;H), \\ \mathbf{Y}_{,t}(\cdot\,;k) &\to \mathbf{Y}_{,t}(\cdot) \quad \text{weakly}^* \text{ in } L^\infty(0,T;V') \end{aligned}$$

for some $\mathbf{Y} \in C([0,T];H) \cap C^1([0,T];V')$. It follows from (1.2) and (4.7) that $\mathbf{Y}$ satisfies

$$\begin{aligned} (\mathbf{Y}_{,t}(t), \Phi(t))_{V'} - (\mathbf{Y}(t), \Phi_{,t}(t)) &= \int_0^t \int_\Gamma f_\infty \phi_m \, d\Gamma dt, \\ (\mathbf{Y}_{,t}(T) - \dot{\mathbf{Y}}_T, \Phi(T))_{V'} - (\mathbf{Y}(T) - \mathbf{Y}_T, \Phi_{,t}(T)) &= 0 \quad \forall (\Phi^0, \Phi^1) \in V \times H, \end{aligned}$$

where Φ is the solution of (1.3). Therefore $\mathbf{Y}$ is the solution of (1.1) corresponding to $f = f_\infty$ and $\mathbf{Y}$ satisfies (4.2). Set

$$\mathcal{U}_{\text{ad}} = \{f \in L^2(\Sigma) : \mathbf{Y} \text{ satisfies (1.1) and (4.2)}\}.$$

We have

$$J_k(f_k) \le \inf_{f\in\mathcal{U}_{\text{ad}}} J_k(f) = \frac{1}{2} \inf_{f\in\mathcal{U}_{\text{ad}}} \|f\|^2_{L^2(\Sigma)}. \tag{4.10}$$

It follows from (4.8) that f_∞ is the control of minimum $L^2(\Sigma)$ norm in $\mathcal{U}_{\text{ad}}$. Because this control is unique (since $\mathcal{U}_{\text{ad}}$ is closed and convex and the mapping $f \mapsto J(f)$ is continuous on $L^2(\Sigma)$), it follows from (4.8) and (4.10) that

$$f_k \to f_\infty \quad \text{strongly in } L^2(\Sigma) \text{ as } k \to \infty \tag{4.11}$$

and that

$$\|f_\infty\|_{L^2(\Sigma)} = \inf_{f\in\mathcal{U}_{\text{ad}}} \|f\|_{L^2(\Sigma)}.$$

From (4.9) we then obtain

$$(\mathbf{Y}(\cdot\,;k), \mathbf{Y}_{,t}(\cdot\,;k)) \to (\mathbf{Y}(\cdot), \mathbf{Y}_{,t}(\cdot)) \quad \text{in } C([0,T];H\times V').$$

As $f_k = -p_m(\cdot\,;k)|_\Sigma$, assumption (4.3) together with (4.11) shows that

$$(\mathbf{P}(T;k), \mathbf{P}_{,t}(T;k)) \to (\mathbf{P}_T, \dot{\mathbf{P}}_T) \quad \text{strongly in } H\times V'$$

for some $(\mathbf{P}_T, \dot{\mathbf{P}}_T)$. Since the linear map $(\mathbf{P}(T;k), \mathbf{P}_{,t}(T;k)) \mapsto (\mathbf{P}(\cdot\,;k), \mathbf{P}_{,t}(\cdot\,;k)) : H \times V' \mapsto C([0,T];H \times V')$ is continuous we have

$$\begin{aligned} \mathbf{P}(\cdot\,;k) &\to \mathbf{P}(\cdot) \quad \text{in } C([0,T];H), \\ \mathbf{P}_{,t}(\cdot\,;k) &\to \mathbf{P}_{,t}(\cdot) \quad \text{in } C([0,T];V'), \end{aligned}$$

where $\mathbf{P} = (p_1, \dots, p_m)$ is the unique solution in $C([0,T];H) \cap C^1([0,T];V')$ of

$$\begin{gathered} p_{i,tt} - a_i \Delta p_i = 0 \quad \text{in } Q_i, \\ p_i = p_j, \quad a_i \frac{\partial p_i}{\partial \nu_i} + a_j \frac{\partial p_j}{\partial \nu_j} = 0 \quad \text{on } \Sigma_{ij}, \\ a_m \frac{\partial p_m}{\partial \nu_m} = 0 \quad \text{on } \Sigma, \\ \mathbf{P}(T) = \mathbf{P}_T, \quad \mathbf{P}_{,t}(T) = \dot{\mathbf{P}}_T. \end{gathered} \tag{4.12}$$

Since $\mathbf{P}(\cdot\,;k) \to \mathbf{P}(\cdot)$ strongly in $C([0,T];H)$ and $p_m(\cdot\,;k)|_\Sigma \to -f_\infty$ strongly in $L^2(\Sigma)$, by definition

$$p_m|_\Sigma = -f_\infty.$$

For $i = 1, 2, \dots$, we may choose that $\hat{f}_i \in L^2(\Sigma)$ be such that $\hat{f}_i \to f_\infty$ strongly in $L^2(\Sigma)$ and the corresponding solution $\hat{\mathbf{Y}}_i$ of (1.1) satisfies $(\hat{\mathbf{Y}}_i, \hat{\mathbf{Y}}_{i,t}) \in C([0,T]; V \times H)$. One then has

$$(\hat{\mathbf{Y}}_i, \hat{\mathbf{Y}}_{i,t}) \to (\mathbf{Y}, \mathbf{Y}_{,t}) \quad \text{in } C([0,T]; H \times V').$$

An integration by parts gives

$$0 = (\mathbf{P}_{,t}(T;k), \hat{\mathbf{Y}}_i(T)) - (\mathbf{P}(T;k), \hat{\mathbf{Y}}_{i,t}(T)) + \int_\Sigma p_m(\cdot\,;k) \hat{f}_i \, d\Sigma.$$

Upon passing to the limit first in i and then in k we obtain

$$(\dot{\mathbf{P}}_T, \mathbf{Y}_T)_{V'} - (\mathbf{P}_T, \dot{\mathbf{Y}}_T) = \int_\Sigma |p_m|^2 d\Sigma,$$

that is, $(\mathbf{P}_T, \dot{\mathbf{P}}_T)$ is the unique solution of the equation

$$((\mathbf{P}_T, \dot{\mathbf{P}}_T), (-\dot{\mathbf{Y}}_T, \mathbf{Y}_T))_{H \times V'} = \int_\Sigma |p_m|^2 d\Sigma. \tag{4.13}$$

□

Remark 4.2. Of course, the solution $(\mathbf{P}_T, \dot{\mathbf{P}}_T)$ of (4.13) is exactly the one obtained from the Hilbert uniqueness method. Suppose that $(\mathbf{P}_T, \dot{\mathbf{P}}_T) \in V \times H$, and let $\mathbf{Y}$, $\mathbf{P}$ be the solution of the global problem (4.4), omitting the condition on $(\mathbf{Y}(T), \mathbf{Y}_{,t}(T))$. By Holmgren's theorem, for $T > T_0$ one may define a norm $\|(\mathbf{P}_T, \dot{\mathbf{P}}_T)\|_F$ by setting

$$\|(\mathbf{P}_T, \dot{\mathbf{P}}_T)\|_F^2 = \int_\Sigma |p_m|^2 d\Sigma,$$

and a corresponding Hilbert space

$$F = \text{ completion of } V \times H \text{ in } \| \cdot \|_F.$$

By (4.3) we have $F \subset H \times V'$, so $H \times V \subset F'$. Then (4.13) is the same as

$$(\mathbf{P}_T, \dot{\mathbf{P}}_T) = \Lambda^{-1}(-\dot{\mathbf{Y}}_T, \mathbf{Y}_T),$$

where Λ is the canonical isomorphism $F \mapsto F'$.

5. The limit of the local optimality system. In this section we wish to study the limiting behavior as $k \to \infty$ of the solution $y_i^{n+1}(\cdot\,;k)$, $p_i^{n+1}(\cdot\,;k)$ of the local problem (3.1)–(3.5). To this end we first consider the limiting behavior as $k \to \infty$ of the solution $z_i(\cdot\,;k)$, $q_i(\cdot\,;k)$ of the system

$$\begin{aligned} z_{i,tt} - a_i \Delta z_i &= 0 \quad \text{in } Q_i, \\ a_i \frac{\partial z_i}{\partial \nu_i} &= f_1^k + \lambda_i(\cdot\,;k) \quad \text{on } S_i, \\ a_i \frac{\partial z_i}{\partial \nu_i} &= f_2^k \quad \text{on } \Sigma_i, \\ z_i(0;k) = z_{i,t}(0;k) &= 0, \end{aligned} \tag{5.1}$$

$$\begin{aligned} q_{i,tt} - a_i \Delta q_i &= 0 \quad \text{in } Q_i, \\ a_i \frac{\partial q_i}{\partial \nu_i} &= \beta z_i + \mu_i(\cdot\,;k) \quad \text{on } S_i, \\ a_i \frac{\partial q_i}{\partial \nu_i} &= 0 \quad \text{on } \Sigma_i, \\ q_i(T;k) &= k\mathcal{A}_i^{-1}(z_{i,t}(T;k) - \dot{y}_{iT} + k^{-1}\sigma_i(k)), \\ q_{i,t}(T;k) &= -k(z_i(T;k) - y_{iT}), \end{aligned} \tag{5.2}$$

$$f_1^k = -\beta q_i(\cdot\,;k)|_{S_i}, \quad f_2^k = -q_i(\cdot\,;k)|_{\Sigma_i}, \tag{5.3}$$

where $\lambda_i(\cdot\,;k)$, $\mu_i(\cdot\,;k) \in L^2(S_i)$, $\tau_i(\cdot\,;k) \in L^2(\gamma_i)$ and $\sigma_i(k) \in V_i'$ is defined by

$$(\sigma_i(k), \phi)_{V_i'} = \int_{\gamma_i} \tau_i(\cdot\,;k)\phi(\cdot)\, d\Gamma \quad \forall \phi \in V_i.$$

The system (5.1)–(5.3) is the optimality system for the problem

$$\begin{aligned} J_i^k(f_1, f_2) = \frac{1}{2}\int_{\Sigma_i} |f_2|^2 d\Sigma + \frac{1}{2\beta}\int_{S_i} (|f_1|^2 + |\beta z_i(\cdot\,;k) + \mu_i(\cdot\,;k)|^2) d\Sigma \\ + \frac{k}{2}(\|z_i(T;k) - y_{iT}\|_{H_i}^2 + \|z_{i,t}(T;k) - \dot{y}_{iT} + k^{-1}\sigma_i(k)\|_{V_i'}^2) \to \inf \end{aligned}$$

subject to

$$\begin{aligned} z_{i,tt} - a_i \Delta z_i &= 0 \quad \text{in } Q_i, \\ a_i \frac{\partial z_i}{\partial \nu_i} &= f_1 + \lambda_i(\cdot\,;k) \quad \text{on } S_i, \\ a_i \frac{\partial z_i}{\partial \nu_i} &= f_2 \quad \text{on } \Sigma_i, \\ z_i(0) = z_{i,t}(0) &= 0 \quad \text{in } \Omega_i, \end{aligned} \tag{5.4}$$

where the infimum is taken over $f_1 \in L^2(S_i)$ and $f_2 \in L^2(\Sigma_i)$.

Let ϕ_i be the solution of

$$\begin{aligned} \phi_{i,tt} - a_i \Delta \phi_i &= 0 \quad \text{in } Q_i, \\ a_i \frac{\partial \phi_i}{\partial \nu_i} &= 0 \quad \text{on } S_i \cup \Sigma_i, \\ \phi_i(0) = \phi_i^0, &\quad \phi_{i,t}(0) = \phi_i^1. \end{aligned} \tag{5.5}$$

It is assumed that (5.5) is continuously observable for $T > T_0$ for some $T_0 > 0$:

(5.6)
$$\|\phi_i^0\|_{H_i}^2 + \|\phi_i^1\|_{V_i'}^2 \le C_T \left(\int_{S_i} |\phi_i|^2 d\Sigma + \int_{\Sigma_i} |\phi_i|^2 d\Sigma \right) \quad \forall (\phi_i^0, \phi_i^1) \in V_i \times H_i,$$

for some constant C_T. It is known that under mild conditions on Γ and γ_i (for example, if they are sufficiently smooth), (5.6) holds for $T > (2/\sqrt{a_i})\mathrm{diam}(\Omega_i)$. It follows from (5.6) that if $(y_{iT}, \dot{y}_{iT}) \in V_i \times H_i$ there are controls $f_1 \in L^2(S_i)$, $f_2 \in L^2(\Sigma_i)$ such that the solution of (5.4) satisfies

$$z_i(T;k) = y_{iT}, \quad z_{i,t}(T;k) = \dot{y}_{iT}. \tag{5.7}$$

THEOREM 5.1. *Suppose that* (5.6) *holds,* $T > T_0$, *and let* $(y_{iT}, \dot{y}_{iT}) \in V_i \times H_i$. *Suppose further that as* $k \to \infty$

$$\begin{gathered} \lambda_i(\cdot\,;k) \to \lambda_i^\infty(\cdot), \quad \mu_i(\cdot\,;k) \to \mu_i^\infty(\cdot) \quad \textit{strongly in } L^2(S_i), \\ \left\{ \frac{1}{\sqrt{k}} \tau_i(\cdot\,;k) \right\} \quad \textit{is bounded in } L^2(\gamma_i). \end{gathered} \tag{5.8}$$

Then the solution of (5.1)–(5.3) *satisfies*

$$\begin{gathered} z_i(\cdot\,;k) \to z_i(\cdot), \quad q_i(\cdot\,;k) \to q_i(\cdot) \quad \textit{in } C([0,T];H_i), \\ z_{i,t}(\cdot\,;k) \to z_{i,t}(\cdot), \quad q_{i,t}(\cdot\,;k) \to q_{i,t}(\cdot) \quad \textit{in } C([0,T];V_i'), \\ z_i(\cdot\,;k)|_{S_i} \to z_i(\cdot)|_{S_i}, \quad q_i(\cdot\,;k)|_{S_i} \to q_i(\cdot)|_{S_i} \quad \textit{strongly in } L^2(S_i), \\ q_i(\cdot\,;k)|_{\Sigma_i} \to q_i(\cdot)|_{\Sigma_i} \quad \textit{strongly in } L^2(\Sigma_i), \end{gathered}$$

where z_i, q_i *are the solution of*

$$\begin{gathered} z_{i,tt} - a_i \Delta z_i = 0 \quad \textit{in } Q_i, \\ a_i \frac{\partial z_i}{\partial \nu_i} = -\beta q_i + \lambda_i^\infty \quad \textit{on } S_i, \\ a_i \frac{\partial z_i}{\partial \nu_i} = -q_i \quad \textit{on } \Sigma_i, \\ z_i(0) = z_{i,t}(0) = 0, \qquad z_i(T) = y_{iT}, \;\; z_{i,t}(T) = \dot{y}_{iT}, \end{gathered} \tag{5.9}$$

$$\begin{gathered} q_{i,tt} - a_i \Delta q_i = 0 \quad \textit{in } Q_i, \\ a_i \frac{\partial q_i}{\partial \nu_i} = \beta z_i + \mu_i^\infty \quad \textit{on } S_i, \\ a_i \frac{\partial q_i}{\partial \nu_i} = 0 \quad \textit{on } \Sigma_i, \\ q_i(T) = q_{iT}, \quad q_{i,t}(T) = \dot{q}_{iT}, \end{gathered} \tag{5.10}$$

and where $(q_{iT}, \dot{q}_{iT}) \in H_i \times V_i'$ *is the unique solution of*

$$\begin{aligned} &((q_{iT}, \dot{q}_{iT}), (-\dot{y}_{iT}, y_{iT}))_{H_i \times V_i'} \\ &\quad = \int_{S_i} [\beta(|q_i|^2 + |z_i|^2) - \lambda_i^\infty q_i + \mu_i^\infty z_i] \, d\Sigma + \int_{\Sigma_i} |q_i|^2 d\Sigma. \end{aligned} \tag{5.11}$$

Proof. Let $T > T_0$ and $f_1^0 \in L^2(S_i)$, $f_2^0 \in L^2(\Sigma_i)$ be such that the solution $\hat{z}_i$ of

$$\begin{aligned} \hat{z}_{i,tt} - a_i \Delta \hat{z}_i &= 0 \quad \text{in } Q_i, \\ a_i \frac{\partial \hat{z}_i}{\partial \nu_i} &= f_1^0 \quad \text{on } S_i, \\ a_i \frac{\partial \hat{z}_i}{\partial \nu_i} &= f_2^0 \quad \text{on } \Sigma_i, \\ \hat{z}_i(0) = \hat{z}_{i,t}(0) &= 0 \quad \text{in } \Omega_i \end{aligned}$$

satisfies (5.7), and set $\hat{f}_1^k(\cdot\,;k) = f_1^0(\cdot) - \lambda(\cdot\,;k)$. Then

$$\begin{aligned} J_i^k(f_1^k, f_2^k) \le J_i^k(\hat{f}_1^k, f_2^0) = \frac{1}{2}\|f_2^0\|^2_{L^2(\Sigma_i)} + \frac{1}{2\beta}\|\hat{f}_1^k\|_{L^2(S_i)} \\ + \frac{1}{2\beta}\int_{S_i} |\beta \hat{z}_i(\cdot) + \mu_i(\cdot\,;k)|^2 d\Sigma + \frac{1}{2k}\|\sigma_i(k)\|^2_{V_i'}. \end{aligned}$$

It follows that

$$\begin{aligned} \{f_1^k\}, \quad \{z_i(\cdot\,;k)|_{S_i}\} \quad &\text{are bounded in } L^2(S_i), \\ \{f_2^k\} \quad &\text{is bounded in } L^2(\Sigma_i), \\ \sqrt{k}(z_i(T;k) - y_{iT}) \quad &\text{is bounded in } H_i, \\ \sqrt{k}(z_{i,t}(T;k) - \dot{y}_{iT} + k^{-1}\sigma_i(k)) \quad &\text{is bounded in } V_i'. \end{aligned}$$

Therefore

$$\begin{aligned} z_i(T;k) &\to y_{iT} \quad \text{strongly in } H_i, \\ z_{i,t}(T;k) &\to \dot{y}_{iT} \quad \text{strongly in } V_i', \end{aligned} \tag{5.12}$$

and on a certain subsequence

$$\begin{aligned} f_1^k &\to f_1^\infty \quad \text{weakly in } L^2(S_i), \\ f_2^k &\to f_2^\infty \quad \text{weakly in } L^2(\Sigma_i), \\ z_i(\cdot\,;k)|_{S_i} &\to z_i^\infty(\cdot) \quad \text{weakly in } L^2(S_i). \end{aligned} \tag{5.13}$$

It follows from (5.8), (5.13), the regularity results of [20], and the compactness result [32] utilized above that

$$\begin{aligned} z_i(\cdot\,;k) &\to z_i(\cdot) \quad \text{in } C([0,T]; L^2(\Omega_i)), \\ z_{i,t}(\cdot\,;k) &\to z_{i,t}(\cdot) \quad \text{in } C([0,T]; (H^1(\Omega_i))'), \end{aligned} \tag{5.14}$$

where z_i satisfies

$$\begin{aligned} z_{i,tt} - a_i \Delta z_i &= 0 \quad \text{in } Q_i, \\ a_i \frac{\partial z_i}{\partial \nu_i} &= f_1^\infty + \lambda_i^\infty \quad \text{on } S_i, \\ a_i \frac{\partial z_i}{\partial \nu_i} &= f_2^\infty \quad \text{on } \Sigma_i, \\ z_i(0) = z_{i,t}(0) = 0, \qquad z_i(T) &= y_{iT}, \;\; z_{i,t}(T) = \dot{y}_{iT}. \end{aligned} \tag{5.15}$$

We have

$$\begin{aligned}(5.16)\quad &\frac{1}{\beta}\|f_1^\infty\|^2_{L^2(S_i)} + \|f_2^\infty\|^2_{L^2(\Sigma_i)} + \frac{1}{\beta}\|\beta z_i^\infty + \mu_i^\infty\|^2_{L^2(S_i)} \\ &\quad \le \liminf\Big(\frac{1}{\beta}\|f_1^k\|^2_{L^2(S_i)} + \|f_2^k\|^2_{L^2(\Sigma_i)} + \frac{1}{\beta}\|\beta z_i(\cdot\,;k) + \mu_i^k(\cdot)\|^2_{L^2(S_i)}\Big) \\ &\quad \le 2\liminf J_i^k(f_1^k, f_2^k).\end{aligned}$$

Set

$$\begin{aligned}\mathcal{U}_{\rm ad} = \{&f_1 \in L^2(S_i),\ f_2 \in L^2(\Sigma_i) : \\ &\text{the solution of (2.1) with } \lambda_i = \lambda_i^\infty \text{ satisfies (5.7)}\}.\end{aligned}$$

For all $f_1, f_2 \in \mathcal{U}_{\rm ad}$ we have

$$\begin{aligned}J_i^k(f_1^k, f_2^k) \le J_i^k(f_1, f_2) &= \frac{1}{2}\|f_2\|^2_{L^2(\Sigma_i)} + \frac{1}{2\beta}\|f_1\|^2_{L^2(S_i)} \\ &\quad + \frac{1}{2\beta}\int_{S_i} |\beta z_i(\cdot\,;k) + \mu_i^k(\cdot)|^2 d\Sigma + \frac{1}{2k}\|\sigma_i(k)\|^2_{V_i'},\end{aligned}$$

where $z_i(\cdot\,;k)$ is the solution of (5.4). Since $\lambda_i(\cdot\,;k) \to \lambda_i^\infty(\cdot)$ strongly in $L^2(S_i)$,

$$z_i(\cdot\,;k)|_{S_i} \to z_i(\cdot)|_{S_i} \quad \text{strongly in } L^2(S_i),$$

where $z_i(\cdot)$ is the solution of (2.1) with $\lambda_i = \lambda_i^\infty$. Therefore

$$\begin{aligned}(5.17)\qquad \limsup J_i^k(f_1^k, f_2^k) &\le \frac{1}{2}\|f_2\|^2_{L^2(\Sigma_i)} + \frac{1}{2\beta}\|f_1\|^2_{L^2(S_i)} \\ &\quad + \frac{1}{2\beta}\int_{S_i} |\beta z_i + \mu_i^\infty|^2 d\Sigma \quad \forall (f_1, f_2) \in \mathcal{U}_{\rm ad}.\end{aligned}$$

It follows from (5.13), (5.16), and (5.17) that as $k \to \infty$,

$$\begin{aligned}(5.18)\qquad f_1^k &\to f_1^\infty \quad \text{strongly in } L^2(S_i), \\ f_2^k &\to f_2^\infty \quad \text{strongly in } L^2(\Sigma_i), \\ z_i(\cdot\,;k)|_{S_i} &\to z_i^\infty \quad \text{strongly in } L^2(S_i)\end{aligned}$$

and that

$$J_i(f_1^\infty, f_2^\infty) \le J_i(f_1, f_2) \quad \forall (f_1, f_2) \in \mathcal{U}_{\rm ad},$$

where

$$J_i(f_1, f_2) = \frac{1}{2}\|f_2\|^2_{L^2(\Sigma_i)} + \frac{1}{2\beta}\|f_1\|^2_{L^2(S_i)} + \frac{1}{2\beta}\int_{S_i} |\beta z_i + \mu_i^\infty|^2 d\Sigma$$

and where z_i satisfies

$$\begin{aligned}z_{i,tt} - a_i \Delta z_i &= 0 \quad \text{in } Q_i, \\ a_i \frac{\partial z_i}{\partial \nu_i} &= f_1 + \lambda_i^\infty \quad \text{on } S_i, \\ a_i \frac{\partial z_i}{\partial \nu_i} &= f_2 \quad \text{on } \Sigma_i, \\ z_i(0) = z_{i,t}(0) = 0 \quad z_i(T) &= y_{iT}, \quad z_{i,t}(T) = \dot{y}_{iT}.\end{aligned}$$

The solution $q_i(\cdot\,;k)$ of (5.2) may be expressed as $q_i = r_i + s_i$, where r_i satisfies (5.2) except that

$$r_i(T;k) = r_{i,t}(T;k) = 0,$$

and s_i satisfies (5.2) except that

$$a_i \frac{\partial s_i}{\partial \nu_i} = 0 \quad \text{on } S_i.$$

By virtue of the regularity results in [20] we have

$$\|r_i(\cdot\,;k)\|^2_{L^2(\Sigma_i)} + \|r_i(\cdot\,;k)\|^2_{L^2(S_i)} \le C_T \int_{S_i} |\beta z_i(\cdot\,;k) + \mu_i(\cdot\,;k)|^2 d\Sigma, \tag{5.19}$$

and by virtue of the observability assumption (5.6) we have, for $T > T_0$,

$$\|q_i(T;k)\|^2_{H_i} + \|q_{i,t}(T;k)\|^2_{V_i'} \le C_T \left(\int_{S_i} |s_i|^2 d\Sigma + \int_{\Sigma_i} |s_i|^2 d\Sigma \right). \tag{5.20}$$

It follows from (5.19), (5.20) that

$$\begin{aligned} &\|q_i(T;k)\|^2_{H_i} + \|q_{i,t}(T;k)\|^2_{V_i'} \\ &\quad \le C_T \left(\int_{S_i} (|q_i(\cdot\,;k)|^2 + |\beta z_i(\cdot\,;k) + \mu_i(\cdot\,;k)|^2) d\Sigma + \int_{\Sigma_i} |q_i(\cdot\,;k)|^2 d\Sigma \right) \\ &\quad = C_T \left(\int_{S_i} (|f_1^k|^2 + |\beta z_i(\cdot\,;k) + \mu_i(\cdot\,;k)|^2) d\Sigma + \int_{\Sigma_i} |f_2^k|^2 d\Sigma \right). \end{aligned} \tag{5.21}$$

From (5.8), (5.18), and (5.21) we may conclude that

$$\begin{aligned} q_i(T;k) &\to q_{iT} \quad \text{strongly in } H_i, \\ q_{i,t}(T;k) &\to \dot{q}_{iT} \quad \text{strongly in } V_i' \end{aligned}$$

for some $(q_{iT}, \dot{q}_{iT}) \in H_i \times V_i'$ and then that

$$\begin{aligned} q_i(\cdot\,;k) &\to q_i(\cdot) \quad \text{in } C([0,T];H_i), \\ q_{i,t}(\cdot\,;k) &\to q_{i,t}(\cdot) \quad \text{in } C([0,T];V_i'), \end{aligned}$$

where q_i is the solution of

$$\begin{aligned} q_{i,tt} - a_i \Delta q_i &= 0 \quad \text{in } Q_i, \\ a_i \frac{\partial q_i}{\partial \nu_i} &= \beta z_i + \mu_i^\infty \quad \text{on } S_i, \\ a_i \frac{\partial q_i}{\partial \nu_i} &= 0 \quad \text{on } \Sigma_i, \\ q_i(T) = q_{iT}, &\quad q_{i,t}(T) = \dot{q}_{iT}. \end{aligned}$$

Further,

$$\begin{aligned} \beta q_i(\cdot\,;k)|_{S_i} = -f_1^k &\to -f_1^\infty \quad \text{strongly in } L^2(S_i), \\ q_i(\cdot\,;k)|_{\Sigma_i} = -f_2^k &\to -f_2^\infty \quad \text{strongly in } L^2(\Sigma_i), \end{aligned}$$

so, by definition,

$$\beta q_i|_{S_i} = -f_1^\infty, \quad q_i|_{\Sigma_i} = -f_2^\infty.$$

A formal integration by parts (which may be justified by approximating f_1^∞, f_2^∞, μ_i^∞, λ_i^∞ by smoother data) gives

$$0 = (\dot{y}_{iT}, q_{iT})_{H_i} - (\dot{q}_{iT}, y_{iT})_{V_i'} - \int_{S_i} \left(q_i a_i \frac{\partial z_i}{\partial \nu_i} - z_i a_i \frac{\partial q_i}{\partial \nu_i} \right) d\Sigma - \int_{\Sigma_i} q_i a_i \frac{\partial z_i}{\partial \nu_i}\, d\Sigma,$$

that is,

$$\begin{aligned} (5.22) \qquad & ((q_{iT}, \dot{q}_{iT}), (-\dot{y}_{iT}, y_{iT}))_{H_i \times V_i'} \\ & = \int_{S_i} [\beta(|q_i|^2 + |z_i|^2) - \lambda_i^\infty q_i + \mu_i^\infty z_i] d\Sigma + \int_{\Sigma_i} |q_i|^2 d\Sigma. \qquad \square \end{aligned}$$

Remark 5.2. The above analysis shows that (5.9), (5.10), (5.11) is the optimality system for the problem $\inf_{(f_1,f_2)\in\mathcal{U}_{\text{ad}}} J_i(f_1, f_2)$.

Let us comment further on (5.22). The pair $(q_{iT}, \dot{q}_{iT})$ may be calculated in a manner reminiscent of the Hilbert uniqueness method. In fact, this pair is chosen so that the solution of

$$\begin{aligned} (5.23) \qquad z_{i,tt} - a_i \Delta z_i &= 0 \quad \text{in } Q_i, \\ a_i \frac{\partial z_i}{\partial \nu_i} &= -\beta q_i + \lambda_i^\infty \quad \text{on } S_i, \\ a_i \frac{\partial z_i}{\partial \nu_i} &= -q_i \quad \text{on } \Sigma_i, \\ z_i(0) = z_{i,t}(0) &= 0, \end{aligned}$$

$$\begin{aligned} (5.24) \qquad q_{i,tt} - a_i \Delta q_i &= 0 \quad \text{in } Q_i, \\ a_i \frac{\partial q_i}{\partial \nu_i} &= \beta z_i + \mu_i^\infty \quad \text{on } S_i, \\ a_i \frac{\partial q_i}{\partial \nu_i} &= 0 \quad \text{on } \Sigma_i, \\ q_i(T) = q_{iT}, \quad & q_{i,t}(T) = \dot{q}_{iT} \end{aligned}$$

satisfies

$$(5.25) \qquad z_i(T) = y_{iT}, \qquad z_{i,t}(T) = \dot{y}_{iT}.$$

For arbitrary $(q_{iT}, \dot{q}_{iT}) \in V_i \times H_i$, the system (5.23), (5.24) has a unique solution since it is the optimality system for the problem

$$\inf_{f\in L^2(S_i)} \left(\frac{1}{2\beta} \int_{S_i} (|f|^2 + |\beta q_i - \lambda_i^\infty|^2)\, d\Sigma + \frac{1}{2} \int_{\Sigma_i} |q_i|^2 d\Sigma \right)$$

subject to

$$\begin{aligned} q_{i,tt} - a_i \Delta q_i &= 0 \quad \text{in } Q_i, \\ a_i \frac{\partial q_i}{\partial \nu_i} &= f + \mu_i^\infty \quad \text{on } S_i, \\ a_i \frac{\partial q_i}{\partial \nu_i} &= 0 \quad \text{on } \Sigma_i, \\ q_i(T) = q_{iT}, \quad & q_{i,t}(T) = \dot{q}_{iT}. \end{aligned}$$

The solution of (5.23), (5.24) may be written $z_i = z_i^1 + z_i^2$, $q_i = q_i^1 + q_i^2$, where

$$(z_i^1, q_i^1) \text{ is the solution corresponding to } \lambda_i^\infty = \mu_i^\infty = 0,$$
$$(z_i^2, q_i^2) \text{ is the solution corresponding to } q_{iT} = \dot{q}_{iT} = 0.$$

By Holmgren's theorem, we may define a norm $\|(q_{iT}, \dot{q}_{iT})\|_{F_i}$ by setting

$$\|(q_{iT}, \dot{q}_{iT})\|_{F_i}^2 := \int_{S_i} \beta(|q_i^1|^2 + |z_i^1|^2) d\Sigma + \int_{\Sigma_i} |q_i^1|^2 d\Sigma,$$

and a corresponding Hilbert space

$$F_i = \text{ completion of } V_i \times H_i \text{ in } \| \cdot \|_{F_i}.$$

For $T > T_0$, one has (cf. (5.19)–(5.21))

$$\|(q_{iT}, \dot{q}_{iT})\|_{H_i \times V_i'} \le C_T \|(q_{iT}, \dot{q}_{iT})\|_{F_i};$$

hence $F_i \subset H_i \times V_i'$, $H_i \times V_i \subset F_i'$. In addition, by Green's formula we have

$$((q_{iT}, \dot{q}_{iT}), (-z_{i,t}^1(T), z_i^1(T))) = \|(q_{iT}, \dot{q}_{iT})\|_{F_i}^2.$$

Therefore, if $(-\dot{z}_{iT}, z_{iT}) \in F_i'$ and if we set

$$(q_{iT}, \dot{q}_{iT}) = \Lambda_i^{-1}(-\dot{z}_{iT}, z_{iT}),$$

where Λ_i is the canonical isomorphism $F_i \mapsto F_i'$, then we will have

$$z_i^1(T) = z_{iT}, \qquad z_{i,t}^1(T) = \dot{z}_{iT}.$$

It follows that (5.25) will be satisfied by choosing

$$(q_{iT}, \dot{q}_{iT}) = \Lambda_i^{-1}((-\dot{y}_{iT}, y_{iT}) - (-z_{i,t}^2(T), z_i^2(T))). \tag{5.26}$$

It may be checked that (5.26) is exactly the same as (5.22).

5.1. Application of Theorem 5.1 to domain decomposition. We now apply Theorem 5.1 to the solution $y_i^{n+1}(\cdot\,; k)$, $p_i^{n+1}(\cdot\,; k)$ of the optimality system (3.1), (3.2), where $\lambda_{ij}^n(\cdot\,; k)$, $\mu_{ij}^n(\cdot\,; k)$, $\tau_{ij}^n(\cdot\,; k)$ are given by (3.3), (3.5) and $\sigma_i^n(k)$ by (3.4). We assume that

$$\lambda_{ij}^0, \quad \mu_{ij}^0, \quad \tau_{ij}^0 \quad \text{are independent of } k. \tag{5.27}$$

Then, as $k \to \infty$ the solution $y_i^1(\cdot\,; k)$, $p_i^1(\cdot\,; k)$ converges in the manner described in Theorem 5.1 to the solution $y_i^1(\cdot)$, $p_i^1(\cdot)$ of the system

$$y_{i,tt}^1 - a_i \Delta y_i^1 = 0, \quad p_{i,tt}^1 - a_i \Delta p_i^1 = 0 \quad \text{in } Q_i,$$
$$a_i \frac{\partial y_i^1}{\partial \nu_i} + \beta p_i^1 = \lambda_{ij}^0, \quad a_i \frac{\partial p_i^1}{\partial \nu_i} - \beta y_i^1 = \mu_{ij}^0 \quad \text{on } \Sigma_{ij},$$
$$a_i \frac{\partial y_i^1}{\partial \nu_i} + p_i^1 = 0, \quad a_i \frac{\partial p_i^1}{\partial \nu_i} = 0 \quad \text{on } \Sigma_i,$$
$$y_i^1(0) = y_{i,t}^1(0) = 0, \quad p_i^1(T) = p_{iT}^1, \quad p_{i,t}(T) = \dot{p}_{iT}^1,$$
$$y_i^1(T) = y_{iT}, \quad y_{i,t}^1(T) = \dot{y}_{iT},$$

where $(p^1_{iT}, \dot{p}^1_{iT}) \in H_i \times V_i'$ is the solution of

$$((p^1_{iT}, \dot{p}^1_{iT}), (-\dot{y}_{iT}, y_{iT}))_{H_i \times V_i'}$$
$$= \sum_{j:\Gamma_{ij} \neq \emptyset} \int_{\Sigma_{ij}} [\beta(|p^1_i|^2 + |y^1_i|^2) d\Sigma - \lambda^0_{ij} p^1_i + \mu^0_{ij} y^1_i] d\Sigma + \int_{\Sigma_i} |p^1_i|^2 d\Sigma.$$

According to Theorem 5.1,

$$y^1_i(\cdot\,; k)|_{S_i} \to y^1_i(\cdot)|_{S_i}, \quad p^1_i(\cdot\,; k)|_{S_i} \to p^1_i(\cdot)|_{S_i} \quad \text{strongly in } L^2(S_i),$$

and therefore

$$a_i \frac{\partial y^1_i(\cdot\,; k)}{\partial \nu_i} \to \lambda^0_{ij}(\cdot) - \beta p^1_i(\cdot) = a_i \frac{\partial y^1_i(\cdot)}{\partial \nu_i},$$
$$a_i \frac{\partial p^1_i(\cdot\,; k)}{\partial \nu_i} \to \mu^0_{ij}(\cdot) + \beta y^1_i(\cdot) = a_i \frac{\partial p^1_i(\cdot)}{\partial \nu_i} \quad \text{strongly in } L^2(\Sigma_{ij}).$$

As a result,

$$\begin{aligned} \lambda^1_{ij}(\cdot\,; k) &\to (1-\epsilon)\left(- a_j \frac{\partial y^1_j(\cdot)}{\partial \nu_j} + \beta p^1_j(\cdot)\right)|_{\Sigma_{ij}} + \epsilon \left(a_i \frac{\partial y^1_i(\cdot)}{\partial \nu_i} + \beta p^1_i(\cdot)\right)|_{\Sigma_{ij}} \\ &:= \lambda^1_{ij}(\cdot) \quad \text{strongly in } L^2(\Sigma_{ij}), \\ \mu^1_{ij}(\cdot\,; k) &\to (1-\epsilon)\left(- a_j \frac{\partial p^1_j(\cdot)}{\partial \nu_j} - \beta y^1_j(\cdot)\right)|_{\Sigma_{ij}} + \epsilon \left(a_i \frac{\partial p^1_i(\cdot)}{\partial \nu_i} - \beta y^1_i(\cdot)\right)|_{\Sigma_{ij}} \\ &:= \mu^1_{ij}(\cdot) \quad \text{strongly in } L^2(\Sigma_{ij}). \end{aligned} \tag{5.28}$$

The proof of Theorem 5.1 showed that

$$\sqrt{k}(y^1_{i,t}(T; k) - \dot{y}_{iT} + k^{-1}\sigma^0_i(k))$$

is bounded in V_i' and therefore

$$\frac{1}{\sqrt{k}} p^1_i(T; k) = \mathcal{A}_i^{-1}[\sqrt{k}(y^1_{i,t}(T; k) - \dot{y}_{iT} + k^{-1}\sigma^0_i(k))] \quad \text{is bounded in } V_i.$$

In particular,

$$\frac{1}{\sqrt{k}} p^1_i(T; k)|_{\gamma_i} \quad \text{is bounded in } L^2(\gamma_i).$$

Since (see Remark 2.1)

$$a_i \frac{\partial p^1_i(T; k)}{\partial \nu_i} + \alpha p^1_i(T; k) = \tau^0_{ij}(\cdot) \quad \text{on } \Gamma_{ij},$$

it follows that

$$\frac{1}{\sqrt{k}} a_i \frac{\partial p^1_i(T; k)}{\partial \nu_i} \quad \text{is bounded in } L^2(\gamma_i);$$

hence

$$\begin{aligned} \frac{1}{\sqrt{k}} \tau^1_{ij}(\cdot\,; k) = {} & \frac{1-\epsilon}{\sqrt{k}} \left(-a_j \frac{\partial p^1_j(T; k)}{\partial \nu_j} + \alpha p^1_j(T; k)\right)\Big|_{\Gamma_{ij}} \\ & + \frac{\epsilon}{\sqrt{k}} \left(a_i \frac{\partial p^1_i(T; k)}{\partial \nu_i} + \alpha p^1_i(T; k)\right)\Big|_{\Gamma_{ij}} \quad \text{is bounded in } L^2(\Gamma_{ij}). \end{aligned}$$

We may now apply Theorem 5.1 to conclude that as $k \to \infty$ the solution $y_i^2(\cdot\,;k)$, $p_i^2(\cdot\,;k)$ of (3.1), (3.2) with $n+1=2$ converges in the manner described in that theorem to the solution $y_i^2(\cdot)$, $p_i^2(\cdot)$ of

$$\begin{gathered}
y_{i,tt}^2 - a_i \Delta y_i^2 = 0, \quad p_{i,tt}^2 - a_i \Delta p_i^2 = 0 \quad \text{in } Q_i, \\
a_i \frac{\partial y_i^2}{\partial \nu_i} + \beta p_i^2 = \lambda_{ij}^1, \quad a_i \frac{\partial p_i^2}{\partial \nu_i} - \beta y_i^2 = \mu_{ij}^1 \quad \text{on } \Sigma_{ij}, \\
a_i \frac{\partial y_i^2}{\partial \nu_i} + p_i^2 = 0, \quad a_i \frac{\partial p_i^2}{\partial \nu_i} = 0 \quad \text{on } \Sigma_i, \\
y_i^2(0) = y_{i,t}^2(0) = 0, \quad p_i^2(T) = p_{iT}^2, \quad p_{i,t}(T) = \dot{p}_{iT}^2, \\
y_i^2(T) = y_{iT}, \quad y_{i,t}^2(T) = \dot{y}_{iT},
\end{gathered}$$

where λ_{ij}^1, μ_{ij}^1 are given by (5.28) and where $(p_{iT}^2, \dot{p}_{iT}^2)$ is the solution of

$$\begin{aligned}
&((p_{iT}^2, \dot{p}_{iT}^2), (-\dot{y}_{iT}, y_{iT}))_{H_i \times V_i'} \\
&\quad = \sum_{j:\Gamma_{ij} \neq \emptyset} \int_{\Sigma_{ij}} [\beta(|p_i^2|^2 + |y_i^2|^2) d\Sigma - \lambda_{ij}^1 p_i^2 + \mu_{ij}^1 y_i^2] d\Sigma + \int_{\Sigma_i} |p_i^2|^2 d\Sigma.
\end{aligned}$$

One may now proceed inductively to obtain the following result.

THEOREM 5.3. *Assume* (5.6) *and* (5.27), *and let* $(y_{iT}, \dot{y}_{iT}) \in V_i \times H_i$. *Then as* $k \to \infty$ *the solutions* $\{y_i^{n+1}(\cdot\,;k), p_i^{n+1}(\cdot\,;k)\}_{n=0}^{\infty}$ *converge in the sense of Theorem* 5.1 *to the solutions* $\{y_i^{n+1}(\cdot), p_i^{n+1}(\cdot)\}_{n=0}^{\infty}$ *of*

$$\begin{gathered}
y_{i,tt}^{n+1} - a_i \Delta y_i^{n+1} = 0 \quad \textit{in } Q_i, \\
a_i \frac{\partial y_i^{n+1}}{\partial \nu_i} + \beta p_i^{n+1} = \lambda_{ij}^n \quad \textit{on } \Sigma_{ij}, \\
a_i \frac{\partial y_i^{n+1}}{\partial \nu_i} + p_i^{n+1} = 0 \quad \textit{on } \Sigma_i, \\
y_i^{n+1}(0) = y_{i,t}^{n+1}(0) = 0,
\end{gathered} \tag{5.29}$$

$$\begin{gathered}
p_{i,tt}^{n+1} - a_i \Delta p_i^{n+1} = 0 \quad \textit{in } Q_i, \\
a_i \frac{\partial p_i^{n+1}}{\partial \nu_i} - \beta y_i^{n+1} = \mu_{ij}^n \quad \textit{on } \Sigma_{ij}, \\
a_i \frac{\partial p_i^{n+1}}{\partial \nu_i} = 0 \quad \textit{on } \Sigma_i, \\
p_i^{n+1}(T) = p_{iT}^{n+1}, \quad p_{i,t}^{n+1}(T) = \dot{p}_{iT}^{n+1},
\end{gathered} \tag{5.30}$$

$$y_i^{n+1}(T) = y_{iT}, \quad y_{i,t}^{n+1}(T) = \dot{y}_{iT}, \tag{5.31}$$

where

$$\begin{aligned}
\lambda_{ij}^n &= (1-\epsilon)\left(-a_j \frac{\partial y_j^n}{\partial \nu_j} + p_j^n\right)|_{\Sigma_{ij}} + \epsilon\left(a_i \frac{\partial y_i^n}{\partial \nu_i} + p_i^n\right)|_{\Sigma_{ij}}, \\
\mu_{ij}^n &= (1-\epsilon)\left(-a_j \frac{\partial p_j^n}{\partial \nu_j} - y_j^n\right)|_{\Sigma_{ij}} + \epsilon\left(a_i \frac{\partial p_i^n}{\partial \nu_i} - y_i^n\right)|_{\Sigma_{ij}},
\end{aligned} \tag{5.32}$$

and where $(p_{iT}^{n+1}, \dot{p}_{iT}^{n+1}) \in H_i \times V_i'$ *is the solution of*

$$((p_{iT}^{n+1}, \dot{p}_{iT}^{n+1}), (-\dot{y}_{iT}, y_{iT}))_{H_i \times V_i'} = \int_{\Sigma_i} |p_i^{n+1}|^2 d\Sigma + \sum_{j:\Gamma_{ij} \neq \emptyset} \int_{\Sigma_{ij}} [\beta(|p_i^{n+1}|^2 + |y_i^{n+1}|^2) d\Sigma - \lambda_{ij}^n p_i^{n+1} + \mu_{ij}^n y_i^{n+1}] d\Sigma. \tag{5.33}$$

6. Convergence to the solution of the global optimality system. In this section it is proved that the solution of (5.29)–(5.33) converges as $n \to \infty$ to the solution of (4.4), (4.5). The proof follows the same lines as the proof of Theorem 3.1. First we introduce the space

$$\mathcal{X} := \prod_{i=1}^{m} L^2(S_i) \times L^2(S_i)$$

with norm $\|\cdot\|_{\mathcal{X}}$ given by

$$\|X\|_{\mathcal{X}}^2 = \sum_{i=1}^{m} \frac{1}{\beta} \int_{S_i} (|\lambda_i|^2 + |\mu_i|^2) d\Sigma,$$

where $X = \{(\lambda_i, \mu_i) : i = 1, \dots, m\}$. For $X \in \mathcal{X}$ define a linear mapping $\mathcal{T} : \mathcal{X} \mapsto \mathcal{X}$ as follows. Let (z_i, q_i) be the solution of

$$\begin{aligned} z_{i,tt} - a_i \Delta z_i &= 0 \quad \text{in } Q_i, \\ a_i \frac{\partial z_i}{\partial \nu_i} &= -\beta q_i + \lambda_i \quad \text{on } S_i, \\ a_i \frac{\partial z_i}{\partial \nu_i} &= -q_i \quad \text{on } \Sigma_i, \\ z_i(0) = z_{i,t}(0) &= 0, \end{aligned}$$

$$\begin{aligned} q_{i,tt} - a_i \Delta q_i &= 0 \quad \text{in } Q_i, \\ a_i \frac{\partial q_i}{\partial \nu_i} &= \beta z_i + \mu_i \quad \text{on } S_i, \\ a_i \frac{\partial q_i}{\partial \nu_i} &= 0 \quad \text{on } \Sigma_i, \\ q_i(T) = q_{iT}, &\quad q_{i,t}(T) = \dot{q}_{iT}, \end{aligned}$$

where $(q_{iT}, \dot{q}_{iT})$ is chosen so that (see (5.26))

$$z_i(T) = z_{i,t}(T) = 0. \tag{6.1}$$

Set

$$\mathcal{T}X = \left\{ \left(\left(-a_j \frac{\partial z_j}{\partial \nu_j} + \beta q_j \right) |_{\Sigma_{ij}}, \left(-a_j \frac{\partial q_j}{\partial \nu_j} - \beta z_j \right) |_{\Sigma_{ij}} \right) : i = 1, \dots, m;\ j : \Gamma_{ij} \neq \emptyset \right\}.$$

We note that $\mathcal{T}X = X$ if and only if the global transmission conditions on Σ_{ij} are satisfied for all i, j. Clearly $X = 0$ is a fixed point of $\mathcal{T}$. By the same calculation as in the proof of Lemma 2.1, we have

$$\|\mathcal{T}X\|_{\mathcal{X}}^2 = \|X\|_{\mathcal{X}}^2 - 4 \int_{\Sigma} |q_m|^2 d\Sigma, \tag{6.2}$$

so that $\mathcal{T}$ is nonexpansive.

We introduce λ_{ij}, μ_{ij}, $\tilde{y}_j^n$, $\tilde{p}_j^n$, $\tilde{\lambda}_{ij}^n$, $\tilde{\mu}_{ij}^n$ as above, and set

$$\tilde{p}_{iT}^n = p_{iT}^n - p_{iT}, \quad \tilde{q}_{iT}^n = \dot{p}_{iT}^n - \dot{p}_{iT},$$

where $p_{iT} = \mathbf{P}_T|_{\Omega_i} \in H_i$, $\dot{p}_{iT} = \mathcal{A}_i(A^{-1}\dot{\mathbf{P}}_T|_{\Omega_i}) \in V_i'$. Then $\tilde{y}_j^n$, $\tilde{p}_j^n$ satisfy

$$\begin{gathered}
\tilde{y}_{i,tt}^{n+1} - a_i\Delta\tilde{y}_i^{n+1} = 0, \quad \tilde{p}_{i,tt}^{n+1} - a_i\Delta\tilde{p}_i^{n+1} = 0 \quad \text{in } Q_i,\\
a_i\frac{\partial\tilde{y}_i^{n+1}}{\partial\nu_i} + \beta\tilde{p}_i^{n+1} = \tilde{\lambda}_{ij}^n, \quad a_i\frac{\partial\tilde{p}_i^{n+1}}{\partial\nu_i} - \beta\tilde{y}_i^{n+1} = \tilde{\mu}_{ij}^n \quad \text{on } \Sigma_{ij},\\
a_i\frac{\partial\tilde{y}_i^{n+1}}{\partial\nu_i} + \tilde{p}_i^{n+1} = 0, \quad a_i\frac{\partial\tilde{p}_i^{n+1}}{\partial\nu_i} = 0 \quad \text{on } \Sigma_i,\\
\tilde{y}_i^{n+1}(0) = \tilde{y}_{i,t}^{n+1}(0) = \tilde{y}_i^{n+1}(T) = \tilde{y}_{i,t}^{n+1}(T) = 0,\\
\tilde{p}_i^{n+1}(T) = \tilde{p}_{iT}^{n+1}, \quad \tilde{p}_{i,t}^{n+1}(T) = \tilde{q}_{iT}^{n+1}.
\end{gathered} \tag{6.3}$$

We observe that the iterations on Σ_{ij} may be expressed as the fixed point iteration

$$X^{n+1} = (1-\epsilon)\mathcal{T}X^n + \epsilon X^n.$$

According to Proposition 2.1, if $0 < \epsilon < 1$, then $X^n \to X$ weakly in $\mathcal{X}$, where X is a fixed point of $\mathcal{T}$.

THEOREM 6.1. *Assume that* (4.3) *and* (5.6) *hold,* $T > T_0$, $0 < \epsilon < 1$, *and let* $(\mathbf{Y}_T, \dot{\mathbf{Y}}_T) \in V \times H$. *Then as* $n \to \infty$ *we have*

$$\begin{gathered}
\tilde{y}_i^n \to 0 \quad \textit{in } C([0,T];H_i),\\
\tilde{y}_{i,t}^n \to 0 \quad \textit{in } C([0,T];V_i'),\\
\tilde{p}_i^n \to 0 \quad \textit{weakly}^* \textit{ in } L^\infty(0,T;H_i),\\
\tilde{p}_{i,t}^n \to 0 \quad \textit{weakly}^* \textit{ in } L^\infty(0,T;V_i'),\\
\tilde{p}_m^n|_\Sigma \to 0 \quad \textit{strongly in } L^2(\Sigma).
\end{gathered}$$

Proof. We proceed as in the proof of Theorem 3.1 and calculate with the aid of (6.2) that

$$\begin{aligned}
\|X^{n+1}\|_{\mathcal{X}}^2 &= (1-\epsilon)^2\|\mathcal{T}X^n\|_{\mathcal{X}}^2 + \epsilon^2\|X^n\|_{\mathcal{X}}^2 + 2\epsilon(1-\epsilon)(X^n, \mathcal{T}X^n)_{\mathcal{X}}\\
&= ((1-\epsilon)^2+\epsilon^2)\|X^n\|_{\mathcal{X}}^2 - 4(1-\epsilon)^2\int_\Sigma |\tilde{p}_m^n|^2 d\Sigma + 2\epsilon(1-\epsilon)(X^n, \mathcal{T}X^n)_{\mathcal{X}}.
\end{aligned} \tag{6.4}$$

Again, by the same calculation as in Lemma 2.2, we obtain

$$\begin{aligned}
\|X^n\|_{\mathcal{X}}^2 = \sum_{i=1}^m \frac{1}{\beta}\int_{S_i}\left(\beta^2(|\tilde{y}_i^n|^2 + |\tilde{p}_i^n|^2) + \left|a_i\frac{\partial\tilde{y}_i^n}{\partial\nu_i}\right|^2 + \left|a_i\frac{\partial\tilde{p}_i^n}{\partial\nu_i}\right|^2\right)d\Sigma\\
+ 2\int_\Sigma |\tilde{p}_m^n|^2 d\Sigma.
\end{aligned} \tag{6.5}$$

Therefore

$$\begin{aligned}
\|X^{n+1}\|_{\mathcal{X}}^2 = ((1-\epsilon)^2+\epsilon^2)E^n\\
- 2(1-2\epsilon)\int_\Sigma |\tilde{p}_m^n|^2 d\Sigma + 2\epsilon(1-\epsilon)(X^n, \mathcal{T}X^n)_{\mathcal{X}},
\end{aligned} \tag{6.6}$$

where

$$E^n := \sum_{i=1}^m \frac{1}{\beta} \int_{S_i} \left(\beta^2 (|\tilde{y}_i^n|^2 + |\tilde{p}_i^n|^2) + \left| a_i \frac{\partial \tilde{y}_i^n}{\partial \nu_i} \right|^2 + \left| a_i \frac{\partial \tilde{p}_i^n}{\partial \nu_i} \right|^2 \right) d\Sigma.$$

From (6.6) and (6.5) with n replaced by $n+1$ we obtain

$$\begin{aligned} E^{n+1} &= ((1-\epsilon)^2 + \epsilon^2) E^n \\ &\quad - 2 \int_\Sigma [(1-2\epsilon)|\tilde{p}_m^n|^2 + |\tilde{p}_m^{n+1}|^2] d\Sigma + 2\epsilon(1-\epsilon)(X^n, \mathcal{T} X^n)_{\mathcal{X}}. \end{aligned} \tag{6.7}$$

One has

$$\begin{aligned} (X^n, \mathcal{T} X^n)_{\mathcal{X}} = \sum_{i=1}^m \sum_{j:\Gamma_{ij} \neq \emptyset} \frac{1}{\beta} \int_{\Sigma_{ij}} \Bigg[& -a_i \frac{\partial \tilde{y}_i^n}{\partial \nu_i} a_j \frac{\partial \tilde{y}_j^n}{\partial \nu_j} - a_i \frac{\partial \tilde{p}_i^n}{\partial \nu_i} a_j \frac{\partial \tilde{p}_j^n}{\partial \nu_j} \\ & + \beta^2 (\tilde{p}_i^n \tilde{p}_j^n + \tilde{y}_i^n \tilde{y}_j^n) \Bigg] d\Sigma, \end{aligned}$$

so that

(6.8)

$$\begin{aligned} & ((1-\epsilon)^2 + \epsilon^2) E^n + 2\epsilon(1-\epsilon)(X^n, \mathcal{T} X^n)_{\mathcal{X}} \\ & = \sum_{i=1}^m \sum_{\substack{j:\Gamma_{ij} \neq \emptyset \\ j>i}} \frac{1}{\beta} \int_{\Sigma_{ij}} \Bigg[((1-\epsilon)^2 + \epsilon^2) \left(\left| a_i \frac{\partial \tilde{y}_i^n}{\partial \nu_i} \right|^2 + \left| a_j \frac{\partial \tilde{y}_j^n}{\partial \nu_j} \right|^2 \right. \\ & \quad \left. + \left| a_i \frac{\partial \tilde{p}_i^n}{\partial \nu_i} \right|^2 + \left| a_j \frac{\partial \tilde{p}_j^n}{\partial \nu_j} \right|^2 \right) - 4\epsilon(1-\epsilon) \left(a_i \frac{\partial \tilde{y}_i^n}{\partial \nu_i} a_j \frac{\partial \tilde{y}_j^n}{\partial \nu_j} + a_i \frac{\partial \tilde{p}_i^n}{\partial \nu_i} a_j \frac{\partial \tilde{p}_j^n}{\partial \nu_j} \right) \\ & \quad + \beta^2 ((1-\epsilon)^2 + \epsilon^2)(|\tilde{y}_i^n|^2 + |\tilde{y}_j^n|^2 + |\tilde{p}_i^n|^2 + |\tilde{p}_j^n|^2) + 4\beta^2 \epsilon(1-\epsilon)(\tilde{y}_i^n \tilde{y}_j^n + \tilde{p}_i^n \tilde{p}_j^n) \Bigg] d\Sigma \\ & \leq [(1-\epsilon)^2 + \epsilon^2 + 2\epsilon(1-\epsilon)] E^n = E^n. \end{aligned}$$

Substitution of (6.8) into (6.7) yields

$$E^{n+1} \leq E^n - 2 \int_\Sigma (|\tilde{p}_m^{n+1}|^2 + (1-2\epsilon)|\tilde{p}_m^n|^2)\, d\Sigma,$$

and so, by iteration,

$$E^{n+1} \leq E^1 - \sum_{\ell=1}^{n+1} \int_\Sigma c_\ell(\epsilon) |\tilde{p}_m^\ell|^2 d\Sigma, \tag{6.9}$$

where

$$c_1(\epsilon) = 1 - 2\epsilon, \quad c_{n+1}(\epsilon) = 1, \quad c_\ell(\epsilon) = 2(1-\epsilon),\ \ell = 2, \ldots, n.$$

In particular, for $0 \leq \epsilon < 1$ we have

$$\sum_{\ell=1}^\infty \int_\Sigma |\tilde{p}_m^\ell|^2 d\Sigma < \infty$$

so that

$$\tilde{p}_m^n|_\Sigma \to 0 \quad \text{strongly in } L^2(\Sigma) \text{ as } n \to \infty. \tag{6.10}$$

We also deduce from (6.9) that on a subsequence

$$\tilde{p}_i^\ell|_{S_i} \to \tilde{p}_i^\infty, \quad \tilde{y}_i^\ell|_{S_i} \to \tilde{y}_i^\infty, \quad a_i \frac{\partial \tilde{y}_i^\ell}{\partial \nu_i}\bigg|_{S_i} \to \tilde{Y}_i, \quad a_i \frac{\partial \tilde{p}_i^\ell}{\partial \nu_i}\bigg|_{S_i} \to \tilde{P}_i \tag{6.11}$$

weakly in $L^2(S_i)$. Since for $0 < \epsilon < 1$ we know that X^n converges weakly in $\mathcal{X}$ to a fixed point of $\mathcal{T}$, it follows from (6.3) that

$$\tilde{P}_i - \tilde{y}_i^\infty = -\tilde{P}_j - \tilde{y}_j^\infty, \quad \tilde{Y}_i + \tilde{p}_i^\infty = -\tilde{Y}_j + \tilde{p}_j^\infty \quad \text{on } \Sigma_{ij}$$

and then that

$$\tilde{Y}_i + \tilde{Y}_j = 0, \quad \tilde{P}_i + \tilde{P}_j = 0, \quad \tilde{y}_i^\infty = \tilde{y}_j^\infty, \quad \tilde{p}_i^\infty = \tilde{p}_j^\infty \quad \text{on } \Sigma_{ij}. \tag{6.12}$$

From (5.6) and (6.11) it is seen that

$$(\tilde{p}_{iT}^n, \tilde{q}_{iT}^n) \quad \text{is bounded in } H_i \times V_i'$$

and then that

$$(\tilde{p}_i^n, \tilde{p}_{i,t}^n) \quad \text{is bounded in } C([0,T]; H_i \times V_i').$$

Therefore, on a subsequence,

$$\begin{gathered}(\tilde{p}_{iT}^\ell, \tilde{q}_{iT}^\ell) \to (\tilde{p}_{iT}, \tilde{q}_{iT}) \quad \text{weakly in } H_i \times V_i', \\ \tilde{p}_i^\ell \to \tilde{p}_i \quad \text{weakly}^* \text{ in } L^\infty(0,T; H_i), \\ \tilde{p}_{i,t}^\ell \to \tilde{p}_{i,t} \quad \text{weakly}^* \text{ in } L^\infty(0,T; V_i').\end{gathered}$$

In addition, it follows as in the proof of Theorem 3.1 that on a subsequence

$$\begin{gathered}\tilde{y}_i^\ell \to \tilde{y}_i \quad \text{in } C([0,T]; H_i), \\ \tilde{y}_{i,t}^\ell \to \tilde{y}_{i,t} \quad \text{in } C([0,T]; V_i').\end{gathered}$$

The functions $\tilde{y}_i$, $\tilde{p}_i$ satisfy

$$\begin{gathered}\tilde{y}_{i,tt} - a_i \Delta \tilde{y}_i = 0, \quad \tilde{p}_{i,tt} - a_i \Delta \tilde{p}_i = 0 \quad \text{in } Q_i, \\ a_i \frac{\partial \tilde{y}_i}{\partial \nu_i} = -\tilde{Y}_i, \quad a_i \frac{\partial \tilde{p}_i}{\partial \nu_i} = -\tilde{P}_i \quad \text{on } S_i, \\ a_i \frac{\partial \tilde{y}_i}{\partial \nu_i} = 0, \quad a_i \frac{\partial \tilde{p}_i}{\partial \nu_i} = 0 \quad \text{on } \Sigma_i, \\ \tilde{y}_i(0) = \tilde{y}_{i,t}(0) = 0, \quad \tilde{p}_i(T) = \tilde{p}_{iT}, \quad \tilde{p}_{i,t}(T) = \tilde{q}_{iT},\end{gathered} \tag{6.13}$$

and

$$\tilde{p}_i|_{S_i} = \tilde{p}_i^\infty, \quad \tilde{y}_i|_{S_i} = \tilde{y}_i^\infty.$$

By virtue of (6.12) it is seen that $\tilde{\mathbf{Y}} := (\tilde{y}_1, \dots, \tilde{y}_m)$ is the solution of (1.1) with $f = 0$, and therefore $\tilde{\mathbf{Y}}(t) = 0$, $\tilde{\mathbf{Y}}_t(t)|_V = 0$, $0 \le t \le T$. In addition, $\tilde{\mathbf{P}} := (\tilde{p}_1, \dots, \tilde{p}_m)$ is the solution of

$$
\begin{gathered}
\tilde{p}_{i,tt} - a_i \Delta \tilde{p}_i = 0 \quad \text{in } Q_i,\ i = 1, 2, \\
\tilde{p}_i = \tilde{p}_j, \quad a_i \frac{\partial \tilde{p}_i}{\partial \nu_i} + a_j \frac{\partial \tilde{p}_j}{\partial \nu_j} = 0 \quad \text{on } \Sigma_{ij}, \\
a_i \frac{\partial \tilde{p}_i}{\partial \nu_i} = 0 \quad \text{on } \Sigma_i, \\
\tilde{\mathbf{P}}(T) = \tilde{\mathbf{P}}_{iT}, \quad \tilde{\mathbf{P}}_{,t}(T) = \tilde{\mathbf{Q}}_{iT},
\end{gathered}
\tag{6.14}
$$

where

$$
\tilde{\mathbf{P}}_{iT} = (\tilde{p}_{1T}, \dots, \tilde{p}_{mT}), \quad \tilde{\mathbf{Q}}_{iT} = (\tilde{q}_{1T}, \dots, \tilde{q}_{mT}).
$$

From (6.10) and (6.13) we have

$$
\tilde{p}_m|_\Sigma = 0,
$$

and then (4.3) gives $\tilde{\mathbf{P}}_{iT} = \tilde{\mathbf{Q}}_{iT} = 0$, hence $\tilde{\mathbf{P}}(t) = 0$, $0 \le t \le T$. □

REFERENCES

[1] A. Bamberger, R. Glowinski, and Q. H. Tran, *A domain decomposition method for the acoustic wave equation with discontinuous coefficients and grid change*, SIAM J. Numer. Anal., 34 (1997), pp. 603–639.

[2] J.-D. Benamou, *Décomposition de domaine pour le contrôle de systèmes gouvernés par des équations d'evolution*, C. R. Acad. Sci. Paris Sér. I Math., 324 (1997), pp. 1065–1070.

[3] J.-D. Benamou, *A domain decomposition method for the optimal control of systems governed the Helmholtz equation*, in Mathematical and Numerical Aspects of Wave Propagation, G. Cohen, ed., SIAM, Philadelphia, 1995, pp. 653–662.

[4] J.-D. Benamou, *A domain decomposition method with coupled transmission conditions for the optimal control of systems governed by elliptic partial differential equations*, SIAM J. Numer. Anal., 33 (1996), pp. 2401–2416.

[5] J.-D. Benamou, *A domain decomposition method for control problems*, in DD9 Proceedings 1996, P. Bjørstad et al., eds., DDM.org, Bergen, 1998, pp. 266–273.

[6] J.-D. Benamou, *Résolution d'un cas test de contrôle optimal pour un système gouvernè par l'équation des ondes à l'aide d'une méthode de décompostion de domaine*, INRIA, Rapport de Rechereche 3095, 1997.

[7] J.-D. Benamou, *Domain decomposition, optimal control of systems governed by partial differential equations and synthesis of feedback laws*, J. Optim. Theory Appl., 102 (1999), pp. 15–36.

[8] J.-D. Benamou and B. Desprès, *A domain decomposition method for the Helmholtz equation and related optimal control problems*, J. Comput. Physics, 136 (1997), pp. 68–82.

[9] A. Bounaim, *A Lagrangian approach to a DDM for an optimal control problem*, in DD9 Proceedings 1996, Bergen, P. Bjørstad et al., eds., DDM.org, 1998, pp. 283–289.

[10] E. J. Dean and R. Glowinski, *Domain decomposition of wave problems using mixed finite elements*, in DD9 Proceedings 1996, Bergen, P. Bjørstad et al., eds., DDM.org, 1998, pp. 326–333.

[11] B. Després, *Méthodes de décomposition de domaine pour les problèmes de propagation d'ondes en régimes harmoniques*, Ph.D. thesis, Université de Paris 9, 1991.

[12] B. Després, *Domain decompostion method and the Helmholtz problem*, in Mathematical and Numerical Aspects of Wave Propagation Phenomena, G. Cohen et al., eds., SIAM, Philadelphia, 1991, pp. 44–52.

[13] R. Glowinski and P. Le Tallec, *Augmented Lagrangian interpretation of the nonoverlapping Schwarz alternating method*, in The Third International Symposium on Domain Decomposition Methods for Partial Differential Equations, T. Chan and R. Glowinski, eds., SIAM, Philadelphia, 1990, pp. 224–231.

[14] R. Glowinski and J.-L. Lions, *Exact and approximate controllability for distributed parameter systems* I, in Acta Numerica 1994, A. Iserles, ed., Cambridge Univ. Press, Cambridge, 1994, pp. 269–378.

[15] R. Glowinski and J.-L. Lions, *Exact and approximate controllability for distributed parameter systems* II, in Acta Numerica 1995, A. Iserles, ed., Cambridge Univ. Press, Cambridge, 1995, pp. 159–333.

[16] R. Glowinski, J. Périaux, Z.-C. Shi, and O. Widlund, *Domain Decomposition Methods in Science and Engineering*, John Wiley & Sons, New York, 1997.

[17] V. Komornik, *Exact Controllability and Stabilization: The Multiplier Method*, Masson, Paris, 1994.

[18] J. E. Lagnese, *Boundary controllability in problems of transmission for a class of second order hyperbolic systems*, ESIAM Control Optim. Calc. Var., 2 (1997), pp. 343–357.

[19] J. E. Lagnese, G. Leugering, and E. J. P. G. Schmidt, *Modeling, Analysis and Control of Dynamic Elastic Multi-Link Structures*, Birkhäuser, Boston, 1994.

[20] I. Lasiecka and R. Triggiani, *Regularity theory of hyperbolic equations with nonhomogeneous Neumann boundary conditions, Part* II*: General boundary data*, J. Differential Equations, 94 (1991), pp. 112–164.

[21] G. Leugering, *On domain decomposition of controlled networks of elastic strings with joint masses*, in Control and Estimation of Distributed Parameter Systems, F. Kappel, ed., Internat. Ser. Numer. Math. 126, Birkhäuser, Basel, 1998, pp. 191–205.

[22] G. Leugering, *On domain decomposition of optimal control problems for dynamic networks of elastic strings*, Comput. Optim. Appl., to appear.

[23] G. Leugering, *Dynamic domain decomposition of optimal control problems for networks of strings and Timoshenko beams*, SIAM J. Control Optim., 37 (1999), pp. 1649–1675.

[24] J.-L. Lions, *Contrôlabilité Exacte, Perturbations et Stabilisation de Systémes Distribués:* Tome 1*, Contrôlabilité Exacte*, Rech. Math. Appl., 8, Masson, Paris, 1988.

[25] J.-L. Lions and O. Pironneau, *Sur le contrôle paralléle des système distribués*, C. R. Acad. Sci. Paris Sér. I Math., 327 (1998), pp. 993–998.

[26] J.-L. Lions and O. Pironneau, *Algorithmes parallèles pour la solution des problèmes aux limites*, C. R. Acad. Sci. Paris Sér. I Math., 327 (1998), pp. 947–952.

[27] P.-L. Lions, *On the Schwarz alternating method* 3: *A variant for nonoverlapping subdomain*, in The Third International Symposium on Domain Decomposition Methods for Partial Differential Equations, T. Chan and R. Glowinski, eds., SIAM, Philadelphia, 1990, pp. 202–223.

[28] S. Miyatake, *Mixed problems for hyperbolic equations of second order*, J. Math. Kyoto Univ., 13 (1973), pp. 435–487.

[29] Z. Opial, *Weak convergence of the sequence of successive approximations for nonexpansive mappings,* Bull. Amer. Math. Soc., 73 (1967), pp. 591–597.

[30] M. Papadrakakis, *Parallel Solution Methods in Computational Mechanics*, John Wiley & Sons, New York, 1997.

[31] A. Quarteroni and A. Valli, *Numerical Approximation of Partial Differential Equations*, Springer-Verlag, Berlin, 1994.

[32] J. Simon, *Compact sets in* $L^p(0,T;B)$, Ann. Mat. Pura Appl., 146 (1987), pp. 65–96.

[33] D. Tataru, *On the regularity of boundary traces for the wave equation*, Ann. Scuola Norm. Pisa Cl. Sci (4), 26 (1998), pp. 185–206.

SIAM J. CONTROL OPTIM.
Vol. 38, No. 2, pp. 538–565

STRONG CONVERGENCE OF BLOCK-ITERATIVE OUTER APPROXIMATION METHODS FOR CONVEX OPTIMIZATION*

PATRICK L. COMBETTES†

Abstract. The strong convergence of a broad class of outer approximation methods for minimizing a convex function over the intersection of an arbitrary number of convex sets in a reflexive Banach space is studied in a unified framework. The generic outer approximation algorithm under investigation proceeds by successive minimizations over the intersection of convex supersets of the feasibility set determined in terms of the current iterate and variable blocks of constraints. The convergence analysis involves flexible constraint approximation and aggregation techniques as well as relatively mild assumptions on the constituents of the problem. Various well-known schemes are recovered as special realizations of the generic algorithm and parallel block-iterative extensions of these schemes are devised within the proposed framework. The case of inconsistent constraints is also considered.

Key words. block-iterative, convex feasibility problem, convex programming, constrained minimization, cutting plane, fixed point, inconsistent constraints, outer approximation, projection onto an intersection of convex sets, reflexive Banach space, surrogate cut, uniformly convex function

AMS subject classifications. 49M27, 65J05, 65K05, 90C25

PII. S036301299732626X

1. Introduction.

Let $\mathcal{X}$ be a real reflexive Banach space, let $J\colon \mathcal{X} \to]-\infty,+\infty]$ be a proper function, and let $(S_i)_{i\in I}$ be an arbitrary family of closed convex subsets of $\mathcal{X}$. We investigate a broad class of block-iterative outer approximation methods for solving the program

$$\text{(P)} \qquad \text{find } \overline{x} \in S \triangleq \bigcap_{i\in I} S_i \quad \text{such that } J(\overline{x}) = \inf_{x\in S} J(x) \triangleq \overline{J}$$

under the following assumptions:

(A1) J is lower semicontinuous and convex.

(A2) For some closed convex set $E \supset S$, there exists a point $u \in S \cap \operatorname{dom} J$ such that the set $C \triangleq \{x \in E \mid J(x) \leq J(u)\}$ is bounded and J is uniformly convex with modulus of convexity c on C, i.e., [53], [54]

$$(1.1) \qquad \big(\forall (x,y) \in C^2\big) \quad J\left(\frac{x+y}{2}\right) \leq \frac{J(x)+J(y)}{2} - c(\|x-y\|),$$

where $c\colon \mathbb{R}_+ \to \mathbb{R}_+$ is nondecreasing and $(\forall \tau \in \mathbb{R}_+)\ c(\tau) = 0 \Leftrightarrow \tau = 0$.

(A3) For every $i \in I$, $S_i = \{x \in \mathcal{X} \mid g_i(x) \leq 0\}$, where g_i belongs to the class $\mathcal{G}$ of

*Received by the editors September 17, 1997; accepted for publication (in revised form) October 29, 1998; published electronically February 2, 2000. This work was supported by the National Science Foundation under grant MIP-9705504.

http://www.siam.org/journals/sicon/38-2/32626.html

†Laboratoire d'Analyse Numérique, Université Pierre et Marie Curie — Paris 6, 75005 Paris, France and City College and Graduate Center, City University of New York, New York, NY 10031 (plc@ann.jussieu.fr).

all functions $g\colon \mathcal{X} \to]-\infty, +\infty]$ such that

$$\text{(1.2)} \qquad \begin{cases} \text{(i) } \{x \in \mathcal{X} \mid g(x) \leq 0\} \text{ is nonempty and convex.} \\ \text{(ii) For every sequence } (y_n)_{n\geq 0} \subset \mathcal{X} \\ \qquad \begin{cases} y_n \overset{n}{\rightharpoonup} y \\ \overline{\lim}_n g(y_n) \leq 0 \end{cases} \quad \Rightarrow \quad g(y) \leq 0. \end{cases}$$

Assumptions (A1)–(A2) are rather standard and ensure in particular that (P) admits a unique solution $\overline{x}$ [1]. Assumption (A3) provides an explicit description of the constraint sets $(S_i)_{i\in I}$ as lower level sets of functions $(g_i)_{i\in I} \subset \mathcal{G}$. As will be seen in section 2, the class $\mathcal{G}$ is quite broad, and (A3) therefore covers a wide range of constraints encountered in convex optimization problems.

In the past four decades, various outer approximation methods for constrained minimization problems have been proposed, following their introduction by Cheney and Goldstein [8] and Kelley [31] in the form of cutting plane algorithms. The underlying principle is to replace (P) by a sequence of minimizations over simple closed convex supersets $(Q_n)_{n\geq 0}$ of the feasibility set S. Typically, the approximation at iteration n can be written as $Q_n = D_n \cap H_n$, where D_n and H_n are two closed convex supersets of S, the latter being termed a cut. Outer approximation methods can be divided into two main categories, namely, cutoff methods and constraints disintegration methods. In cutoff method [37], $D_{n+1} = Q_n$ and Q_{n+1} therefore results from the accumulation of all previous cuts. Several classical algorithms fit in this framework that differ in the way the cuts are defined, e.g., [8], [30], [31], [52], [55]. Naturally, a limitation of cutoff methods is that the minimization of J over the sets $(Q_n)_{n\geq 0}$ becomes increasingly demanding in terms of both computational load and storage requirements. This shortcoming prompted the development of filtered cutoff methods in which some of the old cuts can be discarded under various hypotheses, thereby keeping the complexity of the outer approximations manageable, e.g., [5], [16], [19], [49], [50]. These methods are cumulative in the sense that every cut must be retained until it is definitely dropped. By contrast, in the somewhat less well known constraints disintegration methods, D_n is a half-space depending solely on x_n and a subgradient of J at x_n. Such schemes were first proposed by Haugazeau in the 1960s for the minimization of quadratic forms in Hilbert spaces [26] and several variants have since been proposed for this particular problem [14], [27], [41], [44], [45]. The extension to convex functions was dealt with in [35] in Banach spaces and rediscovered in Euclidean spaces in [29] and [39].

The goal of the present work is to develop a general framework for outer approximation methods that captures and extends the above algorithms. Our investigation will not only provide a unified strong convergence analysis of existing outer approximation methods for solving (P) but also yield flexible generalizations of these methods in the form of parallel block-iterative algorithms.

The paper is built around the following generic outer approximation scheme. For brevity, $\mathfrak{m}(A)$ denotes the minimizer of J over a convex set A and $\mathfrak{C}(A)$ the family of all closed convex supersets of A.

ALGORITHM 1.1. *A sequence $(x_n)_{n\geq 0}$ is constructed as follows, where E is supplied by* (A2).

Step 0. *Set $D_0 = E$, $x_0 = \mathfrak{m}(D_0)$, and $n = 0$.*

Step 1. *Take a nonempty finite index set $I_n \subset I$ and generate H_n such that*

$$H_n \in \mathfrak{C}\left(\bigcap_{i\in I_n} S_i\right). \tag{1.3}$$

Step 2. *Set $Q_n = E \cap D_n \cap H_n$ and $x_{n+1} = \mathfrak{m}(Q_n)$.*
Step 3. *Generate D_{n+1} such that*

$$D_{n+1} \in \mathfrak{C}(S) \quad \text{and} \quad x_{n+1} = \mathfrak{m}(D_{n+1}). \tag{1.4}$$

Step 4. *Set $n = n + 1$ and go to Step* 1.

Associated with this algorithm is the following terminology.

DEFINITION 1.1. *Let H_n and D_{n+1} be two subsets of $\mathcal{X}$. Then H_n (respectively, D_{n+1}) will be said to be a* cut *(respectively, a* base*) for Algorithm* 1.1 *at iteration n if* (1.3) *(respectively,* (1.4)*) holds.*

At iteration n, x_n and an outer approximation D_n to S are given such that x_n minimizes J over D_n. A finite block of sets $(S_i)_{i\in I_n}$ is then selected and H_n is constructed as an outer approximation to their intersection. The update x_{n+1} is the minimizer of J over $Q_n = E \cap D_n \cap H_n$. We observe that, since $u \in Q_n$, x_{n+1} is the minimizer of J over $Q_n \cap C$. Consequently, since J is weakly lower semicontinuous (by (A1)) and since $Q_n \cap C$ is nonempty and weakly compact (bounded, closed, and convex by (A1)–(A2) in the reflexive space $\mathcal{X}$), the existence of x_{n+1} follows from Weierstrass' theorem [1, Thm. 2.1.1]; its uniqueness follows from the strict convexity of J over $Q_n \cap C$, which is secured by (A2). The iteration is completed by generating a new outer approximation D_{n+1} to S over which J achieves its infimum at x_{n+1}.

The remainder of the paper is divided into six sections. In section 2 basic notation and definitions are introduced and assumptions (A1)–(A3) are illustrated through specific examples. In section 3 we establish the strong convergence of Algorithm 1.1 to the solution $\overline{x}$ of (P) for two types of control sequence $(I_n)_{n\geq 0}$ under certain "tightness" conditions. Four frameworks are then considered individually. In section 4, two general cut construction techniques are described, namely, exact-constraint cuts in section 4.1 and surrogate cuts in section 4.2. In the former case, the cuts are drawn directly from the pool of constraint sets $(S_i)_{i\in I}$, whereas in the latter they are constructed as surrogate half-spaces based on approximate projections of the current iterate onto the selected block of sets. Section 5 is devoted to the construction of bases. In section 5.1 the bases are cumulative, as in cutoff methods, whereas in section 5.2 the bases are instantaneous, as in constraints disintegration methods. By coupling a cut construction strategy from section 4.1 or 4.2 with a base construction strategy from section 5.1 or 5.2, we obtain in section 6 four general realizations of the abstract Algorithm 1.1. In each case, strong convergence theorems are given and existing methods are exhibited as special cases. As a by-product, a block-iterative algorithm for projecting onto an intersection of convex sets in a Hilbert space is presented in detail. Finally, problems with inconsistent constraints and feasibility problems are discussed in section 7.

2. Preliminaries.

2.1. Notation, definitions, and basic facts. The definitions and results stated hereafter can be found in [1].

$\mathbb{N}$ is the set of nonnegative integers, $\mathbb{N}^*$ the set of positive integers, $\mathbb{R}_+$ the set of nonnegative reals, $\mathbb{R}_+^*$ the set of positive reals, and $\mathbb{R}^N$ the standard N-dimensional

Euclidean space. $\mathcal{X}$ is a real reflexive Banach space, and Id its identity operator. $\operatorname{bd} A$ denotes the boundary of a set $A \subset \mathcal{X}$, A° its interior, $\mathfrak{m}(A)$ the minimizer of J over A (i.e., $\mathfrak{m}(A) \in A$ and $(\forall x \in A)\ J(\mathfrak{m}(A)) \leq J(x)$) provided such a point exists and is unique, and $\mathfrak{C}(A)$ the family of all closed convex supersets of A. The norm of $\mathcal{X}$ and that of its topological dual $\mathcal{X}'$ is denoted by $\|\cdot\|$, the associated distance by d, and the canonical bilinear form on $\mathcal{X} \times \mathcal{X}'$ by $\langle \cdot, \cdot \rangle$. The expressions $x_n \overset{n}{\rightharpoonup} x$ and $x_n \overset{n}{\to} x$ denote, respectively, the weak and strong convergence to x of a sequence $(x_n)_{n\geq 0}$ and $\mathfrak{W}(x_n)_{n\geq 0}$ its set of weak cluster points. The closed ball of center x and radius γ in $\mathcal{X}$ or $\mathcal{X}'$ is denoted by $B(x, \gamma)$ and the normalized duality mapping of $\mathcal{X}$ by Δ, i.e.,

$$(\forall x \in \mathcal{X}) \quad \Delta(x) = \big\{x' \in \mathcal{X}' \mid \|x\|^2 = \langle x, x' \rangle = \|x'\|^2\big\}. \tag{2.1}$$

It follows from the reflexivity of $\mathcal{X}$ that Δ is surjective ($\Delta^{-1}(x') \neq \emptyset$ for every $x' \in \mathcal{X}'$). Δ is single valued if $\mathcal{X}'$ is strictly convex.

Let $F\colon \mathcal{X} \to]-\infty, +\infty]$ be a proper function, i.e., $\operatorname{dom} F = \{x \in \mathcal{X} \mid F(x) < +\infty\} \neq \emptyset$. F is subdifferentiable at $x \in \operatorname{dom} F$ if its subdifferential at this point,

$$\partial F(x) = \big\{t' \in \mathcal{X}' \mid\ (\forall y \in \mathcal{X})\ \langle y - x, t' \rangle + F(x) \leq F(y)\big\}, \tag{2.2}$$

is not empty. A subgradient of F at x is an element of $\partial F(x)$. The lower level set of F at height $\lambda \in \mathbb{R}$ is $\operatorname{lev}_{\leq \lambda} F = \{x \in \mathcal{X} \mid F(x) \leq \lambda\}$. F is quasi-convex if its lower level sets $(\operatorname{lev}_{\leq \lambda} F)_{\lambda \in \mathbb{R}}$ are convex and it is (respectively, weakly) lower semicontinuous if they are (respectively, weakly) closed. Now suppose that $A \subset \mathcal{X}$ is a nonempty convex set and that F is convex and continuous at a point in $A \cap \operatorname{dom} F$, and let $p \in A$. Then

$$F(p) = \inf_{y \in A} F(y) \quad \Leftrightarrow \quad \big(\exists t' \in \partial F(p)\big)(\forall y \in A) \quad \langle p - y, t' \rangle \leq 0. \tag{2.3}$$

In particular, fix $x \in \mathcal{X}$ and let $F\colon y \mapsto \|x - y\|^2/2$. Then (2.3) yields

$$\|x - p\| = d(x, A) \quad \Leftrightarrow \quad \big(\exists q' \in \Delta(x - p)\big)(\forall y \in A) \quad \langle y - p, q' \rangle \leq 0 \tag{2.4}$$

and p is called a projection of x onto A. Such a point exists if A is closed and it is unique if in addition $\mathcal{X}$ is strictly convex, as is the case when $\mathcal{X}$ is uniformly convex, i.e.,

$$(\forall \epsilon \in]0, 2])(\exists \delta \in]0, 2])\big(\forall (x, y) \in B(0, 1)^2\big) \quad \|x - y\| \geq \epsilon \Rightarrow \|x + y\| \leq 2 - \delta, \tag{2.5}$$

and a fortiori when $\mathcal{X}$ is a Hilbert space.

If $\mathcal{X}$ is a Hilbert space, the identifications $\mathcal{X}' = \mathcal{X}$ and $\Delta = \mathrm{Id}$ will be made and the scalar product of $\mathcal{X}$ will also be denoted by $\langle \cdot, \cdot \rangle$. Thus, expressions such as $\langle x, y' \rangle$, where $(x, y) \in \mathcal{X}^2$ and $y' \in \Delta(y)$, will reduce to $\langle x, y \rangle$.

2.2. On assumptions (A1)–(A3). We first describe basic scenarios covered by assumptions (A1)–(A2). It should be noted at this point that the boundedness of C in (A2) is mentioned only for the sake of clarity and that it is actually implicit. Indeed, if $F\colon \mathcal{B} \to]-\infty, +\infty]$ is lower semicontinuous and uniformly convex on a closed convex set $A \subset \mathcal{B}$, where $\mathcal{B}$ is a reflexive Banach space, then $A \cap \operatorname{lev}_{\leq F(w)} F$ is bounded for every $w \in A \cap \operatorname{dom} F$ [53, Thm. 1(1)].

PROPOSITION 2.1. *Assumptions* (A1) *and* (A2) *are satisfied in each of the following cases.*

(i) *J is lower semicontinuous and convex and, for some $E \in \mathfrak{C}(S)$, there exists $u \in S \cap \operatorname{dom} J$ such that $C = E \cap \operatorname{lev}_{\leq J(u)} J$ is compact and J is strictly convex and continuous on C.*

(ii) $\mathcal{X} = \mathbb{R}^N$, *$J$ is finite and strictly convex, and either of the following conditions is fulfilled:*
  (a) *$E = \mathcal{X}$ and, for some $\overline{\lambda} \in \mathbb{R}$, $\operatorname{lev}_{\leq \overline{\lambda}} J$ is nonempty and bounded.*
  (b) *$E \in \mathfrak{C}(S)$ is bounded.*

(iii) *$\mathcal{X}$ is a Hilbert space, $E = \mathcal{X}$, and J is a coercive quadratic form, i.e., $J\colon x \mapsto a(x,x)/2 - \langle x, b\rangle$, where $b \in \mathcal{X}$ and a is a symmetric bounded bilinear form on $\mathcal{X}^2$ that satisfies*

$$\big(\exists \gamma \in \mathbb{R}_+^*\big)(\forall x \in \mathcal{X}) \quad a(x,x) \geq \gamma \|x\|^2. \tag{2.6}$$

(iv) *$\mathcal{X}$ is uniformly convex, $E = \mathcal{X}$, and $J\colon x \mapsto \int_0^{\|x-w\|} \varphi(t)dt$, where $w \in \mathcal{X}$ and $\varphi\colon \mathbb{R}_+ \to \mathbb{R}_+$ is increasing.*

(v) *Let $(\Omega, \mathcal{F}, \mu)$ be a complete finite measure space, let $p \in]1, +\infty[$, and let X be a separable, real reflexive Banach space with norm $\|\cdot\|_{\mathsf{X}}$ and Borel σ-algebra $\mathcal{B}$. $\mathcal{X} = \mathbf{L}^p_{\mathsf{X}}$ is the Lebesgue space of (equivalence classes of μ — almost everywhere (a.e.) equal) measurable functions $x\colon (\Omega, \mathcal{F}) \to (\mathsf{X}, \mathcal{B})$ such that $\int_\Omega \|x(\omega)\|^p_{\mathsf{X}} \mu(d\omega) < +\infty$ and $J\colon x \mapsto \int_\Omega \varphi(\omega, x(\omega))\mu(d\omega)$, where the integrand $\varphi\colon \Omega \times \mathsf{X} \to]-\infty, +\infty]$ fulfills the following conditions:*
  (a) *φ is measurable relative to the product σ-algebra $\mathcal{F} \times \mathcal{B}$.*
  (b) *$(\forall x \in \mathcal{X})$ $\int_\Omega |\varphi(\omega, x(\omega))|\mu(d\omega) < +\infty$.*
  (c) *The functions $(\varphi(\omega, \cdot))_{\omega \in \Omega}$ are lower semicontinuous, proper, and uniformly convex on X with common modulus of convexity c_0, hence*

$$(\forall \omega \in \Omega)(\forall (\mathsf{x}, \mathsf{y}) \in \big(\operatorname{dom} \varphi(\omega, \cdot)\big)^2) \quad \varphi\left(\omega, \frac{\mathsf{x}+\mathsf{y}}{2}\right) \leq \frac{\varphi(\omega, \mathsf{x}) + \varphi(\omega, \mathsf{y})}{2} - c_0(\|\mathsf{x}-\mathsf{y}\|_{\mathsf{X}}). \tag{2.7}$$

Moreover, c_0 is continuous and $\underline{\lim}_{\tau \to +\infty} c_0(\tau)/\tau^p > 0$.

Proof. (i) is a consequence of [36, Thm. 4.1.8.(1)]. (ii) $\Rightarrow$ (i): J is convex and, by [46, Cor. 10.1.1], continuous. Moreover, for any $u \in S$, $E \cap \operatorname{lev}_{\leq J(u)} J$ is compact. This follows from the compactness of $\operatorname{lev}_{\leq \lambda} J$ for any $\lambda \in \mathbb{R}$ in (a) [46, Cor. 8.7.1] and from that of E in (b). (iii): $a(\cdot,\cdot)$ is a scalar product on $\mathcal{X}$ with associated norm $|||\cdot|||\colon x \mapsto \sqrt{a(x,x)}$. The parallelogram identity applied to $|||\cdot|||$ and (2.6) then shows its uniform convexity on $\mathcal{X}$ with modulus of convexity $\alpha \mapsto \gamma\alpha^2/4$. Hence, $|||\cdot|||$ satisfies (A1)–(A2) and so does J. (iv): Without loss of generality, let $w = 0$. The function $\psi\colon \alpha \mapsto \int_0^\alpha \varphi(t)dt$ is well defined, finite, increasing, convex, and continuous on $\mathbb{R}_+$ [46, Thm. 24.2]. Hence, $J = \psi \circ \|\cdot\|$ is convex and continuous, and (A1) is satisfied. Finally, (A2) is satisfied due to the uniform convexity of J on any closed ball [54, Thm. 4.1(ii)] and therefore on $\operatorname{lev}_{\leq J(u)} J$ for any $u \in S$. (v): $\mathcal{X}$ is a reflexive Banach space with norm $\|\cdot\|\colon x \mapsto (\int_\Omega \|x(\omega)\|^p_{\mathsf{X}} \mu(d\omega))^{1/p}$ [20, Thm. 8.20.5]. Moreover, J is finite, continuous, and convex on $\mathcal{X}$ [47, Thm. 22(a)], which gives (A1). As regards (A2), we claim that J is uniformly convex on $\mathcal{X}$. Indeed, take arbitrarily $(x, y) \in \mathcal{X}^2$. Then it follows from (b) that $\varphi(\cdot, x(\cdot)) < +\infty$ and $\varphi(\cdot, y(\cdot)) < +\infty$ μ-a.e. Consequently, by virtue of (2.7), for μ almost every $\omega \in \Omega$, it holds that

$$\varphi\left(\omega, \frac{x(\omega)+y(\omega)}{2}\right) \leq \frac{\varphi(\omega, x(\omega)) + \varphi(\omega, y(\omega))}{2} - c_0\left(\|x(\omega) - y(\omega)\|_{\mathsf{X}}\right), \tag{2.8}$$

where, under our assumptions, the function $\omega \mapsto c_0(\|x(\omega) - y(\omega)\|_{\mathsf{X}})$ is measurable. Upon integrating (2.8), we obtain

$$J\left(\frac{x+y}{2}\right) \leq \frac{J(x)+J(y)}{2} - \int_{\Omega} c_0\left(\|x(\omega)-y(\omega)\|_{\mathsf{X}}\right)\mu(d\omega). \tag{2.9}$$

Now fix $\varepsilon \in \mathbb{R}_+^*$ arbitrarily. Then, since $\mu(\Omega) < +\infty$ and $\underline{\lim}_{\tau\to+\infty} c_0(\tau)/\tau^p > 0$, it follows from [54, Lem. 4.4] that there exists $c \in \mathbb{R}_+^*$ depending only on ε such that $\int_\Omega c_0(\|x(\omega)-y(\omega)\|_{\mathsf{X}})\mu(d\omega) \geq c$ whenever $\|x-y\| \geq \varepsilon$. This proves the claim. ∎

Scenario (ii) is an important practical instance of (i) in which (P) takes the form of a semi-infinite convex program, as commonly found in numerical applications. In scenario (iii), since there exists $w \in \mathcal{X}$ such that $(\forall x \in \mathcal{X})\langle x, b\rangle = a(x,w)$ [7, Chap. V], we can write $J\colon x \mapsto a(x-w,x-w)/2 - a(w,w)/2$. (P) can therefore be looked upon as the problem of finding the projection of w onto the intersection of the closed convex sets $(S_i)_{i\in I}$ relative to the norm $|||\cdot|||\colon x \mapsto \sqrt{a(x,x)}$. Alternatively, since for every $y \in \mathcal{X}$ $\langle y, \nabla J(\overline{x})\rangle = a(y,\overline{x}) - \langle y, b\rangle$ [15, Chap. VII], (2.3) shows that (P) is equivalent to solving the variational inequality

$$\text{find}\,\overline{x} \in S = \bigcap_{i\in I} S_i \quad \text{such that } (\forall x \in S) \quad a(x-\overline{x},\overline{x}) \geq \langle x-\overline{x}, b\rangle, \tag{2.10}$$

which arises in numerous areas of mathematical sciences [1], [7], [15]. Next, scenario (iv) describes the problem of projecting w onto the intersection of the closed convex sets $(S_i)_{i\in I}$ in a uniformly convex Banach space. It is noted that if $J\colon x \mapsto \|x-w\|^2/2$, then $\partial J\colon x \mapsto \Delta(x-w)$ [1]. Finally, scenario (v) is of interdisciplinary interest and covers problems in areas such as stochastic programming, economics, and control theory; see, e.g., [1], [43], [47]. It should be added that $t' \in \partial J(x) \Leftrightarrow t'(\cdot) \in \partial\varphi(\cdot, x(\cdot))$ μ — a.e. [47, Thm. 22(c)] and that $\mathcal{X}$ is a Hilbert space if X is a Hilbert space and $p = 2$.

We now turn to assumption (A3). The motivation for introducing the class of functions $\mathcal{G}$ stems from its ability to capture in the convenient form of functional inequalities a wide range of convex constraints arising in theoretical and practical optimization problems. As illustrated below, constraint sets in the form of lower level sets of quasi-convex functions or of fixed point sets of quasi-nonexpansive operators, as found for instance in [4], [10], [11], [12], [29], [34], [35], and [51], are included. Let us also call attention to the fact that (1.2)(ii) implies that $\text{lev}_{\leq 0} g$ is weakly closed for every $g \in \mathcal{G}$.

PROPOSITION 2.2. *Let $g\colon \mathcal{X} \to]-\infty,+\infty]$ be a function such that, for some $w \in \mathcal{X}$, $g(w) \leq 0$. Then $g \in \mathcal{G}$ if one of the conditions below is fulfilled.*

(i) *$\text{lev}_{\leq 0} g$ is convex and g is weakly lower semicontinuous.*

(ii) *g is lower semicontinuous and quasi-convex.*

(iii) *$\text{lev}_{\leq 0} g$ is closed and convex and the constraint "$g(x) \leq 0$" is correct* [37]*:*

$$\left(\forall (y_n)_{n\geq 0} \subset \mathcal{X}\right)\ \overline{\lim}_n\, g(y_n) \leq 0 \Rightarrow \underline{\lim}_n\, d(y_n, \text{lev}_{\leq 0} g) = 0. \tag{2.11}$$

(iv) *$g\colon x \mapsto \|Tx - x\|$ is the displacement function of an operator $T\colon \mathcal{X} \to \mathcal{X}$ whose fixed point set $\text{Fix}\, T \triangleq \{x \in \mathcal{X} \mid Tx = x\}$ is convex and such that $T - \text{Id}$ is demiclosed at the origin:*

$$\left(\forall (y_n)_{n\geq 0} \subset \mathcal{X}\right) \quad \begin{cases} y_n \overset{n}{\rightharpoonup} y \\ Ty_n - y_n \overset{n}{\to} 0 \end{cases} \quad \Rightarrow \quad y \in \text{Fix}\, T. \tag{2.12}$$

These conditions are fulfilled in each of the following cases.

(a) $\mathcal{X}$ *is uniformly convex and* T *is nonexpansive:* $(\forall (x,y)\in\mathcal{X}^2)\ \|Tx-Ty\| \leq \|x-y\|$.

(b) $\operatorname{Fix} T$ *is closed and convex and, for every sequence* $(y_n)_{n\geq 0}\subset\mathcal{X}$, $Ty_n - y_n \stackrel{n}{\to} 0 \Rightarrow \varliminf_n d(y_n, \operatorname{Fix} T) = 0$.

(c) $T-\operatorname{Id}$ *is demiclosed at the origin and there exists* $\eta\in\mathbb{R}_+^*$ *such that*

$$(\forall x\in\mathcal{X})\big(\exists z'\in\Delta(x-Tx)\big)(\forall y\in\operatorname{Fix} T)\quad \langle x-y, z'\rangle \geq \eta\|Tx-x\|^2. \tag{2.13}$$

(d) $\mathcal{X}$ *is a Hilbert space,* $T-\operatorname{Id}$ *is demiclosed at the origin, and* T *is quasi-nonexpansive:*

$$\big(\forall (x,y)\in\mathcal{X}\times\operatorname{Fix} T\big)\quad \|Tx-y\|\leq\|x-y\|. \tag{2.14}$$

(e) $\mathcal{X}$ *is a Hilbert space and* T *is firmly nonexpansive:*

$$\big(\forall (x,y)\in\mathcal{X}^2\big)\quad \|Tx-Ty\|^2 \leq \|x-y\|^2 - \|(T-\operatorname{Id})x-(T-\operatorname{Id})y\|^2. \tag{2.15}$$

Proof. (i) Since g is weakly lower semicontinuous, $y_n \stackrel{n}{\rightharpoonup} y \Rightarrow g(y)\leq \varlimsup_n g(y_n)$. Hence (1.2) (ii) holds. (ii) $\Rightarrow$ (i) is immediate. (iii) Since $\operatorname{lev}_{\leq 0} g$ is convex, $d(\cdot, \operatorname{lev}_{\leq 0} g)$ is convex and Lipschitzian and therefore weakly lower semicontinuous. Accordingly,

$$y_n \stackrel{n}{\rightharpoonup} y \quad\Rightarrow\quad d(y,\operatorname{lev}_{\leq 0} g)\leq \varliminf_n d(y_n, \operatorname{lev}_{\leq 0} g). \tag{2.16}$$

Hence, if we further assume $\varlimsup_n g(y_n)\leq 0$, (2.11) gives $d(y,\operatorname{lev}_{\leq 0} g)=0$; i.e., $g(y)\leq 0$ since $\operatorname{lev}_{\leq 0} g$ is closed. (iv) is immediate. (a) is proved in [24, Lem. 3.4 and Thm. 8.4]. (b) follows from (iii). (c): Let

$$(\forall x\in\mathcal{X})\quad Q_x = \big\{y\in\mathcal{X}\mid \langle x-y, z'\rangle\geq\eta\|Tx-x\|^2\big\} \tag{2.17}$$

(z' being as in (2.13)) and $Q=\bigcap_{x\in\mathcal{X}} Q_x$. Then Q is convex as an intersection of half-spaces. Let us show $\operatorname{Fix} T = Q$. $\operatorname{Fix} T\subset Q$ results at once from (2.13). Conversely, let $x\in Q$. Then $x\in Q_x$ and therefore $0\geq\eta\|Tx-x\|^2$. Thus, $Tx=x$ and, in turn, $Q\subset\operatorname{Fix} T$. (d) $\Rightarrow$ (c): In Hilbert spaces, (2.13) becomes

$$\big(\forall (x,y)\in\mathcal{X}\times\operatorname{Fix} T\big)\quad \langle x-y, x-Tx\rangle\geq\eta\|Tx-x\|^2. \tag{2.18}$$

The identity $2\langle x-y, x-Tx\rangle = \|Tx-x\|^2+\|x-y\|^2-\|Tx-y\|^2$ shows that (2.18) is equivalent to

$$\big(\forall (x,y)\in\mathcal{X}\times\operatorname{Fix} T\big)\quad \|Tx-y\|^2\leq\|x-y\|^2-(2\eta-1)\|Tx-x\|^2, \tag{2.19}$$

which reduces to (2.14) for $\eta=1/2$. (e) $\Rightarrow$ (c): T is nonexpansive and $T-\operatorname{Id}$ is therefore demiclosed by (a). In addition, (2.15) $\Rightarrow$ (2.19) with $\eta=1$. ∎

3. Convergence analysis. This section is devoted to establishing the strong convergence of Algorithm 1.1 under suitable conditions. Our starting point is the following proposition, which collects some basic properties of the algorithm.

PROPOSITION 3.1. *Let* $(x_n)_{n\geq 0}$ *be an arbitrary orbit of Algorithm* 1.1. *Then:*

(i) $(\forall n\in\mathbb{N})\ J(x_n)\leq J(x_{n+1})\leq \overline{J}$.

(ii) $(x_n)_{n\geq 0}\subset C$.

(iii) $\mathfrak{W}(x_n)_{n\geq 0}\neq\varnothing$.

(iv) $(J(x_n))_{n\geq 0}$ *converges and* $\lim_n J(x_n) \leq \overline{J}$.
(v) $(\exists n \in \mathbb{N})\ x_n \in S \quad \Rightarrow \quad (\forall k \in \mathbb{N})\ x_{n+k} = \overline{x}$.
(vi) $\mathfrak{W}(x_n)_{n\geq 0} \subset S \quad \Rightarrow \quad x_n \stackrel{n}{\rightarrow} \overline{x}$.
(vii) $x_{n+1} - x_n \stackrel{n}{\rightarrow} 0$.
(viii) $d(x_n, H_n) \stackrel{n}{\rightarrow} 0$.

Proof. (i) results from the inclusions $(\forall n \in \mathbb{N})D_n \supset Q_n \supset S$. (ii), (iv), and (v) follow from (i). (ii) $\Rightarrow$ (iii): It follows from (A1)–(A2) and the reflexivity of $\mathcal{X}$ that C is weakly compact. (vi): Assume $\mathfrak{W}(x_n)_{n\geq 0} \subset S$ and take $x \in \mathfrak{W}(x_n)_{n\geq 0}$, say $x_{n_k} \stackrel{k}{\rightharpoonup} x$. By virtue of (A1), J is weakly lower semicontinuous and it follows from (iv) that $J(x) \leq \underline{\lim}_k J(x_{n_k}) = \lim_n J(x_n) \leq \overline{J}$. However, $x \in S$ and $\overline{x}$ is the unique solution to (P). Hence, $x = \overline{x}$, $\mathfrak{W}(x_n)_{n\geq 0} = \{\overline{x}\}$, and, since C is weakly compact, (ii) yields $x_n \stackrel{n}{\rightharpoonup} \overline{x}$. Repeating the above argument, we obtain $\overline{J} = J(\overline{x}) \leq \underline{\lim}_n J(x_n)$ and, by (iv), $J(x_n) \stackrel{n}{\rightarrow} \overline{J}$. Since $(x_n + \overline{x})/2 \stackrel{n}{\rightharpoonup} \overline{x}$, the weak lower semicontinuity of J and (1.1) yield

$$\begin{aligned} \overline{J} &\leq \underline{\lim}_n J\left(\frac{x_n + \overline{x}}{2}\right) \\ &\leq \overline{\lim}_n \frac{J(x_n) + \overline{J}}{2} - \overline{\lim}_n c(\|x_n - \overline{x}\|) \\ &= \overline{J} - \overline{\lim}_n c(\|x_n - \overline{x}\|). \end{aligned} \tag{3.1}$$

Hence, $c(\|x_n - \overline{x}\|) \stackrel{n}{\rightarrow} 0$ and, by (A2), $x_n \stackrel{n}{\rightarrow} \overline{x}$. (vii): For every $n \in \mathbb{N}$, $(x_n, x_{n+1}) \in D_n^2$ and therefore $y_n = (x_n + x_{n+1})/2 \in D_n$. Since $x_n = \mathfrak{m}(D_n)$, (1.1) then yields

$$J(x_n) \leq J(y_n) \leq \frac{J(x_n) + J(x_{n+1})}{2} - c(\|x_{n+1} - x_n\|). \tag{3.2}$$

Hence, (iv) implies $c(\|x_{n+1} - x_n\|) \stackrel{n}{\rightarrow} 0$ and, in turn, $x_{n+1} - x_n \stackrel{n}{\rightarrow} 0$. (vii) $\Rightarrow$ (viii): $(\forall n \in \mathbb{N})x_{n+1} \in H_n \Rightarrow \|x_{n+1} - x_n\| \geq d(x_n, H_n)$. ∎

Item (i) above shows that Algorithm 1.1 is an ascent method. On the other hand, item (vi) guarantees the strong convergence of any orbit to the solution of (P) as long as each of its weak cluster points satisfies all the constraints. In view of (1.3), for this condition to hold, the control sequence $(I_n)_{n\geq 0}$ determining the blocks of constraints activated over the course of the iterations must sweep through the index set I in a coherent fashion; three suitable control modes will be considered in Definition 3.1. In addition, the constraint sets $(S_i)_{i\in I}$ must be tightly approximated by the cuts $(H_n)_{n\geq 0}$ in a sense that will be made precise in Definition 3.2.

DEFINITION 3.1. *Algorithm* 1.1 *operates under*

- admissible *control if* I *is countable and there exist positive integers* $(M_i)_{i\in I}$ *such that*

$$\big(\forall (i,n) \in I \times \mathbb{N}\big) \quad i \in \bigcup_{k=n}^{n+M_i-1} I_k; \tag{3.3}$$

- chaotic *control if* I *is countable and*

$$I = \overline{\lim}_n I_n \triangleq \bigcap_{n\geq 0} \bigcup_{k\geq n} I_k; \tag{3.4}$$

- coercive *control if*

$$\left(\exists (\mathrm{i}(n))_{n\geq 0} \in \bigtimes_{n\geq 0} I_n\right) \quad \overline{\lim}_n g_{\mathrm{i}(n)}(x_n) \leq 0 \Rightarrow \overline{\lim}_n \sup_{i\in I} g_i(x_n) \leq 0. \tag{3.5}$$

In addition, Algorithm 1.1 *is* serial *if* $(I_n)_{n\geq 0}$ *reduces to a sequence of singletons* $(\{i(n)\})_{n\geq 0}$. *The above admissibility and coercivity conditions then read*

$$(\forall i \in I)(\exists M_i \in \mathbb{N}^*)(\forall n \in \mathbb{N}) \quad i \in \{\mathrm{i}(n), \ldots, \mathrm{i}(n+M_i-1)\} \tag{3.6}$$

and

$$\overline{\lim}_n g_{\mathrm{i}(n)}(x_n) \leq 0 \Rightarrow \overline{\lim}_n \sup_{i\in I} g_i(x_n) \leq 0, \tag{3.7}$$

respectively.

The coercive control mode is found in [13] with $(\forall i \in I)$ $g_i\colon x \mapsto d(x, S_i)$. The admissible and chaotic control modes have already been used at various levels of generality in convex feasibility problems [4], [10], [13], [34], [42].

DEFINITION 3.2. *Let* $(x_n)_{n\geq 0}$ *be an arbitrary orbit of Algorithm* 1.1. *Then the algorithm will be said to be*

- tight *if, for every* $i \in I$ *and every increasing sequence* $(n_k)_{k\geq 0} \subset \mathbb{N}$ *such that* $i \in \bigcap_{k\geq 0} I_{n_k}$, *we have* $\overline{\lim}_k g_i(x_{n_k}) \leq 0$;
- strongly tight *if* $\overline{\lim}_n \max_{i\in I_n} g_i(x_n) \leq 0$.

It is clear that strong tightness implies tightness. We show below that, when I is finite, the distinction between the two notions disappears.

PROPOSITION 3.2. *Suppose that* I *is finite. Then Algorithm* 1.1 *is tight if and only if it is strongly tight.*

Proof. To show necessity, take an arbitrary orbit $(x_n)_{n\geq 0}$ and suppose that the algorithm is not strongly tight, i.e., that $\epsilon \triangleq \overline{\lim}_n \max_{i\in I_n} g_i(x_n) > 0$. Define a sequence $(\mathrm{i}(n))_{n\geq 0} \subset I$ by $(\forall n \in \mathbb{N})$ $g_{\mathrm{i}(n)}(x_n) = \max_{i\in I_n} g_i(x_n)$. Then, since I is finite, there exists an index $i \in I$ and an increasing sequence $(n_k)_{k\geq 0} \subset \mathbb{N}$ such that $(\forall k \in \mathbb{N})$ $\mathrm{i}(n_k) = i$ and $g_i(x_{n_k}) \stackrel{k}{\to} \epsilon$, in contradiction of the tightness assumption. □

We are now ready to state and prove the following strong convergence result.

THEOREM 3.1. *Let* $(x_n)_{n\geq 0}$ *be an arbitrary orbit of Algorithm* 1.1 *generated under either of the following conditions:*

(i) *tightness, I countable, and admissible control;*

(ii) *strong tightness and coercive control.*

Then $x_n \stackrel{n}{\to} \overline{x}$.

Proof. By virtue of Proposition 3.1(vi), it suffices to show $\mathfrak{W}(x_n)_{n\geq 0} \subset S$. Fix arbitrarily $i \in I$ and $x \in \mathfrak{W}(x_n)_{n\geq 0}$, say $x_{n_k} \stackrel{k}{\to} x$. Then it is enough to show $x \in S_i$, i.e., that $g_i(x) \leq 0$. (i): By (3.3), there exist $M_i \in \mathbb{N}^*$ and an increasing sequence $(p_k)_{k\geq 0} \subset \mathbb{N}$ such that

$$(\forall k \in \mathbb{N}) \quad n_k \leq p_k \leq n_k + M_i - 1 \qquad \text{and} \qquad i \in I_{p_k}. \tag{3.8}$$

Hence

$$(\forall k \in \mathbb{N}) \quad \|x_{p_k} - x_{n_k}\| \leq \sum_{l=n_k}^{n_k+M_i-2} \|x_{l+1} - x_l\| \tag{3.9}$$

and Proposition 3.1(vii) yields $x_{p_k} - x_{n_k} \xrightarrow{k} 0$. Consequently, $x_{p_k} \xrightarrow{k} x$. On the other hand, the tightness condition gives $\overline{\lim}_k g_i(x_{p_k}) \leq 0$ and (A3) then yields $g_i(x) \leq 0$, as desired. (ii): The strong tightness condition gives $\overline{\lim}_n \max_{j\in I_n} g_j(x_n) \leq 0$. However, since the control is coercive, we obtain

$$\begin{aligned} \overline{\lim}_n \max_{j\in I_n} g_j(x_n) \leq 0 &\Rightarrow \overline{\lim}_n g_{\mathrm{i}(n)}(x_n) \leq 0 \\ &\Rightarrow \overline{\lim}_n \sup_{j\in I} g_j(x_n) \leq 0 \\ &\Rightarrow \overline{\lim}_k g_i(x_{n_k}) \leq 0, \end{aligned} \tag{3.10}$$

where the sequence $(\mathrm{i}(n))_{n\geq 0}$ is as in (3.5). It then follows from (A3) that $g_i(x) \leq 0$, which completes the proof. □

We conclude this section by supplying a theoretical condition under which Theorem 3.1(i) can be extended to the chaotic control mode (3.4).

PROPOSITION 3.3. *Suppose that Algorithm* 1.1 *is tight and that I is countable, and let $(x_n)_{n\geq 0}$ be any of its orbits generated under chaotic control. Then, if $(x_n)_{n\geq 0}$ admits at most one weak cluster point, $x_n \xrightarrow{n} \overline{x}$.*

Proof. It follows from Proposition 3.1(ii) and the weak compactness of C that if $(x_n)_{n\geq 0}$ admits at most one weak cluster point, then it converges weakly, say $x_n \xrightarrow{n} x$. Now, fix $i \in I$ arbitrarily. According to Proposition 3.1(vi), it remains to show $g_i(x) \leq 0$. By condition (3.4), there exists an increasing sequence $(n_k)_{k\geq 0} \subset \mathbb{N}$ such that $i \in \bigcap_{k\geq 0} I_{n_k}$. In turn, tightness implies $\overline{\lim}_k g_i(x_{n_k}) \leq 0$ and, since $x_{n_k} \xrightarrow{k} x$, (A3) yields $g_i(x) \leq 0$. □

The execution of iteration n of Algorithm 1.1 necessitates the construction of a cut H_n at Step 1 and of a base D_{n+1} at Step 3 (see Definition 1.1). This question is addressed in the next two sections.

4. Cut construction schemes. In this section, we describe two techniques to construct cuts for Algorithm 1.1 and provide examples of families of constraint functions $(g_i)_{i\in I}$ that yield tight and strongly tight algorithms in each case.

4.1. Exact-constraint cuts. Here, Algorithm 1.1 is assumed to operate under serial control, say $(\forall n \in \mathbb{N})\ I_n = \{\mathrm{i}(n)\}$. In view of Definition 1.1, the following observation is self-evident.

PROPOSITION 4.1. *The set $H_n = S_{i(n)}$ is a cut for Algorithm* 1.1 *at iteration n.*

When it operates under serial control with cuts generated as above, Algorithm 1.1 will be said to be implemented with *exact-constraint cuts*. We now proceed with some examples of families $(g_i)_{i\in I}$ that yield tight and strongly tight algorithms (see also Proposition 3.2). In Propositions 4.2 and 4.3, γ is the diameter of C in (A2) and $Q = B(u, 2\gamma)$.

PROPOSITION 4.2. *Algorithm* 1.1 *with exact-constraint cuts is tight if, for every $i \in I$, one of the following conditions holds.*

(i) *g_i is uniformly continuous on Q.*
(ii) *g_i is weakly continuous on Q.*
(iii) *$\mathcal{X} = \mathbb{R}^N$ and g_i is finite and convex.*
(iv) *g_i is the displacement function of an operator $T_i\colon \mathcal{X} \to \mathcal{X}$ which satisfies condition* (c) *(in particular* (d) *or* (e)*) in Proposition* 2.2(iv) *with constant $\eta_i \in \mathbb{R}_+^*$.*

Proof. Given an arbitrary orbit $(x_n)_{n\geq 0}$, Propositions 3.1(viii) and 4.1 give $d(x_n, S_{\mathrm{i}(n)}) \xrightarrow{n} 0$. Now take an index $i \in I$ and an increasing sequence $(n_k)_{k\geq 0} \subset \mathbb{N}$

such that, for every $k \in \mathbb{N}$, $i = \underline{\mathrm{i}}(n_k)$. Then $d(x_{n_k}, S_i) \stackrel{k}{\to} 0$ and, in view of Definition 3.2, it must be proved that $\overline{\lim}_k g_i(x_{n_k}) \leq 0$. (i) is similar to Proposition 4.3(i) and thus is omitted. (ii) $\Rightarrow$ (i) follows from the weak compactness of Q. (iii) $\Rightarrow$ (i): g_i is Lipschitzian on Q by [46, Thm. 10.4]. (iv): For every $k \in \mathbb{N}$, let $p_{i,k}$ be a projection of x_{n_k} onto S_i and suppose that (c) in Proposition 2.2(iv) holds with constant $\eta_i \in \mathbb{R}_+^*$. Then, there exists $z'_k \in \Delta(x_{n_k} - T_i x_{n_k})$ such that $\|T_i x_{n_k} - x_{n_k}\|^2 \leq \eta_i^{-1} \langle x_{n_k} - p_{i,k}, z'_k \rangle$. Consequently, $\|T_i x_{n_k} - x_{n_k}\|^2 \leq \eta_i^{-1} \|x_{n_k} - p_{i,k}\| \cdot \|z'_k\| = \eta_i^{-1} d(x_{n_k}, S_i) \cdot \|T_i x_{n_k} - x_{n_k}\|$ and we conclude $g_i(x_{n_k}) \leq \eta_i^{-1} d(x_{n_k}, S_i)$. □

PROPOSITION 4.3. *Algorithm* 1.1 *with exact-constraint cuts is strongly tight if one of the following conditions holds.*

(i) $(g_i)_{i \in I}$ *is uniformly equicontinuous on* Q.
(ii) $(g_i)_{i \in I}$ *is weakly equicontinuous on* Q*: For every* $(x, \epsilon) \in Q \times \mathbb{R}_+^*$*, there exists a weak neighborhood* V *of* x *such that* $(\forall y \in V)(\forall i \in I)\ |g_i(x) - g_i(y)| \leq \epsilon$.
(iii) $(g_i)_{i \in I}$ *is a family of affine functions associated with a family of pointwise bounded continuous linear functions.*
(iv) $\mathcal{X} = \mathbb{R}^N$ *and* $(g_i)_{i \in I}$ *is a family of pointwise bounded convex functions.*
(v) $(g_i)_{i \in I}$ *is a family of displacement functions of operators* $(T_i)_{i \in I}$ *as in Proposition* 4.2(iv) *with* $\eta \triangleq \inf_{i \in I} \eta_i > 0$ *(in particular, each* T_i *satisfies condition* (d) *or* (e) *in Proposition* 2.2(iv)*).*

Proof. Take an arbitrary orbit $(x_n)_{n \geq 0}$. Then, as above, $d(x_n, S_{\mathrm{i}(n)}) \stackrel{n}{\to} 0$ and, in view of Definition 3.2, it must be proved that $\overline{\lim}_n g_{\mathrm{i}(n)}(x_n) \leq 0$. (i): Fix $\epsilon \in \mathbb{R}_+^*$, extract a subsequence $(x_{n_k})_{k \geq 0}$ such that $(g_{\mathrm{i}(n_k)}(x_{n_k}))_{k \geq 0} \subset]0, +\infty]$ (if no such subsequence exists, the proof is complete), and let p_k be a projection of x_{n_k} onto $S_{\mathrm{i}(n_k)}$. Since $u \in S_{\mathrm{i}(n_k)}$ and $(x_{n_k}, u) \in C^2$, we have

$$\|p_k - u\| \leq \|x_{n_k} - u\| + \|x_{n_k} - p_k\| \leq 2\|x_{n_k} - u\| \leq 2\gamma \tag{4.1}$$

and, in turn, $p_k \in B(u, 2\gamma) = Q$. Next, as $x_{n_k} - p_k \stackrel{k}{\to} 0$ and $((x_{n_k}, p_k))_{k \geq 0} \subset Q^2$, the uniform equicontinuity of $(g_i)_{i \in I}$ on Q gives, for k sufficiently large, $\sup_{i \in I} |g_i(x_{n_k}) - g_i(p_k)| \leq \epsilon$ and, therefore, $0 < g_{\mathrm{i}(n_k)}(x_{n_k}) \leq \epsilon$. Since ϵ can be arbitrarily small, strong tightness ensues. (ii) $\Rightarrow$ (i) follows from the weak compactness of Q. (iii) $\Rightarrow$ (i): $(\forall i \in I) g_i \colon x \mapsto \langle x, z'_i \rangle + \alpha_i$, where $(z'_i, \alpha_i) \in \mathcal{X}' \times \mathbb{R}$ and $(\forall x \in \mathcal{X}) \sup_{i \in I} |\langle x, z'_i \rangle| < +\infty$. The uniform boundedness principle [1, Thm. 1.1.4] asserts that $\zeta \triangleq \sup_{i \in I} \|z'_i\| < +\infty$ and, therefore, that $(g_i)_{i \in I}$ is equi-Lipschitzian with constant ζ. (iv) $\Rightarrow$ (i): $(g_i)_{i \in I}$ is equi-Lipschitzian on Q by [46, Th. 10.6]. (v): Following the proof of Proposition 4.2(iv), we obtain $(\forall n \in \mathbb{N})\ g_{\mathrm{i}(n)}(x_n) \leq \eta^{-1} d(x_n, S_{\mathrm{i}(n)})$. □

It should be remarked that in Hilbert spaces, projectors are nonexpansive [24, Chap. 12]. Accordingly, the inequalities $\|p_k - u\| \leq \|x_{n_k} - u\| \leq \gamma$ can be used in lieu of (4.1), and one can take $Q = B(u, \gamma)$ in Propositions 4.2 and 4.3.

Next, we recover the framework proposed by Laurent and Martinet in [35].

EXAMPLE 4.1. *Under the strong tightness condition, Algorithm* 1.1 *implemented with coercive control and exact-constraint cuts contains the setting of* [35]. *There,* (P) *is investigated under assumptions* (A1)–(A2) *with* $E = \mathcal{X}$ *and the special instance of* (A3) *when the functions* $(g_i)_{i \in I}$ *are lower semicontinuous, convex, and satisfy the condition*

$$(\exists \Omega \in \mathbb{R}_+^*)\big(\forall (x, i) \in C \times I\big) \quad g_i(x) \leq \Omega d(x, S_i). \tag{4.2}$$

Furthermore, the serial control rule

$$(\forall n \in \mathbb{N})\ g_{\mathrm{i}(n)}(x_n) \geq \theta \sup_{i \in I} g_i(x_n) - \rho_n, \quad \text{where } 0 < \theta \leq 1 \text{ and } 0 \leq \rho_n \overset{n}{\to} 0 \tag{4.3}$$

is in force. Since $d(x_n, S_{\mathrm{i}(n)}) \overset{n}{\to} 0$, (4.2) $\Rightarrow \overline{\lim}_n g_{\mathrm{i}(n)}(x_n) \leq 0$, *which shows strong tightness. On the other hand, since* (4.3) $\Rightarrow$ (3.7), *the control is coercive. Hence,* [35, Thm. 1] *is a corollary of Theorem* 3.1(ii).

4.2. Surrogate cuts. In this section, the cut H_n at iteration n is constructed as a surrogate half-space (this terminology appears in [23]). The basic idea is, for every $i \in I_n$, to "linearize" g_i by approximating it by a continuous affine function $g_{i,n}$ (determined here geometrically via a projection onto a simple superset of S_i). A surrogate function $\widetilde{g}_n$ is then formed as a convex combination of the family $(g_{i,n})_{i \in I_n}$, and the cut is defined as $H_n = \mathrm{lev}_{\leq \gamma_n} \widetilde{g}_n$ for some $\gamma_n \in \mathbb{R}_+$. We formally define *surrogate cuts* as follows.

PROPOSITION 4.4. *Fix* $(\delta, \varepsilon) \in]0,1[^2$ *and let*

$$H_n = \left\{ x \in \mathcal{X} \mid \sum_{i \in I_n} w_{i,n} \langle x - p_{i,n}, q'_{i,n} \rangle \leq \gamma_n \right\}, \tag{4.4}$$

where the following conditions hold.

(C1) *For every* $i \in I_n$, $p_{i,n}$ *is a projection of* x_n *onto a set* $S_{i,n} \in \mathfrak{C}(S_i)$ *and* $q'_{i,n} \in \Delta(x_n - p_{i,n})$ *satisfies*

$$(\forall x \in S_{i,n}) \quad \langle x - p_{i,n}, q'_{i,n} \rangle \leq 0. \tag{4.5}$$

(C2) $(w_{i,n})_{i \in I_n} \subset [0,1]$, $\sum_{i \in I_n} w_{i,n} = 1$, *and*

$$(\exists j \in I_n) \quad \begin{cases} d(x_n, S_{j,n}) = \max\limits_{i \in I_n} d(x_n, S_{i,n}), \\ w_{j,n} \geq \delta. \end{cases}$$

(C3) $0 \leq \gamma_n \leq (1 - \varepsilon) \sum_{i \in I_n} w_{i,n} d(x_n, S_{i,n})^2$.

Then H_n *is a cut for Algorithm* 1.1 *at iteration* n.

Proof. Let us show that (1.3) holds. First, it is clear that H_n is closed and convex. Second, (C1) yields $S_i \subset S_{i,n} \subset \{x \in \mathcal{X} \mid \langle x - p_{i,n}, q'_{i,n} \rangle \leq 0\}$ for every $i \in I_n$. Hence, by virtue of (C2) and (C3), $H_n \in \mathfrak{C}(\bigcap_{i \in I_n} S_i)$. □

The existence of $(q'_{i,n})_{i \in I_n}$ in (C1) is guaranteed by (2.4), while (2.1) yields

$$(\forall i \in I_n) \quad \|q'_{i,n}\|^2 = d(x_n, S_{i,n})^2 = \langle x_n - p_{i,n}, q'_{i,n} \rangle. \tag{4.6}$$

On the other hand, $(p_{i,n})_{i \in I_n}$ and $(q'_{i,n})_{i \in I_n}$ are uniquely defined if $\mathcal{X}$ and $\mathcal{X}'$ are strictly convex, respectively [1]. In particular, if $\mathcal{X}$ is a Hilbert space, one can identify $q'_{i,n} = x_n - p_{i,n}$ hereafter, and (4.4) becomes

$$H_n = \left\{ x \in \mathcal{X} \mid \sum_{i \in I_n} w_{i,n} \langle x - p_{i,n}, x_n - p_{i,n} \rangle \leq \gamma_n \right\}. \tag{4.7}$$

Surrogate half-spaces have already been used—explicitly or implicitly—for solving convex feasibility problems in Hilbert spaces. Thus, in the methods of [10], [12], [13], [22], [32], [34], [40], [42], [45], the update x_{n+1} is obtained by (under/over) projecting the current iterate x_n onto a half-space whose general form is (4.7). This point will be reexamined in section 7.2.

An important feature of Algorithm 1.1 with surrogate cuts is that it does not require the ability to enforce exactly a constraint "$g_i(x) \leq 0$" selected at Step 1 but merely the ability to move the current iterate x_n toward S_i by means of a projection onto a superset $S_{i,n}$. A wide range of approximating supersets are acceptable, and the construction of $S_{i,n}$ can be adapted to the nature of the function g_i. In the two examples below, $S_{i,n}$ is constructed as an affine half-space and the expressions for $p_{i,n}$, $q'_{i,n}$, and $d(x_n, S_{i,n})$ are derived from the following facts.

LEMMA 4.1 (see [48, Lem. I.1.2]). *Given a nonzero functional $z' \in \mathcal{X}'$ and $\alpha \in \mathbb{R}$, consider the closed affine half-space $A = \{y \in \mathcal{X} \mid \langle y, z' \rangle \leq \alpha\}$. Take $x \notin A$ and let $p = x + (\alpha - \langle x, z' \rangle)z/\|z'\|^2$, where $z \in \Delta^{-1}(z')$. Then $d(x, A) = (\langle x, z' \rangle - \alpha)/\|z'\|$ and p is a projection of x onto A.*

EXAMPLE 4.2. *The function g_i is convex and lower semicontinuous, and subdifferentiable on C. Then, given $t'_{i,n} \in \partial g_i(x_n)$, the function $x \mapsto \langle x - x_n, t'_{i,n} \rangle + g_i(x_n)$ minorizes g_i by* (2.2). *Thus*

$$S_{i,n} = \{x \in \mathcal{X} \mid \langle x_n - x, t'_{i,n} \rangle \geq g_i(x_n)\} \tag{4.8}$$

lies in $\mathfrak{C}(S_i)$. If $x_n \notin S_i$, then $p_{i,n} = x_n - g_i(x_n)t_{i,n}/\|t'_{i,n}\|^2$, $q'_{i,n} = g_i(x_n)t'_{i,n}/\|t'_{i,n}\|^2$, and $d(x_n, S_{i,n}) = g_i(x_n)/\|t'_{i,n}\|$, where $t_{i,n} \in \Delta^{-1}(t'_{i,n})$.

Approximations of type (4.8) go back to [31] and have been used extensively; see, e.g., [4], [12], [29], [32], [34].

EXAMPLE 4.3. *The function g_i is the displacement function of an operator T_i as in Proposition* 2.2(iv)(c) *with constant $\eta_i \in \mathbb{R}_+^*$. Hence $S_i = \operatorname{Fix} T_i$ and, for some $z'_{i,n} \in \Delta(x_n - T_i x_n)$, we have $\langle x_n - x, z'_{i,n} \rangle \geq \eta_i \|T_i x_n - x_n\|^2$ for every $x \in S_i$. Therefore*

$$S_{i,n} = \big\{x \in \mathcal{X} \mid \langle x - x_n - \eta_i(T_i x_n - x_n), z'_{i,n} \rangle \leq 0\big\} \tag{4.9}$$

lies in $\mathfrak{C}(S_i)$. Furthermore, $p_{i,n} = x_n + \eta_i(T_i x_n - x_n)$, $q'_{i,n} = \eta_i z'_{i,n}$, and $d(x_n, S_{i,n}) = \eta_i \|T_i x_n - x_n\|$.

Further examples can be derived from Example 4.3 by considering the special cases (d) or (e) of (c) in Proposition 2.2(iv). For instance, if $\mathcal{X}$ is a Hilbert space and T_i is firmly nonexpansive, (4.9) reads as

$$S_{i,n} = \{x \in \mathcal{X} \mid \langle x - T_i x_n, x_n - T_i x_n \rangle \leq 0\}. \tag{4.10}$$

This particular approximation appears implicitly in [4] and [10], and explicitly in [34].

We preface our study of the tightness of Algorithm 1.1 with surrogate cuts with two basic facts.

PROPOSITION 4.5. $(\forall n \in \mathbb{N})\ \sum_{i\in I_n} w_{i,n}q'_{i,n} = 0 \Leftrightarrow x_n \in \bigcap_{i\in I_n} S_{i,n} \Leftrightarrow x_n \in H_n$.

Proof. Fix $(n,x) \in \mathbb{N}\times S$. Then (C2), (4.6), and (4.5) imply

$$\begin{aligned}\delta \max_{i\in I_n} d(x_n,S_{i,n})^2 &\le \sum_{i\in I_n} w_{i,n} d(x_n,S_{i,n})^2 = \sum_{i\in I_n} w_{i,n}\langle x_n - p_{i,n}, q'_{i,n}\rangle \\ &= \sum_{i\in I_n} w_{i,n}\langle x_n - x, q'_{i,n}\rangle + \sum_{i\in I_n} w_{i,n}\langle x - p_{i,n}, q'_{i,n}\rangle \\ &\le \left\langle x_n - x, \sum_{i\in I_n} w_{i,n}q'_{i,n}\right\rangle. \qquad (4.11)\end{aligned}$$

Hence, $\sum_{i\in I_n} w_{i,n}q'_{i,n} = 0 \Rightarrow \max_{i\in I_n} d(x_n,S_{i,n})^2 = 0 \Rightarrow x_n \in \bigcap_{i\in I_n} S_{i,n}$. The three other implications are easily obtained. □

PROPOSITION 4.6. $\max_{i\in I_n} d(x_n,S_{i,n}) \overset{n}{\to} 0$.

Proof. Fix $n \in \mathbb{N}$ and suppose $x_n \notin \bigcap_{i\in I_n} S_{i,n}$. Then Proposition 4.5, the convexity of $\|\cdot\|$, (C2), and (4.6) yield $0 \neq \|\sum_{i\in I_n} w_{i,n}q'_{i,n}\| \le \sum_{i\in I_n} w_{i,n}\|q'_{i,n}\| \le \max_{i\in I_n} d(x_n,S_{i,n})$. Consequently, we derive from (4.4), Lemma 4.1, (4.6), and (4.11) that

$$\begin{aligned} d(x_n,H_n) &= \frac{\sum_{i\in I_n} w_{i,n}\langle x_n - p_{i,n}, q'_{i,n}\rangle - \gamma_n}{\|\sum_{i\in I_n} w_{i,n}q'_{i,n}\|} \\ &\ge \varepsilon \frac{\sum_{i\in I_n} w_{i,n} d(x_n,S_{i,n})^2}{\max_{i\in I_n} d(x_n,S_{i,n})} \\ &\ge \delta\varepsilon \max_{i\in I_n} d(x_n,S_{i,n}). \qquad (4.12)\end{aligned}$$

On the other hand, if $x_n \in \bigcap_{i\in I_n} S_{i,n}$, then (4.12) is immediate. Since $d(x_n,H_n) \overset{n}{\to} 0$ by Proposition 3.1(viii), the assertion is proved. □

We observe in passing that when Algorithm 1.1 is implemented with surrogate cuts and satisfies the tightness condition then, for every index $i \in I$ and every suborbit $(x_{n_k})_{k\ge 0}$ such that $i \in \bigcap_{k\ge 0} I_{n_k}$, it follows from Proposition 4.6 that $d(x_{n_k},S_{i,n_k}) \overset{k}{\to} 0$ and from (A3) that $x_{n_k} \overset{k}{\to} x \Rightarrow x \in S_i$. This is essentially the *focusing* property introduced in [4].

We wind up this section by furnishing convenient criteria for tightness and strong tightness.

PROPOSITION 4.7. *Algorithm* 1.1 *with surrogate cuts is tight if, for every* $i \in I$ *and every suborbit* $(x_{n_k})_{k\ge 0}$ *such that* $i \in \bigcap_{k\ge 0} I_{n_k}$, *one of the following conditions is fulfilled.*

(i) *It holds that*

$$d\big(x_{n_k},S_{i,n_k}\big) \overset{k}{\to} 0 \quad \Rightarrow \quad d\big(x_{n_k},S_i\big) \overset{k}{\to} 0 \qquad (4.13)$$

and any of conditions (i)–(iii) *in Proposition* 4.2 *is satisfied.*

(ii) g_i *is as in Example* 4.2 *with the additional assumption that its subdifferential is bounded on* C, *i.e., that*

$$\big(\exists \zeta_i \in \mathbb{R}_+^*\big)(\forall x \in C) \quad \partial g_i(x) \subset B(0,\zeta_i), \qquad (4.14)$$

and the sets $(S_{i,n_k})_{k\ge 0}$ *are as in* (4.8).

(iii) *g_i is as in Example* 4.3 *and the sets $(S_{i,n_k})_{k\geq 0}$ are as in* (4.9).

Proof. Take an arbitrary orbit $(x_n)_{n\geq 0}$. Proposition 4.6 asserts that $\max_{i\in I_n} d(x_n, S_{i,n}) \stackrel{n}{\to} 0$. Hence, given $i \in I$ and an increasing sequence $(n_k)_{k\geq 0} \subset \mathbb{N}$ such that $i \in \bigcap_{k\geq 0} I_{n_k}$, we have $d(x_{n_k}, S_{i,n_k}) \stackrel{k}{\to} 0$ and must show $\overline{\lim}_k g_i(x_{n_k}) \leq 0$. (i): (4.13) yields $d(x_{n_k}, S_i) \stackrel{k}{\to} 0$. However, under condition (i) (and in particular condition (ii) or (iii)) of Proposition 4.2, $d(x_{n_k}, S_i) \stackrel{k}{\to} 0 \Rightarrow \overline{\lim}_k g_i(x_{n_k}) \leq 0$. (ii): $(\forall k \in \mathbb{N}) \max\{0, g_i(x_{n_k})\} = \|t'_{i,n_k}\| \cdot d(x_{n_k}, S_{i,n_k}) \leq \zeta_i d(x_{n_k}, S_{i,n_k})$ by (4.14). (iii): $(\forall k \in \mathbb{N})\ g_i(x_{n_k}) = \|T_i x_{n_k} - x_{n_k}\| = \eta_i^{-1} d(x_{n_k}, S_{i,n_k})$. □

PROPOSITION 4.8. *Algorithm* 1.1 *with surrogate cuts is strongly tight if one of the following conditions holds.*

(i) *For any of its orbits $(x_n)_{n\geq 0}$, we have*

$$\max_{i\in I_n} d(x_n, S_{i,n}) \stackrel{n}{\to} 0 \quad \Rightarrow \quad \max_{i\in I_n} d(x_n, S_i) \stackrel{n}{\to} 0, \tag{4.15}$$

and any of conditions (i)–(iv) *in Proposition* 4.3 *is satisfied.*

(ii) *$(g_i)_{i\in I}$ is as in Example* 4.2 *with the additional assumption that the subdifferentials are equibounded on C, i.e., that*

$$\big(\exists \zeta \in \mathbb{R}_+^*\big)(\forall i \in I)(\forall x \in C) \quad \partial g_i(x) \subset B(0, \zeta), \tag{4.16}$$

and the sets $((S_{i,n})_{i\in I_n})_{n\geq 0}$ are as in (4.8).

(iii) *$(g_i)_{i\in I}$ is as in Example* 4.3*, with the additional assumption that $\eta \triangleq \inf_{i\in I} \eta_i > 0$, and the sets $((S_{i,n})_{i\in I_n})_{n\geq 0}$ are as in* (4.9).

Proof. Take an arbitrary orbit $(x_n)_{n\geq 0}$. Then Proposition 4.6 entails $\max_{i\in I_n} d(x_n, S_{i,n}) \stackrel{n}{\to} 0$. Let us show $\overline{\lim}_n \max_{i\in I_n} g_i(x_n) \leq 0$. (i): Define a sequence $(\mathrm{i}(n))_{n\geq 0} \subset I$ by $(\forall n \in \mathbb{N})\ g_{\mathrm{i}(n)}(x_n) = \max_{i\in I_n} g_i(x_n)$. Then (4.15) $\Rightarrow d(x_n, S_{\mathrm{i}(n)}) \stackrel{n}{\to} 0$. However, under condition (i) (and in particular any of conditions (ii)–(iv)) of Proposition 4.3, $d(x_n, S_{\mathrm{i}(n)}) \stackrel{n}{\to} 0 \Rightarrow \overline{\lim}_n g_{\mathrm{i}(n)}(x_n) \leq 0$, as desired. (ii) and (iii): Fix $n \in \mathbb{N}$. Following the proof of Proposition 4.7(ii) and (iii), we obtain, respectively, $\max_{i\in I_n} g_i(x_n) \leq \zeta \max_{i\in I_n} d(x_n, S_{i,n})$ and $\max_{i\in I_n} g_i(x_n) = \max_{i\in I_n} \|T_i x_n - x_n\| \leq \eta^{-1} \max_{i\in I_n} d(x_n, S_{i,n})$. □

It is readily noted that (4.13) and (4.15) are satisfied in particular when exact projections onto the constraint sets are used instead of projections onto approximating supersets.

4.3. Comments. When compared with exact-constraint cuts, surrogate cuts display three advantages. First, they yield versatile block-iterative algorithms that offer great latitude in the selection of the constraints retained at each iteration. Since the pairs $(p_{i,n}, q'_{i,n})_{i\in I_n}$ can be computed simultaneously prior to their aggregation in (4.4), surrogate cuts therefore allow for flexible parallel implementations that can fully take advantage of multiprocessor systems. Second, the processing of a constraint does not require its exact enforcement. Rather, each constraint can be "linearized" by means of a projection onto an outer approximation to the corresponding constraint set. This procedure, illustrated in Examples 4.2 and 4.3, significantly lightens the computational burden of the algorithm when nonaffine constraints are present. Third, surrogate cuts are capable of producing deep cuts, as reported in various theoretical and numerical studies, e.g., [11], [12], [32], [44], [45]. In this connection, the problem of finding optimal weights $(w_{i,n})_{i\in I_n}$ in terms of maximizing $d(x_n, H_n)$ is addressed in [32] and [33].

5. Base construction schemes. Two approaches to the construction of bases for Algorithm 1.1 are described in this section.

5.1. Cumulative bases. Steps 2 and 3 of Algorithm 1.1 suggest an obvious candidate for a base at iteration n, namely, $D_{n+1} = Q_n = E \cap D_n \cap H_n$. This base can be rewritten as

$$D_{n+1} = E \cap \bigcap_{k=0}^{n} H_k. \tag{5.1}$$

In other words, the current base is the intersection of the initial base E with all the previous cuts. This principle is the basis for the cutting plane methods originally proposed in [8] and [31], and reconsidered from a more general viewpoint in [37]. As noted in section 1, a drawback of (5.1) is that the number of cuts accumulated to define the bases grows rapidly as the iterations proceed. In the next proposition, it is pointed out that in the present framework this phenomenon can be mitigated by discarding all the cuts that are inactive at iteration n in the construction of the subsequent bases $(D_{n+k})_{k\geq 1}$.

PROPOSITION 5.1. *Let* $\mathbb{K}_{-1} = \emptyset$ *and, for every* $n \in \mathbb{N}$, $\mathbb{A}_n = \{k \in \mathbb{K}_{n-1} \cup \{n\} \mid x_{n+1} \in \operatorname{bd} H_k\}$. *Then the set*

$$D_{n+1} = E \cap \bigcap_{k\in\mathbb{K}_n} H_k, \quad \text{where } \mathbb{A}_n \subset \mathbb{K}_n \subset \mathbb{K}_{n-1} \cup \{n\} \tag{5.2}$$

is a base for Algorithm 1.1 *at iteration* n.

Proof. We need to check that (1.4) holds for D_{n+1} as in (5.2). First, as $E \in \mathfrak{C}(S)$ and, by (1.3), $(H_k)_{k\in\mathbb{K}_n} \subset \mathfrak{C}(S)$, (5.2) implies $D_{n+1} \in \mathfrak{C}(S)$. Next, to show $x_{n+1} = \mathfrak{m}(D_{n+1})$, note that $x_{n+1} = \mathfrak{m}(Q_n)$ and $Q_n = E \cap D_n \cap H_n = E \cap \bigcap_{k\in\mathbb{K}_{n-1}\cup\{n\}} H_k = D_{n+1} \cap B_n$, where $B_n = \bigcap_{k\in(\mathbb{K}_{n-1}\cup\{n\})\smallsetminus\mathbb{K}_n} H_k$. However, it follows from the definition of $\mathbb{A}_n$ and the inclusion $\mathbb{A}_n \subset \mathbb{K}_n$ that $x_{n+1} \in B_n^\circ$ and therefore that B_n is inactive at x_{n+1}. Accordingly, $x_{n+1} = \mathfrak{m}(D_{n+1} \cap B_n) = \mathfrak{m}(D_{n+1})$. The proof is complete. ∎

In particular, if at every iteration $\mathbb{K}_n = \mathbb{K}_{n-1} \cup \{n\}$, then all the cuts are retained and (5.2) relapses to (5.1). At the other end of the spectrum, the simplest bases are obtained by discarding all the inactive cuts, i.e., by taking $\mathbb{K}_n = \mathbb{A}_n$ at every iteration.

5.2. Instantaneous bases. The construction of D_{n+1} described here was first proposed for quadratic forms in Hilbert spaces by Haugazeau in [26] and extended to the present setting in [35].

PROPOSITION 5.2. *Suppose that:*

(A4) *There exists a point* $v \in S \cap \operatorname{dom} J$ *at which* J *is continuous.*

Then, given $t'_{n+1} \in \partial J(x_{n+1})$ *such that* $(\forall x \in Q_n)\langle x_{n+1} - x, t'_{n+1}\rangle \leq 0$, *the set*

$$D_{n+1} = \big\{x \in \mathcal{X} \mid \langle x_{n+1} - x, t'_{n+1}\rangle \leq 0\big\} \tag{5.3}$$

is a base for Algorithm 1.1 *at iteration* n.

Proof. Since $S \subset Q_n$, (A4) asserts that J is continuous at $v \in Q_n \cap \operatorname{dom} J$, whence, as $x_{n+1} = \mathfrak{m}(Q_n)$, the existence of t'_{n+1} is guaranteed by (2.3). Moreover, the inclusion $Q_n \subset D_{n+1}$ shows that $D_{n+1} \in \mathfrak{C}(S)$. Finally, for $A = D_{n+1}$, (2.3) yields $x_{n+1} = \mathfrak{m}(D_{n+1})$. We have thus established (1.4). ∎

If t'_{n+1} is the zero functional (which may happen only when $x_{n+1} = \mathfrak{m}(\mathcal{X})$), then $D_{n+1} = \mathcal{X}$. On the other hand, if the Gâteaux-derivative, $\nabla J(x_{n+1})$, of J at x_{n+1} exists, then

$$D_{n+1} = \{x \in \mathcal{X} \mid \langle x_{n+1} - x, \nabla J(x_{n+1})\rangle \leq 0\}. \tag{5.4}$$

5.3. Comments. An advantage of cumulative bases is their wide applicability. However, they lead to increasingly complex outer approximations as the algorithm progresses. A partial remedy to this situation is to systematically discard all the inactive cuts. One should, however, beware of its potential side-effect, namely, slower convergence. By contrast, instantaneous bases are very attractive, for they take the form of half-spaces under the relatively mild assumption (A4). Their efficacy can, however, be limited by the search for an acceptable subgradient in (5.3). Of course, this limitation vanishes altogether when J is Gâteaux-differentiable on C, the base being then explicitly given by (5.4).

6. Examples. The analysis of the preceding sections gives rise to four general realizations of Algorithm 1.1 according to whether one selects, on the one hand, exact-constraint or surrogate cuts and, on the other hand, cumulative or instantaneous bases. In this section, these four realizations are presented and the theorems stating their strong convergence to the solution $\overline{x}$ of (P) under the standing assumptions (A1)–(A3) are given. A variety of outer approximation methods are exhibited as special cases and their convergence is deduced from the main theorems. Although we have restricted ourselves to known methods, it is clear that further convergence results can be generated by considering alternative schemes subsumed by Algorithms 6.1–6.4 below.

6.1. Exact-constraint cuts and cumulative bases. If the cuts are generated as in Proposition 4.1 and the bases as in Proposition 5.1, Algorithm 1.1 reads as follows.

ALGORITHM 6.1. *A sequence $(x_n)_{n\geq 0}$ is constructed as follows, where E is supplied by* (A2).

Step 0. *Set $D_0 = E$, $x_0 = \mathfrak{m}(D_0)$, $\mathbb{K}_{-1} = \emptyset$, and $n = 0$.*

Step 1. *Take $\mathrm{i}(n) \in I$.*

Step 2. *Set $x_{n+1} = \mathfrak{m}(D_n \cap S_{\mathrm{i}(n)})$ and $\mathbb{A}_n = \{k \in \mathbb{K}_{n-1} \cup \{n\} \mid x_{n+1} \in \operatorname{bd} S_{\mathrm{i}(k)}\}$.*

Step 3. *Take $\mathbb{A}_n \subset \mathbb{K}_n \subset \mathbb{K}_{n-1} \cup \{n\}$ and set $D_{n+1} = E \cap \bigcap_{k\in\mathbb{K}_n} S_{\mathrm{i}(k)}$.*

Step 4. *Set $n = n + 1$ and go to Step* 1.

The convergence result below is a direct application of Theorem 3.1.

THEOREM 6.1. *Let $(x_n)_{n\geq 0}$ be an arbitrary orbit of Algorithm* 6.1 *generated under either of the following conditions:* (i) *tightness, I countable, and admissible control; or* (ii) *strong tightness and coercive control. Then $x_n \stackrel{n}{\to} \overline{x}$.*

EXAMPLE 6.1 (see [5, Thm. 2.4]). *In Algorithm* 6.1, *$\mathcal{X} = \mathbb{R}^N$, E is bounded, J is finite and strictly convex, I is a compact metric space, $(g_i)_{i\in I}$ is a family of finite convex functions such that $(i, x) \mapsto g_i(x)$ is continuous on $I \times \mathcal{X}$, $\mathbb{K}_n = \mathbb{A}_n$ at Step* 3, *and the* most violated constraint *control mode*

$$(\forall n \in \mathbb{N}) \quad g_{\mathrm{i}(n)}(x_n) = \max_{i\in I} g_i(x_n) \tag{6.1}$$

is in force. Then $x_n \stackrel{n}{\to} \overline{x}$.

Proof. The conditions of Proposition 2.1(ii)(b) are fulfilled, and (A1)–(A2) are therefore satisfied. In addition, it follows from Proposition 2.2(ii) that (A3) is satisfied. For every $x \in \mathcal{X}$, the continuity of $i \mapsto g_i(x)$ on the compact space I yields $\sup_{i\in I} |g_i(x)| < +\infty$. Proposition 4.3(iv) then ensures the strong tightness of the algorithm, while (6.1) is a special instance of the coercive control mode (3.7). The claim is therefore a consequence of Theorem 6.1(ii). □

Examples 6.2 and 6.3 below are, respectively, infinite-dimensional formulations of Kelley's basic cutting plane algorithm [31] and of the Kaplan–Veinott supporting hyperplane algorithm [30], [52]. These algorithms were already shown in [35] to be special instances of the framework described in Example 4.1. The problem under consideration is to find the minimizer $\overline{x}$ of a function $J\colon \mathcal{X} \to]-\infty, +\infty]$ over a closed convex set S under assumptions (A1)–(A2). By expressing S as a suitable intersection of half-spaces $(S_i)_{i\in I}$, this problem will be recast in the form of (P).

EXAMPLE 6.2. *Suppose that $S = \mathrm{lev}_{\leq 0} g$, where $g\colon \mathcal{X} \to]-\infty, +\infty]$ is a lower semicontinuous convex function. Let E be a polyhedron and suppose that there exists $\zeta \in \mathbb{R}_+^*$ such that, for every $x \in C$, $\partial g(x) \subset B(0, \zeta)$. Let $x_0 = \mathfrak{m}(E)$ and define $(x_n)_{n\geq 0}$ by the recursion*

(6.2)
$$(\forall n \in \mathbb{N}) \quad x_{n+1} = \mathfrak{m}\Bigg(E \cap \bigcap_{k=0}^{n} \{x \in \mathcal{X} \mid \langle x_k - x, t_k' \rangle \geq g(x_k)\}\Bigg), \quad \text{where } t_n' \in \partial g(x_n).$$

Then $x_n \overset{n}{\to} \overline{x}$.

Proof. Let $I = \{(y, t') \in \mathcal{X} \times \mathcal{X}' \mid t' \in \partial g(y)\}$ be the graph of ∂g. For every $i = (y, t') \in I$, the continuous affine function $g_i\colon x \mapsto \langle x - y, t' \rangle + g(y)$ minorizes g by virtue of (2.2) with $g_i(y) = g(y)$ and it defines a closed affine half-space $S_i = \mathrm{lev}_{\leq 0} g_i$. We can then write $S = \bigcap_{i\in I} S_i$. Since at iteration $n \in \mathbb{N}$ the function $i \mapsto g_i(x_n)$ is maximized for $\mathrm{i}(n) = (x_n, t_n')$ where $t_n' \in \partial g(x_n)$, (6.2) appears as a particular realization of Algorithm 6.1 with $\mathbb{K}_n = \mathbb{K}_{n-1} \cup \{n\}$ at Step 3 and control rule (6.1). The control is therefore coercive since (6.1) $\Rightarrow$ (3.7). Moreover, if $x_n \notin S$, $g_{\mathrm{i}(n)}(x_n) = g(x_n) = \|t_n'\| \cdot d(x_n, S_{\mathrm{i}(n)}) \leq \zeta d(x_n, S_{\mathrm{i}(n)})$. However, Proposition 3.1(viii) states that $d(x_n, S_{\mathrm{i}(n)}) \overset{n}{\to} 0$. Thus, $\overline{\lim}_n g_{\mathrm{i}(n)}(x_n) \leq 0$ and the algorithm is strongly tight. The announced result then follows from Theorem 6.1(ii). □

EXAMPLE 6.3. *Suppose that $\mathcal{X}$ is a Hilbert space, that S is bounded with $S^\circ \neq \emptyset$, and that E is a polyhedron. Let $x_0 = \mathfrak{m}(E)$ and $w \in S^\circ$, and define $(x_n)_{n\geq 0}$ by the recursion*

$$(\forall n \in \mathbb{N}) \quad x_{n+1} = \mathfrak{m}\left(E \cap \bigcap_{k=0}^{n} H_k\right), \tag{6.3}$$

where $H_n \in \mathfrak{C}(S)$ is either the whole space $\mathcal{X}$ or an affine half-space whose boundary supports S at the point $y_n \in \mathrm{bd}\, S \cap [x_n, w]$, according to whether x_n lies in S or not. Then $x_n \overset{n}{\to} \overline{x}$.

Proof. Let $I = \{i \in \mathcal{X} \mid \|i\| = 1\}$ be the unit sphere in $\mathcal{X}$ and $\sigma\colon i \mapsto \sup_{x\in S} \langle x, i \rangle$ the support function of S. For every $i \in I$, define a closed affine half-space $S_i = \mathrm{lev}_{\leq 0} g_i$, where $g_i\colon x \mapsto \langle x, i \rangle - \sigma(i)$. Then $S = \bigcap_{i\in I} S_i$. By assumption, $B(w, \gamma) \subset S$ for some $\gamma \in \mathbb{R}_+^*$. Now suppose $x_n \notin S$ and let p_n be the projection of x_n onto $\mathrm{bd}\, H_n \ni y_n$. Then $\mathrm{bd}\, H_n = \{y \in \mathcal{X} \mid \langle y - y_n, p_n - x_n \rangle = 0\}$ and $d(w, \mathrm{bd}\, H_n) =$

$\langle w-y_n, p_n-x_n\rangle/\|p_n-x_n\|$. However, $y_n \in [x_n, w]$ and therefore $w-y_n = \alpha_n(y_n-x_n)$, where $\alpha_n = \|w-y_n\|/\|x_n-y_n\|$. Hence, $d(w, \operatorname{bd} H_n) = \alpha_n\langle y_n-x_n, p_n-x_n\rangle/\|p_n-x_n\|$ $= \alpha_n d(x_n, H_n)$. Consequently,

$$(\forall i \in I) \quad d(x_n, H_n) = \frac{d(w, \operatorname{bd} H_n)}{\|w-y_n\|} \cdot \|x_n-y_n\| \geq \eta d(x_n, S_i), \tag{6.4}$$

where $\eta = \gamma/\sup_{y\in\operatorname{bd} S}\|w-y\| > 0$. In addition, for every $i \in I$, $x_n \notin S_i \Rightarrow d(x_n, S_i) = g_i(x_n)$. Hence, $g_{\mathrm{i}(n)}(x_n) \geq \eta \sup_{i\in I} g_i(x_n)$, where $\mathrm{i}(n) \in I$ and $S_{\mathrm{i}(n)} = H_n$, from which it follows that (6.3) is a particular realization of Algorithm 6.1 with $\mathbb{K}_n = \mathbb{K}_{n-1} \cup \{n\}$ at Step 3 and coercive control rule (3.7). Since the affine family $(g_i)_{i\in I}$ is equi-Lipschitzian on $\mathcal{X}$ with constant 1, strong tightness follows from Proposition 4.3(i) and Theorem 6.1(ii) yields $x_n \overset{n}{\to} \overline{x}$. □

6.2. Exact-constraint cuts and instantaneous bases. Algorithm 6.2 below is derived from Algorithm 1.1 by coupling the cuts of Proposition 4.1 together with the bases of Proposition 5.2.

ALGORITHM 6.2. *A sequence $(x_n)_{n\geq 0}$ is constructed as follows, where E is supplied by* (A2).

Step 0. *Set $D_0 = E$, $x_0 = \mathfrak{m}(D_0)$, and $n = 0$.*

Step 1. *Take $\mathrm{i}(n) \in I$.*

Step 2. *Set $x_{n+1} = \mathfrak{m}(E \cap D_n \cap S_{\mathrm{i}(n)})$.*

Step 3. *Take $t'_{n+1} \in \partial J(x_{n+1})$ such that $(\forall x \in E \cap D_n \cap S_{\mathrm{i}(n)})$ $\langle x_{n+1}-x, t'_{n+1}\rangle \leq 0$ and set $D_{n+1} = \{x \in \mathcal{X} \mid \langle x_{n+1}-x, t'_{n+1}\rangle \leq 0\}$.*

Step 4. *Set $n = n+1$ and go to Step 1.*

Convergence follows at once from Theorem 3.1.

THEOREM 6.2. *Let $(x_n)_{n\geq 0}$ be an arbitrary orbit of Algorithm 6.2 generated under either of the following conditions:* (i) *tightness, I countable, and admissible control; or* (ii) *strong tightness and coercive control. Then $x_n \overset{n}{\to} \overline{x}$.*

EXAMPLE 6.4 (see [26, Thm. 2]). *In Algorithm 6.2, $\mathcal{X}$ is a Hilbert space, $E = \mathcal{X}$, J is a coercive quadratic form, $I = \{0, \ldots, M-1\}$ is finite, $(\forall i \in I)$ $g_i\colon x \mapsto d(x, S_i)$, and the* periodic *control mode*

$$(\forall n \in \mathbb{N}) \quad \mathrm{i}(n) = n \ (\text{modulo } M), \tag{6.5}$$

is in force. Then $x_n \overset{n}{\to} \overline{x}$.

Proof. The conditions of Proposition 2.1(iii) are satisfied and, consequently, so are (A1) and (A2). In addition, (A3) is secured by Proposition 2.2(ii) since $(g_i)_{i\in I}$ is a family of continuous and—by the convexity of the sets $(S_i)_{i\in I}$—convex functions. In addition, it follows from Proposition 3.1(viii) that $g_{\mathrm{i}(n)}(x_n) = d(x_n, S_{\mathrm{i}(n)}) \overset{n}{\to} 0$ and therefore that the algorithm is tight. Finally, since (6.5) $\Rightarrow$ (3.6), the control is admissible. Hence, the assertion follows from Theorem 6.2(i). □

EXAMPLE 6.5 (see [41]). *In Algorithm 6.2, $\mathcal{X}$ is a Hilbert space, $E = \mathcal{X}$, J is a coercive quadratic form, I is a compact metric space, $(g_i)_{i\in I}$ is a family of affine functions such that $(i, x) \mapsto g_i(x)$ is continuous on $I \times \mathcal{X}$, and the most violated constraint control mode* (6.1) *is in force. Then $x_n \overset{n}{\to} \overline{x}$.*

Proof. (A1)–(A2) hold by Proposition 2.1(iii). Now fix $i \in I$. Then $g_i\colon x \mapsto \langle x, z_i\rangle + \alpha_i$, where $z_i \in \mathcal{X}$ and $\alpha_i \in \mathbb{R}$, and (A3) holds. For every $x \in \mathcal{X}$, the continuity of $i \mapsto \langle x, z_i\rangle$ on the compact space I implies $\sup_{i\in I}|\langle x, z_i\rangle| < +\infty$. Therefore,

the algorithm is strongly tight by Proposition 4.3(iii) and, since (6.1) $\Rightarrow$ (3.7), it operates under coercive control. The desired conclusion is reached by invoking Theorem 6.2(ii). □

EXAMPLE 6.6 (see [39]). *In Algorithm* 6.2, $\mathcal{X} = \mathbb{R}^N$, J *is finite and strictly convex,* E *is bounded,* J *is strongly convex (i.e.,* $c : \tau \mapsto \kappa\tau^2$ *with* $\kappa \in \mathbb{R}_+^*$ *in* (A2) [36]*) and differentiable on* E, $I = \{0, \dots, M-1\}$ *is finite,* $(\forall i \in I)$ $g_i \colon x \mapsto d(x, S_i)$, *and either of the following conditions is fulfilled:* (i) *the periodic control mode* (6.5) *is in force; or* (ii) *the most violated constraint control mode* (6.1) *is in force. Then* $x_n \overset{n}{\to} \overline{x}$.

Proof. The conditions of Proposition 2.1(ii)(b)—and therefore (A1)–(A2)—hold. As in Example 6.4, (A3) also holds and the algorithm is (strongly) tight. Hence, (i) and (ii) follow from Theorem 6.2(i) and (ii), respectively. □

6.3. Surrogate cuts and cumulative bases. Algorithm 1.1 is implemented with the cuts of Proposition 4.4 and the bases of Proposition 5.1.

ALGORITHM 6.3. *A sequence* $(x_n)_{n\geq 0}$ *is constructed as follows, where* E *is supplied by* (A2).

Step 0. *Take* $(\delta, \varepsilon) \in]0,1[^2$ *and set* $D_0 = E$, $x_0 = \mathfrak{m}(D_0)$, $\mathbb{K}_{-1} = \emptyset$, *and* $n = 0$.

Step 1. *Take a finite index set* $\emptyset \neq I_n \subset I$ *and set* $H_n = \{x \in \mathcal{X} \mid \sum_{i\in I_n} w_{i,n}\langle x - p_{i,n}, q'_{i,n}\rangle \leq \gamma_n\}$, *where*

(C1) *For every* $i \in I_n$, $p_{i,n}$ *is a projection of* x_n *onto a set* $S_{i,n} \in \mathfrak{C}(S_i)$ *and* $q'_{i,n} \in \Delta(x_n - p_{i,n})$ *is such that* $(\forall x \in S_{i,n})$ $\langle x - p_{i,n}, q'_{i,n}\rangle \leq 0$.

(C2) $(w_{i,n})_{i\in I_n} \subset [0,1]$, $\sum_{i\in I_n} w_{i,n} = 1$, *and*

$$(\exists j \in I_n) \quad \begin{cases} d(x_n, S_{j,n}) = \max\limits_{i\in I_n} d(x_n, S_{i,n}), \\ w_{j,n} \geq \delta. \end{cases}$$

(C3) $0 \leq \gamma_n \leq (1-\varepsilon)\sum_{i\in I_n} w_{i,n} d(x_n, S_{i,n})^2$.

Step 2. *Set* $x_{n+1} = \mathfrak{m}(D_n \cap H_n)$ *and* $\mathbb{A}_n = \{k \in \mathbb{K}_{n-1} \cup \{n\} \mid x_{n+1} \in \operatorname{bd} H_k\}$.

Step 3. *Take* $\mathbb{A}_n \subset \mathbb{K}_n \subset \mathbb{K}_{n-1} \cup \{n\}$ *and set* $D_{n+1} = E \cap \bigcap_{k\in\mathbb{K}_n} H_k$.

Step 4. *Set* $n = n+1$ *and go to Step* 1.

Theorem 3.1 now reads as follows.

THEOREM 6.3. *Let* $(x_n)_{n\geq 0}$ *be an arbitrary orbit of Algorithm* 6.3 *generated under either of the following conditions:* (i) *tightness,* I *countable, and admissible control; or* (ii) *strong tightness and coercive control. Then* $x_n \overset{n}{\to} \overline{x}$.

It is noteworthy that Kelley's basic algorithm, presented in Example 6.2 as a special case of Algorithm 6.1, can also be viewed as a special case of Algorithm 6.3 with a single constraint set and cuts as in (4.8). Along the same lines, we present below a formulation of Kelley's algorithm with a finite number of constraints [31] under assumptions (A1)–(A2).

EXAMPLE 6.7. *In Algorithm* 6.3, I *is finite,* $(g_i)_{i\in I}$ *is a family of finite continuous convex functions satisfying* (4.16), *the approximations* (4.8) *are used, and the most violated constraint control mode* (6.1) *is in force. Then* $x_n \overset{n}{\to} \overline{x}$.

Proof. Proposition 4.8(ii) asserts that this particular realization of Algorithm 6.3 is strongly tight. Thus, since the control is coercive, the result follows from Theorem 6.3(ii). □

6.4. Surrogate cuts and instantaneous bases. The fourth implementation of Algorithm 1.1 is obtained by generating the cuts as in Proposition 4.4 and the bases as in Proposition 5.2.

ALGORITHM 6.4. *A sequence $(x_n)_{n\geq 0}$ is constructed as follows, where E is supplied by* (A2).

Step 0. *Take $(\delta,\varepsilon)\in]0,1[^2$ and set $D_0=E$, $x_0=\mathfrak{m}(D_0)$, and $n=0$.*

Step 1. *Take a finite index set $\varnothing\neq I_n\subset I$ and set $H_n=\{x\in\mathcal{X}\mid \sum_{i\in I_n}w_{i,n}\langle x-p_{i,n},q'_{i,n}\rangle\leq\gamma_n\}$, where*

(C1) *For every $i\in I_n$, $p_{i,n}$ is a projection of x_n onto a set $S_{i,n}\in\mathfrak{C}(S_i)$ and $q'_{i,n}\in\Delta(x_n-p_{i,n})$ is such that $(\forall x\in S_{i,n})\ \langle x-p_{i,n},q'_{i,n}\rangle\leq 0$.*

(C2) *$(w_{i,n})_{i\in I_n}\subset[0,1]$, $\sum_{i\in I_n}w_{i,n}=1$, and*

$$(\exists j\in I_n)\quad\begin{cases}d(x_n,S_{j,n})=\max\limits_{i\in I_n}d(x_n,S_{i,n}),\\ w_{j,n}\geq\delta.\end{cases}$$

(C3) *$0\leq\gamma_n\leq(1-\varepsilon)\sum_{i\in I_n}w_{i,n}d(x_n,S_{i,n})^2$.*

Step 2. *Set $x_{n+1}=\mathfrak{m}(E\cap D_n\cap H_n)$.*

Step 3. *Take $t'_{n+1}\in\partial J(x_{n+1})$ such that $(\forall x\in E\cap D_n\cap H_n)\langle x_{n+1}-x,t'_{n+1}\rangle\leq 0$ and set $D_{n+1}=\{x\in\mathcal{X}\mid\langle x_{n+1}-x,t'_{n+1}\rangle\leq 0\}$.*

Step 4. *Set $n=n+1$ and go to Step* 1.

The convergence conditions below are furnished by Theorem 3.1.

THEOREM 6.4. *Let $(x_n)_{n\geq 0}$ be an arbitrary orbit of Algorithm* 6.4 *generated under either of the following conditions:* (i) *tightness, I countable, and admissible control; or* (ii) *strong tightness and coercive control. Then $x_n\overset{n}{\to}\overline{x}$.*

EXAMPLE 6.8. *In Algorithm* 6.4, *$\mathcal{X}$ is a Hilbert space, I is finite, $J\colon x\mapsto\|x-w\|^2/2$ where $w\in\mathcal{X}$, $E=\mathcal{X}$, $(\forall i\in I)$ $g_i\colon x\mapsto d(x,S_i)$, $S_{i,n}=S_i$ in* (C1), *$\gamma_n=0$ in* (C3), *$x_0=w$, and one of the following conditions is fulfilled:*

(i) [27, Thm. 3-2] *The periodic control mode* (6.5) *is in force.*

(ii) [44, Thm. V.1] *The* static *control mode*

$$(\forall n\in\mathbb{N})\quad I_n=I \tag{6.6}$$

is in force and $(\forall i\in I)$ $w_{i,n}=1/\mathrm{card}I$ in (C2).

(iii) [14] *The control mode*

$$(\forall n\in\mathbb{N})\quad I_n=\Big\{i\in I\mid d(x_n,S_i)=\max_{j\in I}d(x_n,S_j)\Big\} \tag{6.7}$$

is in force and $(w_{i,n})_{i\in I_n}\subset]0,1]$ in (C2).

Then $x_n\overset{n}{\to}\overline{x}$.

Proof. First, the above setting fits into that of Proposition 2.1(iii), and therefore (A1)–(A2) hold. In addition, as in Example 6.4, (A3) holds. Furthermore, Proposition 4.6 yields $\max_{i\in I_n}g_i(x_n)=\max_{i\in I_n}d(x_n,S_{i,n})\overset{n}{\to}0$, which establishes the strong tightness of this implementation of Algorithm 6.4. Accordingly, since in (i) and (ii) the control conforms to the admissibility condition (3.3), the first two assertions follow from Theorem 6.4(i). Finally, since (6.7) is an instance of the coercive control mode (3.5) here, (iii) follows from Theorem 6.4(ii). □

EXAMPLE 6.9. *In Algorithm* 6.4, *$\mathcal{X}=\mathbb{R}^N$, J is finite, strictly convex, and differentiable with bounded lower level sets, $E=\mathcal{X}$, I is finite, $(g_i)_{i\in I}$ is a family of finite convex functions, the approximations* (4.8) *are used, and one of the following conditions is fulfilled:*

(i) [29, Thm. 1] *The most violated constraint control mode* (6.1) *is in force.*

(ii) [29, Thm. 2] *The serial admissible control mode* (3.6) *is in force.*
(iii) [29, Thm. 3] *The static control mode* (6.6) *is in force and* $(\forall (n,i) \in \mathbb{N} \times I),\ x_n \in S_i \Rightarrow w_{i,n} = 0$.

Then $x_n \overset{n}{\to} \overline{x}$.

Proof. Note that the conditions of Proposition 2.1(ii)(a) are satisfied and that (A1)–(A3) hold. Moreover, each g_i is continuous and subdifferentiable on $\mathcal{X}$ and (4.14) holds [46, Thm. 24.7]. Since I is finite, Propositions 3.2 and 4.7(ii) therefore imply that the algorithm is strongly tight. Hence, (i) follows from Theorem 6.4(ii), while (ii) and (iii) follow from Theorem 6.4(i). □

It emerges from the discussions of sections 4.3 and 5.3 that Algorithms 6.3 and 6.4, which employ surrogate cuts, are more advantageous numerically than Algorithms 6.1 and 6.2, which employ exact-constraint cuts. When instantaneous bases are easily generated, as is the case when J is differentiable on C, Algorithm 6.4 stands out as the most attractive implementation of Algorithm 1.1. Its chief asset is to generate at every iteration a simple outer approximation, namely, the intersection of two half-spaces (with the initial base E when $E \neq \mathcal{X}$). An application of Algorithm 6.4 to an important concrete problem is demonstrated next.

6.5. Projection onto an intersection of convex sets. Algorithm 6.4 is applied to the problem of finding the projection $\overline{x}$ of a point w onto the intersection of an arbitrary family of intersecting closed convex sets $(S_i)_{i\in I}$ conforming to (A3) in a real Hilbert space $\mathcal{X}$. As $J\colon x \mapsto \|x-w\|^2/2$ in (P), assumptions (A1) and (A4) are clearly satisfied and, in light of Proposition 2.1(iii), so is (A2) with $E = \mathcal{X}$.

Given $(x,y,z) \in \mathcal{X}^3$, it will be convenient to define

$$H(x,y) = \{h \in \mathcal{X} \mid \langle h - y, x - y\rangle \leq 0\} \tag{6.8}$$

and to denote by $\mathfrak{q}(x,y,z)$ the projection of x onto $H(x,y) \cap H(y,z)$. Thus, $H(x,x) = \mathcal{X}$ and, if $x \neq y$, $H(x,y)$ is a closed affine half-space onto which y is the projection of x.

ALGORITHM 6.5. *A sequence* $(x_n)_{n\geq 0}$ *is constructed as follows.*

Step 0. *Take* $(\delta,\varepsilon) \in]0,1[^2$ *and set* $x_0 = w$ *and* $n = 0$.

Step 1. *Take a finite index set* $\emptyset \neq I_n \subset I$ *and set* $z_n = x_n + \lambda_n(\sum_{i\in I_n} w_{i,n}p_{i,n} - x_n)$, *where*

(B1) *For every* $i \in I_n$, $p_{i,n}$ *is the projection of* x_n *onto a set* $S_{i,n} \in \mathfrak{C}(S_i)$.

(B2) $(w_{i,n})_{i\in I_n} \subset [0,1]$, $\sum_{i\in I_n} w_{i,n} = 1$, *and*

$$(\exists j \in I_n) \quad \begin{cases} d(x_n, S_{j,n}) = \max\limits_{i\in I_n} d(x_n, S_{i,n}), \\ w_{j,n} \geq \delta. \end{cases}$$

(B3) $\varepsilon L_n \leq \lambda_n \leq L_n \triangleq \begin{cases} \dfrac{\sum\limits_{i\in I_n} w_{i,n}\|p_{i,n} - x_n\|^2}{\left\|\sum\limits_{i\in I_n} w_{i,n}p_{i,n} - x_n\right\|^2}, & \text{if } x_n \notin \bigcap\limits_{i\in I_n} S_{i,n}, \\ 1 & \text{otherwise.} \end{cases}$

Step 2. *Set* $\pi_n = \langle x_0 - x_n, x_n - z_n\rangle$, $\mu_n = \|x_0 - x_n\|^2$, $\nu_n = \|x_n - z_n\|^2$, $\rho_n =$

$\mu_n\nu_n - \pi_n^2$, *and*

$$x_{n+1} = \mathfrak{q}(x_0, x_n, z_n) \tag{6.9}$$
$$= \begin{cases} z_n & \text{if } \rho_n = 0 \text{ and } \pi_n \geq 0, \\ x_0 + (1 + \pi_n/\nu_n)(z_n - x_n) & \text{if } \rho_n > 0 \text{ and } \pi_n\nu_n \geq \rho_n, \\ x_n + \dfrac{\nu_n}{\rho_n}(\pi_n(x_0 - x_n) + \mu_n(z_n - x_n)) & \text{if } \rho_n > 0 \text{ and } \pi_n\nu_n < \rho_n. \end{cases}$$

Step 3. *Set* $n = n + 1$ *and go to Step* 1.

In this algorithm, the update x_{n+1} is obtained in (6.9) as the projection of $x_0 = w$ onto the intersection of the half-spaces $H(x_0, x_n)$ and $H(x_n, z_n)$.

PROPOSITION 6.1. *In the present context, Algorithm* 6.4 *reduces to Algorithm* 6.5.

Proof. Since $E = \mathcal{X}$, we obtain $x_0 = \mathfrak{m}(\mathcal{X}) = w$ at Step 0 of Algorithm 6.4. Next, recall that the cut H_n at Step 1 of Algorithm 6.4 is given by (4.7). We shall now show $H_n = H(x_n, z_n)$. Assume $x_n \notin \bigcap_{i\in I_n} S_{i,n}$ and define $y_n = x_n - \sum_{i\in I_n} w_{i,n}p_{i,n}$ ($\neq 0$ by Proposition 4.5), $\sigma_n^2 = \sum_{i\in I_n} w_{i,n}d(x_n, S_{i,n})^2$, and $\lambda_n = (\sigma_n^2 - \gamma_n)/\|y_n\|^2$. Then $z_n = x_n - \lambda_n y_n$, $L_n = \sigma_n^2/\|y_n\|^2$, and (B3) $\Leftrightarrow$ (C3). Moreover, for every $x \in \mathcal{X}$, we have

$$\begin{aligned} x \in H_n &\Leftrightarrow \langle x, y_n\rangle \leq \sum_{i\in I_n} w_{i,n}\langle p_{i,n}, x_n - p_{i,n}\rangle + \gamma_n \\ &\Leftrightarrow \langle x, y_n\rangle \leq \langle x_n, y_n\rangle - \lambda_n\|y_n\|^2 \\ &\Leftrightarrow \langle x - z_n, y_n\rangle \leq 0 \\ &\Leftrightarrow \langle x - z_n, x_n - z_n\rangle \leq 0. \end{aligned} \tag{6.10}$$

Consequently, $H_n = H(x_n, z_n)$. Next, observe that (5.4) yields $D_n = H(x_0, x_n)$. Hence, as $E = \mathcal{X}$, (6.9) coincides with Step 2 of Algorithm 6.4; the expression for $\mathfrak{q}(x, y, z)$ in terms of x, y, and z is drawn from [27, Thm. 3-1]. Note that all the possible cases are exhausted in (6.9) since $\rho_n \geq 0$ and, as also shown in [27, Thm. 3-1], $H(x_0, x_n) \cap H(x_n, z_n) = \emptyset \Leftrightarrow \rho_n = 0$ and $\pi_n < 0$. ∎

Naturally, Algorithm 6.5 contains those described in Example 6.8 as particular instances. Unlike them, however, it can handle an infinite number of constraints, approximate projections, and flexible block-iterative control modes. Strong convergence conditions are given in Theorem 6.4.

For comparison purposes, let us now review alternative iterative schemes that generate sequences converging strongly to the sought projection $\overline{x}$. From an algorithmic standpoint, these schemes are initialized with $x_0 = w$ and operate either in the serial format

$$(\forall n \in \mathbb{N}) \quad \mathrm{i}(n) \in I \quad \text{and} \quad x_{n+1} = R_{\mathrm{i}(n),n}x_n \tag{6.11}$$

or in the static parallel format

$$(\forall n \in \mathbb{N}) \quad x_{n+1} = \sum_{i\in I} w_i R_{i,n}x_n \quad \text{with} \quad (w_i)_{i\in I} \subset]0, 1] \quad \text{and} \quad \sum_{i\in I} w_i = 1, \tag{6.12}$$

where I is assumed to be countable and $(R_{i,n})_{(i,n)\in I\times\mathbb{N}}$ is a family of operators from $\mathcal{X}$ into $\mathcal{X}$. Henceforth, $(P_i)_{i\in I}$ designates the family of projectors onto the sets $(S_i)_{i\in I}$.

(1) *Periodic projection method.* Suppose that $(S_i)_{i\in I}$ is a finite family of M closed vector subspaces and set $R_{i,n} = P_i$. Then it was shown in [25] that under the periodic control mode (6.5) the serial projection method (6.11) converges strongly to $\overline{x}$. This result remains valid in the case of closed affine subspaces and it coincides with von Neumann's alternating projection theorem for $M = 2$ (see [18] for further details).

(2) *Dykstra-like methods.* In [6], an extension of the preceding periodic projection method to finite families of closed convex sets was obtained by setting $R_{i,n}x_n = P_i(x_n + y_{i,n})$, where $y_{i,n}$ is the outward normal vector resulting from the previous projection onto S_i. In [21], this serial algorithm was examined from a dual perspective and given an elegant and natural interpretation; moreover, the convergence of its parallel counterpart (6.12) was established. (See also [3] for further analysis.) New developments were reported in [28], where a nonperiodic control mode was used in (6.11) and countably infinite families of sets were considered.

(3) *Anchor point methods.* Suppose that, for every $i \in I$, $S_i = \operatorname{Fix} T_i$ where $T_i\colon \mathcal{X} \to \mathcal{X}$ is firmly nonexpansive, i.e., satisfies (2.15). (Note that if, for some $i \in I$, T_i is merely nonexpansive, it can be replaced by the averaged mapping $T_i^{\mathrm{av}} = (T_i + \mathrm{Id})/2$ which is firmly nonexpansive [24, Thm. 12.1] and satisfies $\operatorname{Fix} T_i^{\mathrm{av}} = \operatorname{Fix} T_i$.) Anchor point methods operate with $R_{i,n}x_n = \alpha_n x_0 + (1-\alpha_n)T_i x_n$, where $(\alpha_n)_{n\geq 0} \subset]0,1]$ converges "slowly" to 0 (e.g., $\alpha_n = 1/(n+1)$ in [2] and [10]). Strong convergence was established in [2] and [38] for the serial version (6.11) under the periodic control rule (6.5) (and I finite) and in [10] for the parallel version (6.12) with I countably infinite.

(4) *Periodic quasi-projection method.* This method, proposed in [26], was described in Example 6.4 as an offspring of Algorithm 6.2. It is equivalent to executing (6.11) under the periodic control mode (6.5) and with $R_{\mathrm{i}(n),n}x_n$ as the "quasi-projection" of x_n onto $S_{\mathrm{i}(n)}$, i.e., the projection of x_n onto $S_{\mathrm{i}(n)} \cap H(x_0, x_n)$.

Overall, Algorithm 6.5 appears to enjoy more flexibility than the above methods in terms of parallel implementation and more versatility in terms of the types of constraints it can handle. Indeed, Dykstra-like and anchor point methods are not well suited for parallel block-processing due to their serial or static parallel structure. The scope of Dykstra-like methods is further limited by the fact that they require the ability to compute projections, which is possible only in special situations. In this regard, anchor point methods are somewhat less restrictive, as any firmly nonexpansive mapping admitting the set under consideration as fixed point set can be used. In addition, Dykstra-like methods require that a normal vector be carried along for each set (except for affine subspaces), which makes their implementation costly in terms of memory allocation and management. Finally, it is noted that the quasi-projection method is a rather conceptual one, the computation of quasi-projections being usually a serious obstacle to its implementability in practice.

7. Further results. In this section, we present convergence results for two variants of (P) in which the original assumptions are altered. $\mathcal{X}$ is assumed to be a Hilbert space.

7.1. Inconsistent constraints. It has been assumed so far that the constraints are consistent, i.e., that $S \neq \emptyset$ in (P). In this section, we place ourselves in the following context: S may be empty and I is finite. As before, $(P_i)_{i\in I}$ are the projectors onto the closed convex sets $(S_i)_{i\in I}$.

As in the convex feasibility problems of [9] and [17], the exact, but possibly empty, feasibility set S can be replaced by the set $\widetilde{S}$ of points which best approximate the constraints in an averaged squared-distance sense. Fix weights $(w_i)_{i\in I} \subset]0,1]$ such that $\sum_{i\in I} w_i = 1$, define a (continuous and convex) *proximity function* $\Phi\colon x \mapsto (1/2)\sum_{i\in I} w_i d(x,S_i)^2$, and let $\widetilde{S}$ be the (closed and convex) set of minimizers of Φ over $\mathcal{X}$. (P) is then replaced by

$$(\widetilde{\mathrm{P}}) \qquad \text{find } \widetilde{x} \in \widetilde{S} \quad \text{such that } J(\widetilde{x}) = \inf_{x\in\widetilde{S}} J(x)$$

under assumptions (A1) and

(A0) $\widetilde{S} \neq \emptyset$,

($\widetilde{\text{A2}}$) for some $\widetilde{E} \in \mathfrak{C}(\widetilde{S})$, there exists a point $\widetilde{u} \in \widetilde{S} \cap \operatorname{dom} J$ such that $\widetilde{C} \triangleq \widetilde{E} \cap \operatorname{lev}_{\leq J(\widetilde{u})} J$ is bounded and J is uniformly convex on $\widetilde{C}$.

Some remarks are in order. First, if $S \neq \emptyset$, then $\widetilde{S} = S$. Second, if $S = \emptyset$, then assumption (A0) holds when one of the sets in $(S_i)_{i\in I}$ is bounded or when they are all closed affine half-spaces [17]. Third, it follows from (A0), (A1), and ($\widetilde{\text{A2}}$) that ($\widetilde{\text{P}}$) admits a unique solution $\widetilde{x}$.

The next step is to regard ($\widetilde{\text{P}}$) as a program of the general form (P) with a single constraint set, namely $\widetilde{S}$. Consequently, ($\widetilde{\text{P}}$) can be solved via Algorithms 6.3 or 6.4 by constructing suitable surrogate cuts for $\widetilde{S}$.

THEOREM 7.1. *Let $(x_n)_{n\geq 0}$ be an arbitrary orbit of Algorithms* 6.3 *or* 6.4 *in which the cut at Step* 1 *is taken to be*

$$(7.1) \qquad H_n = \left\{ x \in \mathcal{X} \mid \left\langle x - \sum_{i\in I} w_i P_i x_n, x_n - \sum_{i\in I} w_i P_i x_n \right\rangle \leq \gamma_n \right\},$$

where $0 \leq \gamma_n \leq (1-\varepsilon)\|x_n - \sum_{i\in I} w_i P_i x_n\|^2$. Then $x_n \overset{n}{\to} \widetilde{x}$.

Proof. The claim follows from Theorems 6.3(i) and 6.4(i). Indeed, the control is admissible since only one constraint set is present. Next, let us show that (7.1) is a valid cut at iteration n. To this end, let $T = \sum_{i\in I} w_i P_i$. Then T is firmly nonexpansive and $\operatorname{Fix} T = \widetilde{S}$ [9]. Hence (A3) holds by Proposition 2.2(iv)(e) and (7.1) is drawn from (4.10). Finally, tightness follows from Proposition 4.7(iii). ∎

7.2. Convex feasibility problems. If, instead of (A2), it is assumed that J is constant on S, then (P) turns into the convex feasibility problem

$$(\mathrm{CFP}) \qquad \text{find } \overline{x} \in S = \bigcap_{i\in I} S_i.$$

A general strategy for solving (CFP) is to construct a sequence $(x_n)_{n\geq 0}$ in which x_{n+1} is a relaxed projection of x_n onto a cut H_n. An implementation of this outer approximation scheme with surrogate cuts leads to the following block-iterative algorithm.

ALGORITHM 7.1. *In Algorithm* 6.5, *pick x_0 arbitrarily at Step* 0, *extend the relaxation range in* (B3) *to "$\varepsilon \leq \lambda_n \leq (2-\varepsilon)L_n$," and reduce Step* 2 *to "$x_{n+1} = z_n$."*

THEOREM 7.2. *Let $(x_n)_{n\geq 0}$ be an arbitrary orbit of Algorithm* 7.1 *generated under either of the following conditions:* (i) *tightness, I countable, and admissible control or* (ii) *strong tightness and coercive control. Then $(x_n)_{n\geq 0}$ converges weakly to a point in S.*

Proof. A slight modification of the results of [13, section 2] shows that $x_{n+1} - x_n \overset{n}{\to} 0$, $\max_{i\in I_n} d(x_n, S_{i,n}) \overset{n}{\to} 0$, and $(x_n)_{n\geq 0}$ converges weakly to a point in S if $\mathfrak{W}(x_n)_{n\geq 0} \subset S$, from which, by arguing along the same lines as in Theorem 3.1(i) (respectively, Theorem 3.1(ii)), we obtain (i) (respectively, (ii)). □

It follows from Propositions 4.7 and 4.8 that Theorem 7.2 covers the weak convergence results of [10], [12], and [13]. A closely related algorithm is proposed in [34, section 11] with similar weak convergence results.

Acknowledgments. The author wishes to express his gratitude to the two anonymous referees for their valuable remarks.

REFERENCES

[1] V. Barbu and Th. Precupanu, *Convexity and Optimization in Banach Spaces*, 2nd ed., D. Reidel, Boston, MA, 1986.
[2] H. H. Bauschke, *The approximation of fixed points of compositions of nonexpansive mappings in Hilbert space*, J. Math. Anal. Appl., 202 (1996), pp. 150–159.
[3] H. H. Bauschke and J. M. Borwein, *Dykstra's alternating projection algorithm for two sets*, J. Approx. Theory, 79 (1994), pp. 418–443.
[4] H. H. Bauschke and J. M. Borwein, *On projection algorithms for solving convex feasibility problems*, SIAM Rev., 38 (1996), pp. 367–426.
[5] J. W. Blankenship and J. E. Falk, *Infinitely constrained optimization problems*, J. Optim. Theory Appl., 19 (1976), pp. 261–281.
[6] J. P. Boyle and R. L. Dykstra, *A method for finding projections onto the intersection of convex sets in Hilbert spaces*, in Advances in Order Restricted Statistical Inference, Lecture Notes in Statist. 37, Springer-Verlag, New York, 1986, pp. 28–47.
[7] H. Brézis, *Analyse Fonctionnelle*, 2nd ed., Masson, Paris, 1993.
[8] E. W. Cheney and A. A. Goldstein, *Newton's method for convex programming and Tchebycheff approximation*, Numer. Math., 1 (1959), pp. 253–268.
[9] P. L. Combettes, *Inconsistent signal feasibility problems: Least-squares solutions in a product space*, IEEE Trans. Signal Process., 42 (1994), pp. 2955–2966.
[10] P. L. Combettes, *Construction d'un point fixe commun à une famille de contractions fermes*, C. R. Acad. Sci. Paris Sér. I Math., 320 (1995), pp. 1385–1390.
[11] P. L. Combettes, *The convex feasibility problem in image recovery*, in Advances in Imaging and Electron Physics, P. Hawkes, ed., vol. 95, Academic Press, New York, 1996, pp. 155–270.
[12] P. L. Combettes, *Convex set theoretic image recovery by extrapolated iterations of parallel subgradient projections*, IEEE Trans. Image Process., 6 (1997), pp. 493–506.
[13] P. L. Combettes, *Hilbertian convex feasibility problem: Convergence of projection methods*, Appl. Math. Optim., 35 (1997), pp. 311–330.
[14] G. Crombez, *Finding projections onto the intersection of convex sets in Hilbert spaces*, Numer. Funct. Anal. Optim., 16 (1995), pp. 637–652.
[15] R. Dautray and J.-L. Lions, *Analyse Mathématique et Calcul Numérique pour les Sciences et les Techniques*, Masson, Paris, 1984. English translation: Springer-Verlag, New York, 1988.
[16] M. A. H. Dempster and R. R. Merkovsky, *A practical geometrically convergent cutting plane algorithm*, SIAM J. Numer. Anal., 32 (1995), pp. 631–644.
[17] A. R. De Pierro and A. N. Iusem, *A parallel projection method for finding a common point of a family of convex sets*, Pesqui. Oper., 5 (1985), pp. 1–20.
[18] F. Deutsch, *The method of alternating orthogonal projections*, in Approximation Theory, Spline Functions and Applications, S. P. Singh, ed., Kluwer, The Netherlands, 1992, pp. 105–121.
[19] B. C. Eaves and W. I. Zangwill, *Generalized cutting plane algorithms*, SIAM J. Control, 9 (1971), pp. 529–542.
[20] R. E. Edwards, *Functional Analysis—Theory and Applications*, 2nd ed., Dover, New York, 1995.

[21] N. Gaffke and R. Mathar, *A cyclic projection algorithm via duality*, Metrika, 36 (1989), pp. 29–54.
[22] U. García-Palomares, *Parallel projected aggregation methods for solving the convex feasibility problem*, SIAM J. Optim., 3 (1993), pp. 882–900.
[23] F. Glover, *A multiphase-dual algorithm for the zero-one integer programming problem*, Oper. Res., 13 (1965), pp. 879–919.
[24] K. Goebel and W. A. Kirk, *Topics in Metric Fixed Point Theory*, Cambridge University Press, Cambridge, UK, 1990.
[25] I. Halperin, *The product of projection operators*, Acta Sci. Math. (Szeged), 23 (1962), pp. 96–99.
[26] Y. Haugazeau, *Sur la minimisation de formes quadratiques avec contraintes*, C. R. Acad. Sci. Paris Sér. A Math., 264 (1967), pp. 322–324.
[27] Y. Haugazeau, *Sur les Inéquations Variationnelles et la Minimisation de Fonctionnelles Convexes*, Thèse, Université de Paris, Paris, France, 1968.
[28] H. Hundal and F. Deutsch, *Two generalizations of Dykstra's cyclic projections algorithm*, Math. Programming, 77 (1997), pp. 335–355.
[29] A. N. Iusem and B. F. Svaiter, *A row-action method for convex programming*, Math. Programming, 64 (1994), pp. 149–171.
[30] A. A. Kaplan, *Determination of the extremum of a linear function on a convex set*, Soviet Math. Dokl., 9 (1968), pp. 269–271.
[31] J. E. Kelley, *The cutting-plane method for solving convex programs*, J. SIAM, 8 (1960), pp. 703–712.
[32] K. C. Kiwiel, *Block-iterative surrogate projection methods for convex feasibility problems*, Linear Algebra Appl., 215 (1995), pp. 225–259.
[33] K. C. Kiwiel, *Monotone Gram matrices and deepest surrogate inequalities in accelerated relaxation methods for convex feasibility problems*, Linear Algebra Appl., 252 (1997), pp. 27–33.
[34] K. C. Kiwiel and B. Łopuch, *Surrogate projection methods for finding fixed points of firmly nonexpansive mappings*, SIAM J. Optim., 7 (1997), pp. 1084–1102.
[35] P. J. Laurent and B. Martinet, *Méthodes duales pour le calcul du minimum d'une fonction convexe sur une intersection de convexes*, in Symposium on Optimization, Lecture Notes in Math. 132, Springer-Verlag, New York, 1970, pp. 159–180.
[36] E. S. Levitin, *Perturbation Theory in Mathematical Programming and Its Applications*, Wiley, New York, 1994.
[37] E. S. Levitin and B. T. Polyak, *Constrained minimization methods*, USSR Comput. Math. Math. Phys., 6 (1966), pp. 1–50.
[38] P.-L. Lions, *Approximation de points fixes de contractions*, C. R. Acad. Sci. Paris Sér. A Math., 284 (1977), pp. 1357–1359.
[39] A. I. Lobyrev, *The minimization of a strictly convex function on an intersection of convex sets*, Optimizacija, 5 (1972), pp. 128–132.
[40] Y. I. Merzlyakov, *On a relaxation method of solving systems of linear inequalities*, USSR Comput. Math. Math. Phys., 2 (1963), pp. 504–510.
[41] W. Oettli, *Solving optimization problems with many constraints by a sequence of subproblems containing only two constraints*, Math. Nachr., 71 (1976), pp. 143–145.
[42] N. Ottavy, *Strong convergence of projection-like methods in Hilbert spaces*, J. Optim. Theory Appl., 56 (1988), pp. 433–461.
[43] N. S. Papageorgiou, *Convex integral functionals*, Trans. Amer. Math. Soc., 349 (1997), pp. 1421–1436.
[44] G. Pierra, *Eclatement de contraintes en parallèle pour la minimisation d'une forme quadratique*, in Lecture Notes in Comput. Sci. 41, Springer-Verlag, New York, 1976, pp. 200–218.
[45] G. Pierra, *Decomposition through formalization in a product space*, Math. Programming, 28 (1984), pp. 96–115.
[46] R. T. Rockafellar, *Convex Analysis*, Princeton University Press, Princeton, NJ, 1970.
[47] R. T. Rockafellar, *Conjugate Duality and Optimization*, SIAM, Philadelphia, PA, 1974.
[48] I. Singer, *Best Approximation in Normed Linear Spaces by Elements of Linear Subspaces*, Springer-Verlag, New York, 1970.
[49] D. M. Topkis, *Cutting-plane methods without nested constraint sets*, Oper. Res., 18 (1970), pp. 404–413.
[50] D. M. Topkis, *A cutting-plane algorithm with linear and geometric rates of convergence*, J. Optim. Theory Appl., 36 (1982), pp. 1–22.
[51] V. V. Vasin, *Iterative methods for solving ill-posed problems with a priori information in Hilbert spaces*, USSR Comput. Math. Math. Phys., 28 (1988), pp. 6–13.

[52] A. F. Veinott, *The supporting hyperplane method for unimodal programming*, Oper. Res., 15 (1967), pp. 147–152.
[53] A. A. Vladimirov, Yu. E. Nesterov, and Yu. N. Čekanov, *Uniformly convex functionals*, Vestnik Moskov. Univ. Ser. XV Vychisl. Mat. Kibernet., 3 (1978), pp. 12–23.
[54] C. Zălinescu, *On uniformly convex functions*, J. Math. Anal. Appl., 95 (1983), pp. 344–374.
[55] W. I. Zangwill, *Nonlinear Programming—A Unified Approach,* Prentice-Hall, Englewood Cliffs, NJ, 1969.

SIAM J. CONTROL OPTIM.
Vol. 38, No. 2, pp. 566–581

VARIATIONAL PRINCIPLES AND WELL-POSEDNESS IN OPTIMIZATION AND CALCULUS OF VARIATIONS*

ALEXANDER D. IOFFE† AND ALEXANDER J. ZASLAVSKI†

Abstract. The concluding result of the paper states that variational problems are generically solvable (and even well-posed in a strong sense) without the convexity and growth conditions always present in individual existence theorems. This and some other generic well-posedness theorems are obtained as realizations of a general variational principle extending the variational principle of Deville–Godefroy–Zizler.

Key words. calculus of variations, variational principle, existence of solutions, well-posed optimization problem

AMS subject classifications. Primary, 49J99, 90C31; Secondary, 54E35

PII. S0363012998335632

1. Introduction.

The Tonelli existence theorem in the calculus of variations [1] and its subsequent modifications (e.g., [2, 3, 4, 6, 5, 7, 8]) are based on two fundamental hypotheses concerning the behavior of the integrand as a function of the last argument (derivative): (1) the integrand should grow superlinearly at infinity and (2) that it should be convex (or exhibit a more special convexity property in case of a multiple integral with vector-valued functions) with respect to the last variable. The first hypothesis guarantees relative weak compactness of level sets of the functional and the second (together with the first) that it is lower semicontinuous (l.s.c.) in the same weak topology. In the absence of the latter the existence can still be proved but for the convexified or relaxed problem, while in the absence of the first no existence can be established at all. A few exceptions (e.g., [9, 7, 10, 11]) with very restrictive additional requirements rather confirm the need of the assumptions for general *individual* existence theorems.

Surprisingly, none of these two fundamental hypotheses is needed *generically*, and not only for the existence but also for uniqueness of a solution and even for well-posedness of the problem in a very strong sense (with respect to some special topology in the space of integrands). This is the content of the concluding theorem of this paper which we call the *antirelaxation theorem*. This theorem extends to multiple integrals and to variational problems with extended-real-valued integrands, in particular, to variational problems with explicit (nonfunctional) constraints on state variables and their derivatives.

The possibility to get generic well-posedness (with respect to variations of the integrand of the integral functional) without the convexity assumption was for the first time established by the second author [12] for a class of optimal control problems satisfying the Cesari growth condition.[1] Meanwhile, the first author was for some

*Received by the editors March 16, 1998; accepted for publication (in revised form) June 15, 1999; published electronically February 2, 2000. The research of the second author was partly supported by US-Israel Binational Scientific Fund grant 94-237.

http://www.siam.org/journals/sicon/38-2/33563.html

†Department of Mathematics, Technion, Haifa 32000, Israel (ioffe@math.technion.ac.il, ajzasl@tx.technion.ac.il).

[1] In a certain specific situation this possibility was mentioned by Asplund as early as in 1968 in his seminal paper [13]. We are thankful to the referee who drew our attention to this part of Asplund's paper.

time involved in a search of a general variational principle [14]. The main stimulus for this study has come from understanding that a common ground for the quoted result of [12] and some other generic well-posedness theorems (e.g., [15, 16]) should be looked for in the form of some "universal" variational principle probably having its prototype in the most recent variational principle of Deville–Godefroy–Zizler [17].

The variational principle of Deville–Godefroy–Zizler can also be interpreted as a statement that for a certain class of functions the minimization problem is uniquely solvable generically, and a closer look at its proof allows the generic well-posedness of the problem to be distilled. In this interpretation, however, the principle applies to a rather narrow class of minimization problems and it turns out that the main restriction comes from the assumption that there must be a continuous bump function by which (or certain transformations of which) we can perturb the functions to be minimized. This assumption effectively excludes such classes of functions or problems as, say, convex functions or problems of calculus of variations.

The key result of this paper, the *generic variational principle* of Theorem 2.2, is a modification of Deville–Godefroy–Zizler's variational principle which allows a substantially broader class of perturbations. As consequences of the theorem we get the already mentioned antirelaxation theorem, the theorem of Beer–Lucchetti [15] about generic well-posedness of minimization for convex l.s.c. functions, its extension to quasi-convex functions (which is probably a new result), a simple proof of a genericity theorem of Asplund [13] along with a characterization of Asplund spaces basically due to Ekeland–Lebourg [18], extensions of some well-posedness theorems collected in [16], in particular a theorem of Revalski [19] dealing with restrictions of continuous functions to closed sets, and the variational principle of Deville–Godefroy–Zizler itself. None of these results, with the obvious exception of the last, follows from the principle of Deville–Godefroy–Zizler. (Note that our main results as well as the variational principles of Ekeland, Borwein–Preiss, and Deville–Godefroy–Zizler apply to arbitrary l.s.c. functions but need a metric or a uniform structure in the domain space. We refer to [20] for a survey of results concerning generic well-posedness of minimization for continuous functions on completely regular topological spaces. It is to be observed that in the context of well-posedness continuity is often a restrictive requirement. The calculus of variations is the most important example of that sort.)

The proofs of all mentioned applications of the main theorem are built according to more or less the same scheme which consists in verification in each case of some basic hypotheses (H) introduced in the next section. We understand that the same scheme can easily be applied to get various variations and strengthenings of the results presented in this paper. (For example, it is easy to extend the antirelaxation theorem to variational problems over functions with values in infinite-dimensional spaces.) But we have stated the main theorem with a number of more complicated and so far unsolved problems in mind; first of all, problems with functional constraints (e.g., the problem of Lagrange in the calculus of variations or optimal control problems) in which constraint functions and maps are also subjects of variation as well as the cost function.

2. Generic variational principle. We shall consider two complete metric spaces (X, ρ) and $(\mathcal{A}, d)$, the first being called the *domain space* and the second the *data space*. We shall further assume that with every $a \in \mathcal{A}$ an l.s.c. function f_a on X is associated with values in $\overline{\mathbb{R}} = [-\infty, \infty]$ and none of these functions is identical $+\infty$. The following is the basic hypotheses about the functions which we adopt in the theorem:

(H) *There is a dense subset $\mathcal{B} \subset \mathcal{A}$, such that for any $a \in \mathcal{B}$, any $\varepsilon > 0$, and any*

$\gamma > 0$ there exist a nonempty open set $U \subset \mathcal{A}$, $x \in X$, $\alpha \in \mathbb{R}$, and $\eta > 0$ such that for any $b \in U$

(i) *$d(a,b) < \varepsilon$ and $\inf f_b > -\infty$;*

(ii) *if $z \in X$ is such that $f_b(z) \le \inf f_b + \eta$, then $\rho(z,x) \le \gamma$ and $|f_b(z) - \alpha| \le \gamma$.*

DEFINITION 2.1. *Given an $a \in \mathcal{A}$, we say that the problem of minimization of f_a on X is well-posed with respect to data in $\mathcal{A}$ (or just with respect to $\mathcal{A}$) if*

(1) *$\inf f_a$ is finite and attained at a unique point $x_a \in X$;*

(2) *for any $\varepsilon > 0$ there is a $\delta > 0$ such that $\inf f_b > -\infty$ for any $b \in \mathcal{A}$ satisfying $d(a,b) < \delta$ and, moreover, any $z \in X$ for which $f_b(z) \le \inf f_b + \delta$ satisfies $\rho(z, x_a) < \varepsilon$, $|f_b(z) - f(x_a)| < \varepsilon$.*

(In a slightly different setting a similar property was introduced in [21].)

THEOREM 2.2 (generic variational principle). *Assume* (H). *Then, the minimization problem for f_a is well-posed with respect to $\mathcal{A}$ for a generic $a \in \mathcal{A}$. In other words, there is a dense G_δ subset $\mathcal{A}' \subset \mathcal{A}$ such that for any $a \in \mathcal{A}'$ the minimization problem for f_a is well-posed with respect to $\mathcal{A}$.*

Proof. Take an $a \in \mathcal{B}$. By (H) for any natural $n = 1, 2, \dots$ there is a nonempty open set $U(a,n) \subset \mathcal{A}$, $x(a,n) \in X$, and numbers $\alpha(a,n)$ and $\eta(a,n) > 0$ such that for any $b \in U(a,n)$ we have

$$d(a,b) < 1/n; \ \inf f_b > -\infty; \tag{2.1}$$

and, whenever $f_b(z) \le \inf f_b + \eta(a,n)$,

$$\rho(z, x(a,n)) < 1/n, \qquad |f_b(z) - \alpha(a,n)| < 1/n. \tag{2.2}$$

Consider the set

$$\mathcal{A}_n = \bigcup_{\substack{a \in \mathcal{B} \\ m \ge n}} U(a,m).$$

Clearly, this set is dense and open, so by the Baire category theorem $\mathcal{A}' = \bigcap \mathcal{A}_n$ is a dense G_δ subset of $\mathcal{A}$.

Let $a \in \mathcal{A}'$. Then there is a sequence $\{a_n\} \subset \mathcal{B}$ and a sequence $\{k_n\} \to \infty$ of natural numbers such that $a \in U(a_n, k_n) = U_n$. Set $x_n = x(a_n, k_n)$, $\alpha_n = \alpha(a_n, k_n)$, and $\eta_n = \eta(a_n, k_n)$. We assume without loss of generality that $\eta_n \to 0$ decreasingly. Let z_n be such that $f_a(z_n) < \inf f_a + \eta_n$. Then $f_a(z_n) < \inf f_a + \eta_m$ if $m \le n$, so by (2.2) $\rho(z_n, x_m) \le 1/m$. It follows that $\rho(z_n, z_{n+k}) \le 2/m$ whenever $n \ge m$, that is, $\{z_n\}$ is a Cauchy sequence. Set $x_a = \lim z_n$. As f_a is l.s.c., we have $f_a(x_a) = \inf f_a$. Clearly, f_a does not have another minimizer for otherwise we would be able to construct a nonconvergent sequence $\{z_n\}$. This proves the first part of the theorem. We further note that by (2.2) $f_a(x_a) = \lim \alpha_n$.

To prove the second part, consider a sequence $\{b_n\} \subset \mathcal{A}$ converging to a, and let w_n be such that $\xi_n = f_{b_n}(w_n) - \inf f_{b_n} \to 0$. Choose $n(m)$ such that $b_n \in U_m$ and $\xi_n \le \eta_m$ for $n \ge n(m)$. For such n we have by (2.2) $\rho(w_n, x_m) \le 1/m$ and $|f_{b_n}(w_n) - \alpha_m| \le 1/m$. As $x_m \to x_a$ and $\alpha_m \to f_a(x_a)$, it follows that $w_n \to x_a$ and $f_{b_n}(w_n) \to f_a(x_a)$. This completes the proof.

Remark 2.3. As follows from the proof, there is no need to require that $(\mathcal{A}, d)$ be a complete metric space; it is sufficient to assume that it is a *Baire* space, that is, a space in which the Baire category theorem holds.

To conclude the section we observe that the variational principle of Deville–Godefroy–Zizler [17] is an immediate consequence of the theorem. To prove this

we consider a Banach space $\mathcal{A}$ of bounded continuous functions on a Banach space X with the following three properties:

(a) the norm topology in $\mathcal{A}$ is not weaker than the topology of uniform convergence on X: $\|a\|_{\mathcal{A}} \geq \sup\{|a(x)| : x \in X\}$;

(b) $\mathcal{A}$ contains compositions of its elements with translations and homotheties of X and $\|a(t+\cdot)\|_{\mathcal{A}} = \|a\|_{\mathcal{A}}$ for each $a \in \mathcal{A}$ and each $t \in X$;

(c) $\mathcal{A}$ contains a bump function, that is to say, a function $\varphi(x)$ supported on the unit ball and satisfying $0 \leq \varphi(x) \leq 1 = \varphi(0)$.

Under these assumptions, the variational principle of Deville–Godefroy–Zizler states that *for any proper l.s.c. and bounded below function f on X the set of $a \in \mathcal{A}$ for which $f + a$ attains a unique minimum on X is a dense G_δ subset of $\mathcal{A}$.*

To prove the statement we set $f_a = f + a$ and, given $a \in \mathcal{A}$, $\varepsilon > 0, \gamma > 0$, choose $\varepsilon_0 \in (0, \min\{\varepsilon/4, \gamma/4\})$ such that $\varepsilon_0 \|\psi\|_{\mathcal{A}} < 4^{-1}\varepsilon$, where $\psi(x) = \varphi(\gamma^{-1}x)$, $x \in X$, take an $\bar{x} \in X$ with $f_a(\bar{x}) < \inf f_a + \varepsilon_0/5$, and set $\bar{a}(x) = a(x) - (4\varepsilon_0/5)\psi(x - \bar{x})$, $U = \{b \in \mathcal{A} : \|b - \bar{a}\|_{\mathcal{A}} < \varepsilon_0/10\}$, $\alpha = f_{\bar{a}}(\bar{x})$, and $\eta = \varepsilon_0/10$. It is an easy matter, then, to verify that for any $b \in U$ and $z \in X$ with $f_b(z) \leq \inf f_b + \eta$

$$\alpha - \varepsilon_0 \leq f_{\bar{a}}(z) < f_b(z) + \varepsilon_0/10 \leq \inf f_b + \varepsilon_0/5 \leq \inf f_{\bar{a}} + 3\varepsilon_0/10 \leq \alpha + 3\varepsilon_0/10 \leq \inf f_a,$$

which implies both $|\alpha - f_b(z)| \leq 2\varepsilon_0 < \gamma$ and $f_{\bar{a}}(z) < \inf f_a$. Then $\|z - \bar{x}\| \leq \gamma$ because otherwise by the definition of $\bar{a}$ $f_a(z) = f_{\bar{a}}(z)$, a contradiction.

We shall see that a similar chain of arguments works in many other cases.

3. Epi-distance topology. In what follows we shall fix a certain "zero element" $\theta \in X$, set $\|x\| = \rho(\theta, x)$, $B(r) = \{x : \|x\| \leq r\}$, and agree to call a set $C \subset X$ *bounded* if $C \subset B(r)$ for some r.

Let $\mathrm{LSC}(X)$ denote the collection of all l.s.c. functions on X with values in $\bar{\mathbb{R}} = [-\infty, \infty]$ which are not identically equal to ∞. We shall next introduce a suitable topological structure in LSC(x) which will allow us to apply Theorem 2.2 to several important classes of l.s.c. functions.

Definition 3.1 (see [22]). *The epi-distance topology in LSC(X) is defined by the uniform structure formed by the sets*

$$U(\varepsilon, K) = \{(f, g) : f, g \in \mathrm{LSC}(X) \text{ and } |\operatorname{dist}((\alpha, x), \operatorname{epi} f) - \operatorname{dist}((\alpha, x), \operatorname{epi} g)| < \varepsilon, \text{ if } |\alpha| \leq K, \ \|x\| \leq K\}.$$

Here we set $\operatorname{dist}((\alpha, x), (\beta, u)) = |\alpha - \beta| + \rho(x, u)$.

Clearly, the uniform space $\mathrm{LSC}(X)$ is metrizable by a complete metric $d(\cdot, \cdot)$. (See e.g., [22]—although the proof is given there only for the case when X is a Banach space, the proof for an arbitrary metric space is not much different.) The following lemma contains crucial information about the behavior of functions close in $\mathrm{LSC}(X)$.

Lemma 3.2. *Let $\bar{f} \in \mathrm{LSC}(X)$ be bounded below. Suppose $\delta \in (0, 1)$ and a $\lambda \in \mathbb{R}$ are such that $\lambda > \inf \bar{f} + \delta$. Set*

$$\mathcal{L}_\lambda = \{x : \bar{f}(x) \leq \lambda\}. \tag{3.1}$$

Then for any $\gamma > 0$, $R > 0$ there is an $\varepsilon > 0$ such that for any other $f \in \mathrm{LSC}(X)$ with $d(f, \bar{f}) \leq \varepsilon$ we have $\inf f \leq \lambda - \delta/2$ and for any $z \in X$ satisfying $f(z) \leq \inf f + \varepsilon$ either $\inf \bar{f} \leq f(z) + \gamma$ and $\rho(z, \mathcal{L}_\lambda) \leq \gamma$ or $\|z\| \geq R$.

Proof. Taking if necessary a larger R, we may assume that $R > |\lambda| + |\inf \bar{f}| + 1$. Set $\sigma = \min\{\delta/4, \gamma/2\}$. Take an x with $\bar{f}(x) < \inf \bar{f} + \sigma$ and choose $\epsilon \in (0, \sigma)$ so small that for f with $d(f, \bar{f}) < \epsilon$ we have

$$|\operatorname{dist}((\alpha, z), \operatorname{epi} f) - \operatorname{dist}((\alpha, z), \operatorname{epi} \bar{f})| < \sigma, \tag{3.2}$$

if $|\alpha| \le R + \|x\|, \|z\| \le R + \|x\|$.

Clearly, $x \in \mathcal{L}_\lambda$ and for any f within ε of $\bar{f}$ we have by (3.2) for $\alpha = \inf \bar{f} + \sigma$

$$\operatorname{dist}((\alpha, x), \operatorname{epi} f) < \sigma,$$

which means that

$$f(x') < \alpha + \sigma = \inf \bar{f} + 2\sigma$$

with some x'. It follows that $\inf f < \lambda - \delta/2$.

Now let z be such that $f(z) \le \inf f + \varepsilon$. If further $||z|| \le R$, then by (3.2)

$$\operatorname{dist}((\max\{f(z), \inf \bar{f} - 1\}, z), \operatorname{epi} \bar{f}) < \sigma.$$

It follows that

$$\bar{f}(u) \le \max\{f(z), \inf \bar{f} - 1\} + \sigma$$

for some u such that $\rho(u, z) < \sigma$. Thus (as $\bar{f}(u) > \inf \bar{f} - 1 + \sigma$)

$$\bar{f}(u) \le f(z) + \sigma < \inf f + \varepsilon + \sigma \le \inf f + \delta/2 < \lambda.$$

This inequality gives $u \in \mathcal{L}_\lambda$; hence $\rho(z, \mathcal{L}_\lambda) \le \rho(u, z) < \sigma \le \gamma$ and $\inf \bar{f} \le \bar{f}(u) \le \inf f + \gamma$ which completes the proof.

4. Generic well-posedness of minimization for three classes of l.s.c. functions. We shall first apply our main theorem to the simplest case when elements of the data space are the functions themselves. Namely, we are ready now to consider the minimization problem for three important classes of l.s.c. functions, namely, convex l.s.c. functions, quasi-convex l.s.c functions, and arbitrary l.s.c functions satisfying a certain uniform growth condition. To this end we first fix some complete metric $d(f, g)$ in $\mathrm{LSC}(X)$ which is compatible with the epi-distance topology.

THEOREM 4.1. *In each of the following three cases the minimization problem for a generic $f \in \mathcal{A}$ is well-posed with respect to the space $\mathcal{A}$:*

(a) *X is a Banach space and $\mathcal{A}$ is the set of all convex l.s.c. functions on X;*

(b) *X is a Banach space and $\mathcal{A}$ is the set of all quasi-convex l.s.c. functions on X;*

(c) *X is a complete metric space and $\mathcal{A}$ is a collection of elements of* $\mathrm{LSC}(X)$ *satisfying $f(x) \ge \varphi(x)$ for all x, where $\varphi \in \mathrm{LSC}(X)$ satisfies $\varphi(x) \to \infty$ as $\|x\| \to \infty$.*

Proof. It is an easy matter to verify in either case that $\mathcal{A}$ is a closed subset of $\mathrm{LSC}(X)$; hence a complete metric space itself.

Let $\tilde{\mathcal{A}}$ denote the collection of those elements of $\mathcal{A}$ which are bounded below. We first observe that $\tilde{\mathcal{A}}$ is dense in $\mathcal{A}$. Indeed, for each $f \in \mathcal{A}$ the functions $f_n(x) = \max\{f(x), -n\}$, $x \in X$, $n = 1, 2, \dots$ belong to $\tilde{\mathcal{A}}$ and converge to f in the epi-distance topology. Let $\mathcal{B}$ denote the collection of those elements of $\tilde{\mathcal{A}}$ which attains its minimum. It can easily be seen that $\mathcal{B}$ is dense in $\mathcal{A}$. Indeed, for each $f \in \tilde{\mathcal{A}}$ the functions $f_n(x) = \max\{f(x), f(x_n)\}$, $x \in X$, $n = 1, 2, \dots$ with $x_n \in X$ satisfying $f(x_n) \le \inf f + n^{-1}$ belong to $\mathcal{B}$ and converge to f. Therefore the theorem will be proved if we verify that (H) holds with the $\mathcal{B}$ defined above.

Thus let $f \in \mathcal{B}$, $\varepsilon > 0$, $\gamma > 0$. Let $\bar{x} \in X$, $f(\bar{x}) = \inf f$. Choose a small $\delta \in (0, 1)$ and set

$$\bar{f}(x) = \sup\{f(x), f(\bar{x}) + \delta\rho(x, \bar{x})\}.$$

We can choose δ so small that $d(f, \bar{f}) < \varepsilon/2$.

Next we take a sufficiently big R—its value will be specified in each of the three cases separately. Set $\lambda = f(\bar{x}) + \delta\gamma/2$ and let, as above, $\mathcal{L}_\lambda = \{x : \bar{f}(x) \le \lambda\}$. Then $\mathcal{L}_\lambda$ belongs to the closed ball of radius $\gamma/2$ around $\bar{x}$ (as $\bar{f}(x) \ge f(\bar{x}) + \delta\rho(x, \bar{x})$). By Lemma 3.2 we can find an $\bar{\varepsilon}$ such that for any $g \in \mathrm{LSC}(X)$ with $d(g, \bar{f}) \le \bar{\varepsilon}$ and any z satisfying $g(z) \le \inf g + \bar{\varepsilon}$ we have

$$\inf g \le \lambda - \delta\gamma/8 \tag{4.1}$$

and either

$$\rho(z, \mathcal{L}_\lambda) \le \gamma/2,\ g(z) \ge \inf \bar{f} - \gamma \tag{4.2}$$

or

$$\|z\| \ge R. \tag{4.3}$$

In the first case, when (4.2) holds, we have by (4.2) $\rho(\bar{x}, z) \le \gamma$ and $g(z) \ge f(\bar{x}) - \gamma$ as desired. (The inequality $g(z) \le f(\bar{x}) + \gamma$ follows from (4.1) and the definition of λ if we assume that $\delta < 2/3$ and $\bar{\varepsilon} < \gamma/4$ which, of course, is always possible.)

To complete the proof of the theorem we have to show that in each case R can be chosen to make (4.3) impossible. Then (H) holds with $U = \{g \in \mathrm{LSC}(X)$: $d(g, \bar{f}) < \bar{\varepsilon}\}$, $\eta = \bar{\varepsilon}$, $\alpha = f(\bar{x})$.

In case (c) this is obvious (as $\varphi(z) \le g(z) \le f(\bar{x}) + \gamma$). Therefore we concentrate on (a) and (b). To simplify the calculation we may assume that all small numbers involved are smaller than one and that g is so close to $\bar{f}$ that

$$|\operatorname{dist}((\beta, x), \operatorname{epi} \bar{f}) - \operatorname{dist}((\beta, x), \operatorname{epi} g)| < 1/2, \tag{4.4}$$

if $\delta\|x - \bar{x}\| \le 2$, $|\beta| \le |f(\bar{x})| + 2$.

It follows that $g(x) > f(\bar{x}) + 1$ whenever $\|x - \bar{x}\| = 2/\delta$. Indeed, assuming the contrary, we get from (4.4) for some x and $\beta = f(\bar{x}) + 1$ that $\operatorname{dist}((\beta, x), \operatorname{epi} \bar{f}) < 1/2$; that is, there is an x' with $\|x - x'\| < 1/2$ such that $\bar{f}(x') \le f(\bar{x}) + 3/2$. However,

$$\begin{aligned} \bar{f}(x') &\ge f(\bar{x}) + \delta\|x' - \bar{x}\| \\ &\ge f(\bar{x}) + \delta(\|x - \bar{x}\| - \|x - x'\|) \\ &\ge f(\bar{x}) + 2 - \delta/2 > f(\bar{x}) + 3/2 \end{aligned}$$

and we arrive at a contradiction. Thus $g(x) > f(\bar{x}) + 1$ for any x with $\|x - \bar{x}\| = 2/\delta$. Now take $R > \|\bar{x}\| + 2/\delta$ and assume that (4.3) holds. By the choice of z, we have $g(z) < f(\bar{x}) + 1$ and, as follows from (4.4) for $\beta = f(\bar{x}) = \bar{f}(\bar{x})$,

$$\operatorname{dist}((\beta, \bar{x}), \operatorname{epi} g) < 1/2,$$

which means that there is an x' with $\|x' - \bar{x}\| < 1/2$ such that $g(x') < f(\bar{x}) + 1/2$. As x' lies within the ball of radius $1/2$ around $\bar{x}$, the line segment joining z and x' contains a point x with $\|x - \bar{x}\| = 2/\delta$. For this point

$$g(x) > f(\bar{x}) + 1 \ge \max\{g(z), g(x')\},$$

which cannot be the case either in (a) or in (b). This completes the proof of the theorem.

Remark 4.2. The first statement of the theorem was proved in [15].

5. Perturbations by linear functions: A theorem of Asplund. It is well known that genericity is not hereditary. In other words, a property which is generic on a space may fail to be generic on a subspace. Therefore it is reasonable, in principle, to consider minimization problems corresponding to more restricted collections of data.

The following result was proved in [13]. *Let X be an Asplund space and X^* its dual; let $g(x^*)$ be a norm l.s.c. extended-real-valued function on X^*; and let $f(x)$ be the restriction to X of the Fenchel conjugate of g. Assume that* $\operatorname{int}\operatorname{dom} f \neq \emptyset$. *Then the problem of minimizing $g_x(x^*) = g(x^*) - \langle x^*, x\rangle$ on X^* is generically well-posed with respect to $x \in \operatorname{dom} f$.*

Here $g^*(x^{**}) = \sup_{x^*}(\langle x^{**}, x^*\rangle - g(x^*))$ is the Fenchel conjugate of g and for any extended-real-valued function f on X, $\operatorname{dom} f = \{x : |f(x)| < \infty\}$.

Recall that X is called an *Asplund* space if every continuous convex function on X is Fréchet differentiable on a dense G_δ-set. There are many known characterizations of Asplund spaces. One of them (which is crucial in the proof in [18] is that a space with a Fréchet differentiable bump function is Asplund) is that every continuous convex function on X is ε-differentiable on a dense set. (A convex function f is ε-differentiable at $\bar{x}$ if there is an $x^* \in X^*$ such that $f(x) \le f(\bar{x}) + \langle x^*, x - \bar{x}\rangle + \varepsilon\|x - \bar{x}\|$ for all x of a neighborhood of $\bar{x}$.)

We shall show that both the quoted well-posedness theorem and the characterization of Asplund spaces follow from the generic variational principle of Theorem 2.2.

THEOREM 5.1. *For any Banach space X the following three statements are equivalent:*

(a) *X is an Asplund space;*

(b) *for any $\varepsilon > 0$, any continuous convex function on X is ε-differentiable on a dense subset of X;*

(c) *for any norm l.s.c. function g on X^* such that the domain of the restriction of its Fenchel conjugate to X has nonempty interior, the problem of minimization of $g_x(x^*) = g(x^*) - \langle x^*, x\rangle$ is generically well-posed with respect to $x \in \operatorname{dom} g^* \bigcap X$.*

(We observe that under the assumptions, $\operatorname{dom} g^* \bigcap X$ belongs to the closure of its interior; hence it is a Baire space—see Remark 2.3.)

Proof. The implication (a)⇒(b) is trivial; (c)⇒(a) follows from Šmulyan's duality between differentiability and rotundity [23]. Under the conditions of (c), if we denote, as above, the restriction of g^* to X by f, this duality reduces to equivalence of the following two properties for any $\bar{x} \in X$ and any $\bar{x}^* \in X^*$ (see [13]):

(i) there is a nonnegative convex l.s.c. function $\xi(t)$ (extended-real-valued) on $[0, \infty)$ such that $\xi(t)/t \to 0$ as $t \to 0$ and

$$f(x) \le f(\bar{x}) + \langle \bar{x}^*, x - \bar{x}\rangle + \xi(\|x - \bar{x}\|)$$

for all x;

(ii) there exists a convex l.s.c. function $\delta(t)$ on $[0, \infty)$ equal to zero at zero and positive for positive t, such that

$$g(x^*) \ge g(\bar{x}^*) + \langle x^* - \bar{x}^*, \bar{x}\rangle + \delta(\|x^* - \bar{x}^*\|)$$

for all x^*.

We have, therefore, to show that (b)⇒(c). Thus assume that (b) holds and g and f are as in (c). This means that $f(x) = -\inf_{x^*} g_x(x^*)$. We claim that for any x belonging to $\operatorname{int}\operatorname{dom} f$ there are $\varepsilon > 0$ and $r > 0$ such that for any $u \in x + B(u, \varepsilon)$ there is a minimizing sequence for g_u which belongs to the ball of radius r (around the origin) in X^*. Indeed, being weak l.s.c., f is continuous at x, so we can choose ε

so small that the entire ball $B(x, 2\varepsilon)$ (of radius 2ε around x) belongs to $\operatorname{int}\operatorname{dom} f$ and f is bounded above on the ball. If the claim is not true, then for any n we can find $u_n \in B(x, \varepsilon)$ within ε and $\{x_n^*\}$ such that $g_{u_n}(x_n^*) \le \inf g_{u_n} + 1/n$ and $\|x_n^*\| \to \infty$. But then we can take an $h \in X$ with the norm less than ε such that $\langle x_n^*, h\rangle \to -\infty$ in which case $-f(u_n) \le g_{u_n+h}(x_n^*) \to -\infty$, which (as $u_n + h \in B(x, 2\varepsilon)$) contradicts to the fact that f is bounded above on the ball.

It follows that for any $x \in \operatorname{int}\operatorname{dom} f$ there is an $r > 0$ such that the Fenchel conjugate f^r of the restriction of g to the ball of radius r around the origin (that is, the function equal to $g(x^*)$ if $\|x^*\| \le r$ and $+\infty$ otherwise) coincides with f in a neighborhood of x. But f^r, being conjugate to a function with bounded domain and bounded below, is everywhere finite, hence continuous, and by (b) densely Fréchet differentiable.

Thus, for any $x \in \operatorname{dom} f$ and any $\varepsilon > 0$ there is an $\bar{x}$ such that $\|\bar{x} - x\| < \varepsilon$ and f is Fréchet differentiable at $\bar{x}$. Let $\bar{x}^*$ be the derivative of f at $\bar{x}$. Then the property (i) above holds and therefore the property (ii).

Now, given a $\gamma > 0$, we choose a $\sigma > 0$ so small that $\sigma < \gamma/2$ and $\delta(t) > 2\sigma t$ if $t > \gamma$ and define a neighborhood U as the open ball of radius σ around $\bar{x}$. For any $u \in U$ we have by (ii)

$$g_u(x^*) \ge g_u(\bar{x}^*) + \delta(\|x^* - \bar{x}^*\|) - \langle x^* - \bar{x}^*, u - \bar{x}\rangle,$$

so that for $u \in U$ we have $g_u(x^*) \ge g_u(\bar{x}^*) + \sigma\|x^* - \bar{x}^*\| \ge g_u(\bar{x}^*) + \sigma\gamma$ if $\|x^* - \bar{x}^*\| > \gamma$. Now taking $\eta < \sigma\gamma/2$, we verify (H).

Thus by Theorem 2.2 the problem of minimizing g_x is generically well-posed with respect to $x \in \operatorname{dom} f$, that is, (c) is true.

6. Minimization subject to a nonfunctional constraint. This section is devoted to minimization problems of the form

$$\text{(P)} \qquad \text{minimize} \quad f(x) \quad \text{subject to} \quad x \in A,$$

with $A \subset X$ closed and f continuous and bounded below. This problem of course reduces to minimization of the restriction $f|_A$ of f to A and therefore can be considered in the framework of unconstrained minimization of l.s.c. extended-real-valued functions. However it is more natural to consider the pair (f, A) as the given data, rather than $f|_A$, and in this context Theorem 4.1 does not allow one to make any conclusion about generic solvability and well-posedness of the problem (as perturbations of $f|_A$ in $\operatorname{LSC}(X)$ may not be represented as restrictions of continuous functions).

In this section we use the notation introduced in the beginning of section 3 with $\|x\|$ standing for the distance from x to a specified "zero" element. Denote by $S(X)$ the collection of all closed subsets of X. We shall consider two complete metrics in $S(X)$: the metric $h_1(A, B)$ compatible with the Hausdorff topology (which is a completely metrizable topology defined by the uniform structure with the base

$$W_1(\varepsilon) = \{(A, B) : \ \rho(x, B) \le \varepsilon,\ x \in A,\ \rho(y, A) \le \varepsilon,\ y \in B\})$$

and a weaker metric $h_2(A, B)$ compatible with the bounded Hausdorff (Attouch–Wets) topology (which is a completely metrizable topology defined by the uniform structure with the base (see, e.g., [22])

$$W_2(\varepsilon, R) = \{(A, B) : \ |\rho(x, A) - \rho(x, B)| < \varepsilon \quad \forall x \in X,\ \|x\| \le R\}.$$

Let $\mathrm{C}^+(X)$ further be the collection of all continuous real-valued functions on X bounded from below. As for sets, we shall consider two metric structures in $\mathrm{C}^+(X)$; the first (complete) metric $r_1(f,g)$ is compatible with the topology of uniform convergence of elements of $\mathrm{C}^+(X)$ on X and the second weaker metric $r_2(f,g)$ can be any metric compatible with the topology of uniform convergence of elements of $\mathrm{C}^+(X)$ on bounded subsets of X (that is, on balls $B(R)=\{x:\|x\|\le R\}$).

Finally, let $\varphi(x)$ be a function on X bounded below and such that $\varphi(x)\to\infty$ as $\|x\|\to\infty$. Set

$$\mathrm{C}(X,\varphi)=\{f(\cdot)\in\mathrm{C}^+(X):\ f(x)\ge\varphi(x)\ \forall x\in X\}.$$

It is a trivial matter to verify that $\mathrm{C}(X,\varphi)$ is a complete subspace of $(\mathrm{C}^+(X),r_2)$. (The latter itself may be incomplete.)

The data spaces to be considered consist of pairs $a=(f,A)$ and we write f_a for $f|_A$ (so that for $b=(g,B)$ we have $f_b=g|_B$). In the theorem below we consider two data spaces of this kind: $(\mathcal{A}_1,d_1(a,b))$, where $\mathcal{A}_1=\mathrm{C}^+(X)\times S(X)$, and $(\mathcal{A}_2,d_2(a,b))$, where $\mathcal{A}_2=\mathrm{C}(X,\varphi)\times S(X)$ and for $a=(f,A),\ b=(g,B)$ the ith distance $(i=1,2)$ $d_i(a,b)$ is defined by $d_i(a,b)=r_i(f,g)+h_i(A,B)$.

THEOREM 6.1. *The problem* (P) *is generically well-posed with respect to both* $\mathcal{A}_1$ *and* $\mathcal{A}_2$.

Proof. We have to verify that in either case (H) holds. Therefore let $a=(f,A),\ \gamma>0$, and a neighborhood V of a, either in $\mathcal{A}_1$ or in $\mathcal{A}_2$ be given. We may assume that in the first case

$$V=\{b=(g,B):\ |f(x)-g(x)|\le\varepsilon\ \ \forall x\in X,\ \rho(x,B)\le\varepsilon\ \ \forall x\in A,\ \rho(y,A)\le\varepsilon\ \ \forall y\in B\}$$

with some $\varepsilon>0$ and in the second case

$$V=\{b=(g,B):\ |f(x)-g(x)|\le\varepsilon,\ x\in B(K),\ (A,B)\in W_2(\varepsilon,K)\}$$

with some $\varepsilon,K>0$. We may assume that $\varepsilon,\gamma<1$.

We choose positive numbers $\delta_1,\ \varepsilon_1,\ \bar\varepsilon$, and an $\bar x\in X$ to make sure that

$$\begin{aligned}&\varepsilon_1,\delta_1<(1/8)\min\{\varepsilon,\gamma\};\\ &\rho(\bar x,A)<\varepsilon_1;\ f(\bar x)\le\inf\{f(x):\ \rho(x,A)<\varepsilon_1\}+\delta_1;\\ &\bar\varepsilon<(\varepsilon_1-\rho(\bar x,A))/4.\end{aligned}$$

Let $\lambda(x)$ be a continuous function on X assuming values between 1 and 0 and such that $\lambda(x)=1$ if $\rho(x,(A\cup\{\bar x\}))\le\bar\epsilon$ and $\lambda(x)=0$ if $\rho(x,A)\ge\varepsilon_1$. Set

$$\begin{aligned}&\bar A=A\bigcup\{\bar x\};\\ &\bar f(x)=(1-\lambda(x))f(x)+\lambda(x)\max\{f(x),f(\bar x)+2\delta_1\min\{1,\gamma^{-1}\rho(x,\bar x)\}\};\\ &\bar a=(\bar f,\bar A).\end{aligned}$$

As in the previous section, we see that

(6.1)

$$f_{\bar a}(\bar x)=f(\bar x)=\bar f(\bar x)=\min f_{\bar a};\ \bar f(x)\ge f(\bar x)+2\delta_1,\quad\text{if } \rho(x,\bar x)\ge\gamma \text{ and } \rho(x,\bar A)\le\bar\varepsilon.$$

It also follows from the definition of $\bar f$ and $\bar x$ that

(6.2)
$$|f(x)-\bar f(x)|\le 3\delta_1\quad\forall\, x\in X.$$

Clearly, there exist a neighborhood U_1 of $\bar{f}$ and a neighborhood U_2 of $\bar{A}$ such that $U_1 \times U_2 \subset V$. We will show that U_1 and U_2 can be chosen so small that (H) holds with

$$U = U_1 \times U_2, \quad x = \bar{x}; \quad \alpha = f(\bar{x}); \quad \eta = \delta_1.$$

Choose a positive number ε_2 such that

$$|\bar{f}(\bar{x}) - \bar{f}(u)| < \delta_1/4, \quad \text{if } \rho(u, \bar{x}) < \varepsilon_2.$$

We can choose U_1 and U_2 sufficiently small to make sure that

(a) for $g \in U_1$

$$|\bar{f}(u) - g(u)| < \delta_1/4$$

for all $u \in X$ in the first case and for all u within R of $\bar{x}$ in the second case, where $R > 4$ is big enough to make sure that $\varphi(x) > f(\bar{x}) + 8\delta_1$ if $\rho(x, \bar{x}) > R$. In the first case we set $R = \infty$,

(b) any $B \in U_2$ contains a point u with $\rho(u, \bar{x}) < \varepsilon_2$ and

(c) for each $B \in U_2$ and each $u \in B$ satisfying $\rho(u, \bar{x}) \le R$:

$$\rho(u, \bar{A}) \le 2^{-1}\bar{\varepsilon}. \tag{6.3}$$

Let $b = (g, B) \in U$. In either case there is $u \in B$ such that $\rho(u, \bar{x}) < \varepsilon_2$ at which

$$\inf f_b \le g(u) \le \bar{f}(u) + \delta_1/4 \le \bar{f}(\bar{x}) + \delta_1/2.$$

If now $z \in B$ is such that $g(z) \le \inf f_b + \delta_1$, then

$$g(z) \le \bar{f}(\bar{x}) + \delta_1/2 + \delta_1, \quad \rho(z, \bar{x}) \le R, \quad \bar{f}(z) \le g(z) + \delta_1/4 < \bar{f}(\bar{x}) + 2\delta_1.$$

As follows from (6.3) and (6.1)

$$\rho(z, \bar{A}) \le \bar{\varepsilon}/2, \ \rho(z, \bar{x}) \le \gamma.$$

Clearly, $|g(z) - f(\bar{x})| \le 2\delta_1 < \gamma$. The proof is completed.

Remark 6.2. For $\mathcal{A}_1$ the theorem was proved in [19].

7. Topology of uniform convergence modulo given growth. We shall now pass to the last application of the generic variational principle, the antirelaxation theorem. But before we have to consider in some details the space of integrands and certain special topologies in this space.

We consider here the collection $\mathrm{LSC}^+(X)$ of nonnegative l.s.c. proper functions on X with another topology (or, more precisely, a series of topologies). Namely, for any $p \ge 1$ we introduce a uniform structure in $\mathrm{LSC}^+(X)$ with the basis formed by the sets

$$U_p(\varepsilon) = \{(f, g) : \ |f(x) - g(x)| \le \varepsilon(\|x\|^p + 1) \quad \forall\, x \in X\}$$

with the convention that $\infty - \infty = 0$. (As above, $\|x\|$ is the distance from x to a "zero" element.) Verification that these sets form a basis of a uniform structure for any p is an easy matter. We shall denote $\mathrm{LSC}^+(X)$ with the corresponding topology by $\mathrm{LSC}_p{}^+(X)$.

Observe that $\mathrm{LSC}_p{}^+(X)$ is a "very" disconnected space. Indeed, fix an $f_0 \in \mathrm{LSC}_p{}^+(X)$ and let

$$\tilde{U} = \bigcup_{\varepsilon>0} \{g \in \mathrm{LSC}(X) :\ (g, f_0) \in U_p(\varepsilon)\}.$$

Clearly, $\tilde{U}$ is a open set. For any $f \notin \tilde{U}$ we have

$$\sup_x \frac{|f(x) - f_0(x)|}{\|x\|^p + 1} = \infty.$$

Therefore if $(g, f) \in U_p(\varepsilon)$ for some ε, the equality above will not change if we replace f by g which means that $g \notin \tilde{U}$. In other words, the complement of $\tilde{U}$ is also an open set. The above argument actually proves the following.

PROPOSITION 7.1. *Functions $f, g \in \mathrm{LSC}(X)$ belong to the same connected component of $\mathrm{LSC}_{\mathrm{p}}^+(X)$ if and only if $(g, f) \in U_p(\varepsilon)$ for some ε; in particular, in this case* $\mathrm{dom}\, f = \mathrm{dom}\, g$.

It is clear that $\mathrm{LSC}_p{}^+(X)$ is metrizable. Moreover it is complete as follows from the proposition below.

PROPOSITION 7.2. *The space $\mathrm{LSC}_p{}^+(X)$ is complete.*

Proof. Let $\{f_m\}$ be a fundamental sequence in $\mathrm{LSC}_p{}^+(X)$. This means that there is a sequence of positive numbers $\{\varepsilon_m\}$ converging to zero such that $(f_n, f_{n+k}) \in U_p(\varepsilon_m)$ for all $n \geq m$ and all k. Let $\mathcal{D}$ be the common domain of all f_m. Clearly, for every x the sequence of values $f_m(x)$ converges. Denote the pointwise limit function by $f(x)$. Its domain is also $\mathcal{D}$ and for any m the inequality

$$|f_m(x) - f(x)| \leq \varepsilon_m(\|x\|^p + 1) \tag{7.1}$$

holds.

It remains to verify that $f \in \mathrm{LSC}_p{}^+(X)$, that is to say, that f is proper and l.s.c. The first is obvious. The second is an immediate consequence of the inequality (7.1) as for any m any x and any sequence $\{x_n\}$ converging to x

$$\begin{aligned} \liminf_{n\to\infty} f(x_n) &\geq \liminf_{n\to\infty} f_m(x_n) - \varepsilon_m(\|x\|^p + 1) \\ &\geq f_m(x) - \varepsilon_m(\|x\|^p + 1) \geq f(x) - 2\varepsilon_m(\|x\|^p + 1). \end{aligned}$$

8. Spaces of integrands. Let (T, Ξ, μ) be a space with a positive measure (which can be infinite). By $\mathcal{I}_p^+(T, X)$ we denote the topological space whose elements are nonnegative extended-real-valued functions $f(t, x)$ on $T \times X$ with the following properties:

(P1) f is $\mathcal{L} \bigotimes \mathcal{B}$-measurable, that is to say, measurable with respect to the σ-algebra generated by products of elements of Ξ and Borel subsets of X;

(P2) $f(t, \cdot)$ is l.s.c. and proper for almost every $t \in T$;

(P3) there is an $x(\cdot) \in L_p(T, X)$ such that $f(t, x(t)) \in L_1$.

The topology in $\mathcal{I}_p^+(T, X)$ is generated by the uniform structure whose basis is formed by the sets

$$\begin{aligned} V_p(\varepsilon) = \{(f, g) :\ &\exists\ \varphi \in B(T, \Xi, \mu),\ \varphi(t) \geq 0 \text{ such that} \\ &|f(t, x) - g(t, x)| < \varepsilon(\|x\|^p + \varphi(t)) \quad \forall x \in X \text{ almost everywhere on } T\}, \end{aligned}$$

where $B(T, \Xi, \mu)$ is the unit ball in $L_1(T, \Xi, \mu)$.

To be more precise, we have to define elements of $\mathcal{I}_p^+(T,X)$ as classes of functions equivalent in the sense that $f \sim f'$ if and only if $f(t,\cdot) = f'(t,\cdot)$ for almost every t. Properties (P1) and (P2) of elements of $\mathcal{I}_p^+(T,X)$ mean that they are *normal integrands* [7].

In this definition, as above, we adopt the convention that $\infty - \infty = 0$; we also observe that φ in the definition may depend on f and g. Again, it is an easy matter to see that the sets $V_p(\varepsilon)$ form a basis of a metrizable uniform structure.

Every integrand $f \in \mathcal{I}_p^+(T,X)$ generates an integral functional

$$J_f(x(\cdot)) = \int_T f(t,x(t))dt$$

on $L_p(T,X)$.

PROPOSITION 8.1. *For any $f \in \mathcal{I}_p^+(T,X)$ the functional J_f is a proper l.s.c. function on $L_p(T,X)$ and $\operatorname{dom} J_f = \operatorname{dom} J_g$ whenever f and g belong to the same connected component of $\mathcal{I}_p^+(T,X)$.*

Proof. J_f is a proper function by (P3). Lower semicontinuity is an immediate consequence of Fatou's lemma and the equality of domains is verified as in the proof of Proposition 7.2.

PROPOSITION 8.2. *The space $\mathcal{I}_p^+(T,X)$ is complete.*

Proof. Let $\{f_m\}$ be a fundamental sequence in $\mathcal{I}_p^+(T,X)$. This means that there is a sequence $\{\varepsilon_m\}$ of positive numbers converging to zero and a (triple indexed) sequence $\{\varphi_{mnk}\}$ of elements of the unit ball of $L_1(T,\Xi,\mu)$ such that almost everywhere in T

$$|f_n(t,x) - f_k(t,x)| \le \varepsilon_m(\|x\|^p + \varphi_{mnk}(t)) \quad \forall x, \tag{8.1}$$

whenever $n \ge m,\ k \ge m$. The strategy of the proof this time will be the following: We shall show that any subsequence of $\{f_m\}$ contains a converging sub-subsequence and limits of any two converging subsequences of $\{f_m\}$ may differ only on a set whose projection to T has μ-measure zero.

Therefore consider an arbitrary subsequence $\{f_{m_s}\}$. Taking a further subsequence if necessary (and writing for convenience g_s instead of f_{m_s} and δ_s instead of ε_{m_s}), we may assume that the series $\sum \delta_s$ converges. Then the functions $\psi_s(t) = \sum_{r=s}^{\infty} \delta_r \varphi_{m_r m_r m_{r+1}}$ decreasingly converge to zero in L_1. As these functions are nonnegative, it follows from the theorem of B. Levi (see [24, p. 75]) that $\psi_s(t) \to 0$ on a set $T' \in \Xi$ whose complement has measure zero.

We have, setting $\gamma_s = \sum_{r=s}^{\infty} \delta_r$,

$$|g_s(t,x) - g_{s+k}(t,x)| \le \gamma_s \|x\|^p + \psi_s(t).$$

It follows that for every $t \in T'$ the sequence $\{g_s(t,\cdot)\}$ is fundamental in $\mathrm{LSC}_p{}^+(X)$. Therefore by Proposition 7.2 the limit function $g(t,\cdot)$ is well defined on $T' \times X$ and l.s.c. in x. Being a pointwise limit of $\mathcal{L} \bigotimes \mathcal{B}$-measurable nonnegative functions, it is also $\mathcal{L} \bigotimes \mathcal{B}$-measurable and nonnegative. We have furthermore

$$|g_s(t,x) - g(t,x)| \le \gamma_s \|x\|^p + \psi_s(t)$$

from which we conclude that g has property (P3); hence $g \in \mathcal{I}_p^+(T,X)$, and g_s converges to g in $\mathcal{I}_p^+(T,X)$.

Now let g and g' be limits of two subsequences of $\{f_m\}$. It follows from (8.1) that $J_g \equiv J_{g'}$ on $L_p(T,X)$. If we had assumed now that $g(t,x) \ne g'(t,x)$ on a set whose projection to T has positive measure, then we would conclude using the standard measurable selection arguments (e.g., [7]) that $J_g(x(\cdot)) \ne J_{g'}(x(\cdot))$ for some $x(\cdot) \in L_p(T,X)$. This completes the proof of the proposition.

9. Generalized Bolza problem: An antirelaxation theorem. We are ready now to state and prove the concluding result quoted at the very beginning of the introduction. Let Ω be a bounded domain in $\mathbb{R}^m$ with Lipschitz boundary. We shall consider the space

$$\mathcal{A}_p = \mathcal{I}_p^+(\partial\Omega, \mathbb{R}^n) \times \mathcal{I}_p^+(\Omega, \mathbb{R}^n \times \mathbb{R}^{mn})$$

with the product topology. (Here Ω is considered with the Lebesgue measure and $\partial\Omega$ with the $(m-1)$-dimensional Lebesgue measure in $\mathbb{R}^m$. As Ω is a bounded domain with Lipschitz boundary, the $(m-1)$-dimensional Lebesgue measure of $\partial\Omega$ is finite.) With every $a = (l, L) \in \mathcal{A}_p$ we associate the generalized Bolza problem of minimizing

$$J_a(x(\cdot)) = \int_{\partial\Omega} l(\xi, x(\xi))d\xi + \int_\Omega L(t, x(t), \nabla x(t))dt$$

over all $x(\cdot) \in W_{1,p}^n(\Omega)$, the latter being the standard Sobolev space. For convenience we shall interpret it as the closure with respect to the norm

$$\|x(\cdot)\|_{1,p} = \left(\int_{\partial\Omega} |x(\xi)|^p d\xi + \int_\Omega (|x(t)|^p + |\nabla x(t)|^p)dt\right)^{1/p}$$

of the collection of continuous $\mathbb{R}^n$-valued functions on $\Omega \bigcup \partial\Omega$ which are continuously differentiable on Ω. (Here we denote by $|\cdot|$ the Euclidean norms of vectors in $\mathbb{R}^m$ and operators $\mathbb{R}^n \mapsto \mathbb{R}^m$ (see, e.g., Theorem 3.4.5 in [5]).)

As in the preceding sections we verify that, given a component of $\mathcal{A}_p$, J_a is either identically equal to ∞ or is proper with the same domain for all a of the component. In the latter case we shall say, abusing the language slightly, that J_a is *proper on the corresponding component*.

THEOREM 9.1. *The generalized Bolza problem of minimizing J_a on $W_{1,p}^n$ is generically well-posed with respect to every component of $\mathcal{A}_p$ on which J_a is proper.*

Proof. Being a product of two complete metrizable uniform spaces, $\mathcal{A}_p$ is also a complete metrizable uniform space. As a basis for the uniform structure in $\mathcal{A}_p$ we can take the sets $U(\delta)$ of pairs (a, a') having the property that there are functions $\varphi(\cdot) \in L_1(\partial\Omega)$ and $\psi(\cdot) \in L_1(\Omega)$ with norms not greater than one such that for almost every t in $\partial\Omega$ and Ω, respectively, the inequalities below are valid:

$$|l(t,x) - l'(t,x)| < \delta(|x|^p + \varphi(t)) \ \forall x \in \mathbb{R}^n,$$
$$|L(t,x,w) - L'(t,x,w)| < \delta(|x|^p + |w|^p + \psi(t)) \ \forall x \in \mathbb{R}^n, \ \forall w \in \mathbb{R}^{mn}.$$

Fix a metric $d(a, b)$ in $\mathcal{A}_p$ compatible with the structure. We have to show that (H) is satisfied for $(\mathcal{A}_p, d)$, J_a, and $X = W_{1,p}^n$. As the set $\mathcal{B}$ of those $a \in \mathcal{A}_p$ for which $J_a(x(\cdot)) \to \infty$ as $\|x(\cdot)\|_{1,p} \to \infty$ is dense in $\mathcal{A}_p$, we need to verify properties (i), (ii) of (H) only for $a \in \mathcal{B}$.

Therefore let an $\varepsilon > 0$, a $\gamma > 0$, and an $a \in \mathcal{B}$ be given. Let K be such that $\|x(\cdot)\|_{1,p}^p \le K$ if $J_a(x(\cdot)) \le \inf J_a + 1$. Let further $1 > \delta_1 > 2^{2(p-1)}\delta_2 > 0$ be such that

$$\gamma > \frac{2\delta_2(2^p K + 3)}{2^{(1-p)}\delta_1 - 2^{(p-1)}\delta_2} \quad \text{and} \quad d(a,b) < \varepsilon/2, \text{ if } (a,b) \in U(\delta_1(K+1)). \tag{9.1}$$

If a belongs to a component of $\mathcal{A}_p$ whose elements generate proper functionals on $W_{1,p}^n$, then $\inf J_a < \infty$. Choose an $x(\cdot) \in X$ such that

$$J_a(x(\cdot)) \le \inf J_a + \delta_2 \tag{9.2}$$

and set

$$\alpha = J_a(x(\cdot)).$$

Set further

$$\begin{aligned}&\bar{l}(t,x) = l(t,x) + 2^{(1-p)}\delta_1|x - x(t)|^p;\\ &\bar{L}(t,x,w) = L(t,x,w) + 2^{(1-p)}\delta_1(|x - x(t)|^p + |w - \nabla x(t)|^p);\\ &\bar{a} = (\bar{l},\bar{L}).\end{aligned}$$

Then for any $y(\cdot) \in W^n_{1,1}$

$$J_{\bar{a}}(y(\cdot)) = J_a(y(\cdot)) + 2^{1-p}\delta_1\|y(\cdot) - x(\cdot)\|^p_{1,p} \tag{9.3}$$

and (as $\|x-y\|^p \le 2^{(p-1)}(\|x\|^p + \|y\|^p)$)

$$\begin{aligned}0 \le \bar{L}(t,x,w) - L(t,x,w) &\le \delta_1(|x|^p + |w|^p + |x(t)|^p + |\nabla x(t)|^p)\\ &\le \delta_1(K+1)(|x|^p + |w|^p + (K+1)^{-1}(|x(t)|^p + |\nabla x(t)|^p));\\ 0 \le \bar{l}(t,x) - l(t,x) &\le \delta_1(|x|^p + |x(t)|^p)\\ &\le \delta_1(K+1)(|x|^p + (K+1)^{-1}|x(t)|^p).\end{aligned}$$

The latter, in view of the choice of $x(\cdot)$ and δ_1, implies that $d(\bar{a},a) < \varepsilon/2$.

Now let $U = \{b \in \mathcal{A} : \ (\bar{a},b) \in U(\delta_2)\}$. Then U is an open set and by (9.1), $d(a,b) < \varepsilon$ for any $b \in U$. We have furthermore for any such b

$$|J_{\bar{a}}(y(\cdot)) - J_b(y(\cdot))| \le \delta_2(\|y(\cdot)\|^p + 2) \tag{9.4}$$

(which is immediate from the definition of $U(\delta)$).

Finally, choose a $z(\cdot)$ such that

$$J_b(z(\cdot)) \le \inf J_b + \delta_2.$$

Then by (9.4) (as $J_a(x(\cdot)) = J_{\bar{a}}(x(\cdot))$)

$$\begin{aligned}J_b(z(\cdot)) &\le \inf J_b + \delta_2 \le J_b(x(\cdot)) + \delta_2\\ &\le J_a(x(\cdot)) + \delta_2(K+3).\end{aligned} \tag{9.5}$$

On the other hand, by (9.3), (9.4) we get

$$\begin{aligned}J_b(z(\cdot)) &\ge J_{\bar{a}}(z(\cdot)) - \delta_2(\|z(\cdot)\|^p + 2)\\ &\ge J_a(z(\cdot)) + 2^{(1-p)}\delta_1\|z(\cdot) - x(\cdot)\|^p\\ &\qquad -2^{(p-1)}\delta_2(\|z(\cdot) - x(\cdot)\|^p + \|x(\cdot)\|^p + 2)\\ &\ge J_a(z(\cdot)) + (2^{(1-p)}\delta_1 - 2^{(p-1)}\delta_2)\|z(\cdot) - x(\cdot)\|^p - \delta_2(2^{(p-1)}K + 2).\end{aligned} \tag{9.6}$$

Comparing (9.5) and (9.6), we get

$$2\delta_2(2^{p-1}K + 3) \ge (2^{1-p}\delta_1 - 2^{p-1}\delta_2)\|z(\cdot) - x(\cdot)\|^p$$

from which we conclude by (9.1) that $\|z(\cdot) - x(\cdot)\|^p \le \gamma$, and also

$$|J_b(z(\cdot) - \alpha| \le \delta_2(2^{p-1}K + 3) < \gamma.$$

Thus, to conclude the proof, we need to take only $\eta \le \delta_2$.

The last result of the paper relates to the standard boundary value problem of calculus of variations:

$$\text{(P)} \qquad \text{minimize} \quad \int_\Omega L(t,x(t),\nabla x(t))dt, \qquad x(\cdot)\in W^n_{1,p}, \quad x|_{\partial\Omega}(\cdot)=0.$$

Let $\mathcal{I}_p(\Omega,\mathbb{R}^n\times\mathbb{R}^{mn})$ stand for the subspace of $\mathcal{I}^+_p(\Omega,\mathbb{R}^n\times\mathbb{R}^{mn})$ consisting of integrands $L(t,x,w)$ for which the infimum in (P) is finite. By Proposition 8.1 if $L\in\mathcal{I}_p(\Omega,\mathbb{R}^n\times\mathbb{R}^{mn})$, then the entire component of $\mathcal{I}^+_p(\Omega,\mathbb{R}^n\times\mathbb{R}^{mn})$ containing L belongs to $\mathcal{I}_p(\Omega,\mathbb{R}^n\times\mathbb{R}^{mn})$. Therefore, $\mathcal{I}_p(\Omega,\mathbb{R}^n\times\mathbb{R}^{mn})$ is a union of components of $\mathcal{I}^+_p(\Omega,\mathbb{R}^n\times\mathbb{R}^{mn})$; hence an open subspace and therefore a completely metrizable uniform space.

THEOREM 9.2. *The standard boundary value problem* (P) *of calculus of variations is generically well-posed with respect to* $\mathcal{I}_p(\Omega,\mathbb{R}^n\times\mathbb{R}^{mn})$.

Proof. Set

$$l_0(t,x)=\begin{cases}0 & \text{if } x=0;\\ \infty & \text{otherwise.}\end{cases}$$

Consider the family $\{\mathcal{C}_\alpha\}$ of all subsets of $\mathcal{A}$ such that every $\mathcal{C}_\alpha$ is the product of the component of $\mathcal{I}^+_p(\partial\Omega,\mathbb{R}^n)$ containing l_0 and a component of $\mathcal{I}_p(\Omega,\mathbb{R}^n\times\mathbb{R}^{mn})$. Then every $\mathcal{C}_\alpha$ is a component of $\mathcal{A}$, J_a is proper on every $\mathcal{C}_\alpha$, and the problem of minimizing J_a is generically well-posed on $\mathcal{C}_\alpha$ by Theorem 9.1.

Set $\mathcal{A}_0=\bigcup\mathcal{C}_\alpha$, and let for any α, $\mathcal{C}'_\alpha$ be a dense G_δ-subset of $\mathcal{C}_\alpha$ such that for any $a\in\mathcal{C}'_\alpha$ the problem of minimizing J_a is well-posed with respect to $\mathcal{C}_\alpha$. For any α, we have $\mathcal{C}'_\alpha=\bigcap_{i=1}^\infty U_{i,\alpha}$, where every $U_{i,\alpha}$ is an open set contained in $\mathcal{C}_\alpha$. Clearly, $\mathcal{A}'_0=\bigcup\mathcal{C}'_\alpha$ is a dense subset of $\mathcal{A}_0$ and on the other hand, as $U_{i,\alpha}\bigcap U_{j,\beta}=\emptyset$ if $\alpha\neq\beta$,

$$\mathcal{A}'_0=\bigcup_\alpha\left(\bigcap_i U_{i,\alpha}\right)=\bigcap_i\left(\bigcup_\alpha U_{i,\alpha}\right),$$

so that $\mathcal{A}'_0$ is also a G_δ subset of $\mathcal{A}$.

Thus, the problem of minimization of J_a on $W^n_{1,1}$ is by Theorem 9.1 generically well-posed with respect to $\mathcal{A}_0$. The theorem now follows from the observation that if l belongs to the component of $\mathcal{I}^+_p(\partial\Omega,\mathbb{R}^n)$ containing l_0, then $l(t,0)$ differs from $l_0(t,0)$ by a function $\varphi(t)$ which is summable on $\partial\Omega$ and $l(t,x)=\infty$ if $x\neq 0$. It follows that $\operatorname{dom}J_a\subset W^{1,1}_0$ for any $a\in\mathcal{A}_0$, and for any $x(\cdot)\in W^{1,1}_0$

$$J_a(x(\cdot))-\int_\Omega L(t,x(t),\nabla x(t))dt=\int_{\partial\Omega}\varphi(t)dt,$$

that is to say, the functional of (P) has the same domain as J_a and differs from the restriction of J_a to $\operatorname{dom}J_a$ by a constant.

REFERENCES

[1] L. TONELLI, *Fondamenti di Calcolo delle Variazioni*, Zanichelli, Bologna, Italy, 1921–1923.

[2] N. NAGUMO, *Uber die gleichmässige Summierbarkeit und ihre Anwendung auf ein Variationsproblem*, Japan J. Math., 6 (1929), pp. 173–182.

[3] E. J. MCSHANE, *Existence theorem for the ordinary problem of the calculus of variations*, Ann. Scoula Norm. Sup. Pisa Cl. Sci. (4), 3 (1934), pp. 181–211.

[4] S. CINQUINI, *Sopra l'esistenza della soluzione nei problemi di Calcolo delle variazioni*, Ann. Scuola Norm. Sup. Pisa Cl. Sci. (4), 5 (1936), pp. 675–682.

[5] Ch. Morrey, *Multiple Integrals in the Calculus of Variations*, Springer-Verlag, New York, 1966.
[6] L. Cesari, *Optimization—Theory and Applications*, Springer-Verlag, New York, 1983
[7] A. D. Ioffe and V. M. Tihomirov, *Theory of Extremal Problems*, Nauka, Moscow, 1974; English translation, North–Holland, Amsterdam, 1979.
[8] F. H. Clarke, *Optimization and Non-smooth Analysis*, Wiley Interscience, New York, 1983.
[9] L. W. Neustadt, *The existence of optimal controls in the absence of convexity conditions*, J. Math. Anal. Appl., 7 (1963), pp. 110–117.
[10] B. Botteron and P. Marcellini, *A general approach to the existence of minimizers of one-dimensional noncoercive integrals of the calculus of variations*, Ann. Inst. H. Poincaré Anal. Non Linéaire, 8 (1991), pp. 197–223.
[11] B. S. Mordukhovich, *Existence theorems in nonconvex optimal control*, in Calculus of Variations and Related Topics, A. Ioffe, S. Reich, and I. Shafrir, eds., CRC Press, Boca Raton, FL, 1999, pp. 175–197.
[12] A. J. Zaslavski, *Existence of Solutions of Optimal Control Problems without Convexity Assumptions*, preprint, 1996.
[13] E. Asplund, *Fréchet differentiability of convex functions*, Acta. Math., 121 (1968), pp. 31–47.
[14] A. D. Ioffe and V. M. Tihomirov, *Several remarks on variational principles*, Matem. Zametki, 61, (1997), pp. 305–311.
[15] G. Beer and R. Lucchetti, *Convex optimization and the epi-distance topology*, Trans., Amer. Math. Soc., 327 (1991), pp. 795–813.
[16] A. L. Dontchev and T. Zolezzi, *Well-Posed Optimization Problems*, Lecture Notes in Math. 1543, Springer-Verlag, Berlin, 1993.
[17] R. Deville, G. Godefroy, and V. Zizler, *Smoothness and Renormings in Banach Spaces*, Longman Scientific and Technical, Harlow, 1993.
[18] I. Ekeland and G. Lebourg, *Generic Fréchet-differentiability and perturbed optimization problems in Banach spaces*, Trans. Amer. Math. Soc., 224 (1976), pp. 193–216.
[19] J. Revalski, *Generic properties concerning well-posed optimization problems*, C.R. Acad. Bulgare Sci., 38 (1985), pp. 1431–1434.
[20] P. S. Kenderov and J. P. Revalski, *Generic well-posedness of optimization problems and the Banach-Mazur game*, in Recent Developments in Well-Posed Variational Problems, R. Lucchetti and J. Revalski, eds., Kluwer, Dordrecht, the Netherlands, 1995, pp. 117–136.
[21] T. Zolezzi, *Well-posedness criteria in optimization with application to the calculus of variations*, Nonlinear Anal., 25 (1995), pp. 437–453.
[22] H. Attouch, R. Lucchetti, and R. J. B. Wets, *The topology of the ρ-Hausdorff distance*, Ann. Mat. Pura Appl., Ser. 4, 160 (1992), pp. 303–320.
[23] V. L. Šmulyan, *Sur la dérivabilité de la norme dans l'espace de Banach*, Dokl. Akad. Nauk, 27 (1940), pp. 643–648.
[24] J. L. Doob, *Measure Theory*, Springer-Verlag, New York, 1994.

SIAM J. CONTROL OPTIM.
Vol. 38, No. 2, pp. 582–612

STOCHASTIC CALCULUS FOR FRACTIONAL BROWNIAN MOTION I. THEORY*

TYRONE E. DUNCAN†, YAOZHONG HU†, AND BOZENNA PASIK-DUNCAN†

Abstract. In this paper a stochastic calculus is given for the fractional Brownian motions that have the Hurst parameter in (1/2, 1). A stochastic integral of Itô type is defined for a family of integrands so that the integral has zero mean and an explicit expression for the second moment. This integral uses the Wick product and a derivative in the path space. Some Itô formulae (or change of variables formulae) are given for smooth functions of a fractional Brownian motion or some processes related to a fractional Brownian motion. A stochastic integral of Stratonovich type is defined and the two types of stochastic integrals are explicitly related. A square integrable functional of a fractional Brownian motion is expressed as an infinite series of orthogonal multiple integrals.

Key words. fractional Brownian motion, stochastic calculus, Itô integral, Stratonovich integral, Itô formula, Wick product, Itô calculus, multiple Itô integrals, multiple Stratonovich integrals

AMS subject classifications. Primary, 60H05; Secondary, 60H30, 60G15, 60G18

PII. S036301299834171X

1. Introduction.

Since the pioneering work of Hurst [14], [15] and Mandelbrot [18], the fractional Brownian motions have played an increasingly important role in many fields of application such as hydrology, economics, and telecommunications.

Let $0 < H < 1$. It is well known that there is a Gaussian stochastic process $(B_t^H, t \geq 0)$ such that

$$\mathbb{E}\left(B_t^H\right) = 0, \qquad \mathbb{E}\left(B_t^H B_s^H\right) = \frac{1}{2}\left\{|t|^{2H} + |s|^{2H} - |t-s|^{2H}\right\} \tag{1.1}$$

for all $s, t \in \mathbb{R}_+$. This process is called a (standard) fractional Brownian motion with Hurst parameter H.

If $H = 1/2$, then the corresponding fractional Brownian motion is the usual standard Brownian motion. If $H > 1/2$, then the process $(B_t^H, t \geq 0)$ exhibits a long-range dependence, that is, if $r(n) = \mathbb{E}[B_1^H(B_{n+1}^H - B_n^H)]$, then $\sum_{n=1}^{\infty} r(n) = \infty$. A fractional Brownian motion is also self-similar, that is, $(B_{\alpha t}^H, t \geq 0)$ has the same probability law as $(\alpha^H B_t^H, t \geq 0)$. A process satisfying this property is called a self-similar process with the Hurst parameter H.

Since in many problems related to network traffic analysis, mathematical finance, and many other fields the processes under study seem empirically to exhibit the self-similar properties, and the long-range dependent properties, and since the fractional Brownian motions are the simplest processes of this kind, it is important to have a systematic study of these processes and to use them to construct other stochastic processes. One way to approach this study is to follow, by analogy, the methods for Brownian motion. In the stochastic analysis, a Brownian motion can be used as the input (white) noise and many other processes (e.g., general diffusion processes) can

*Received by the editors July 9, 1998; accepted for publication (in revised form) March 3, 1999; published electronically February 9, 2000. This research was partially supported by NSF grant DMS 9623439.

http://www.siam.org/journals/sicon/38-2/34171.html

†Department of Mathematics, University of Kansas, 405 Snow Hall, Lawrence, KS 66045-2142 (duncan@math.ukans.edu, bozenna@math.ukans.edu).

be constructed as solutions of stochastic differential equations. One powerful tool for determining these solutions is the Itô formula.

However, it is also known that if a stochastic process $(\pi_t, t \geq 0)$ has the property that the stochastic integral $\int F_t \, d\pi_t$ is well defined for a large class of integrands $(F_t, t \geq 0)$, then this process $(\pi_t, t \geq 0)$ is a semimartingale, e.g., [20]. It is known that the fractional Brownian motions are not semimartingales. Therefore the beautiful classical theory of stochastic analysis [4] is not applicable to fractional Brownian motions for $H \neq 1/2$. It is a significant and challenging problem to extend the results in the classical stochastic analysis to these fractional Brownian motions. There have been a few papers in this direction. Lin [17] and Dai and Heyde [2] introduced stochastic integrals and extended the Itô formula to fractional Brownian motions. Their definitions of a stochastic integral give a stochastic integral of Stratonovich type, which is explained further in section 3. Their Itô formula is the usual chain rule for differentiation.

The stochastic integral $\int_0^t f_s \delta B_s^H$, with respect to the fractional Brownian motions introduced by Lin, Dai, and Heyde, does not satisfy in general the following property: $\mathbb{E} \int_0^t f_s \delta B_s^H = 0$. A new type of stochastic integral, $\int_0^t f_s \, dB_s^H$, is introduced, satisfying $\mathbb{E} \int_0^t f_s \, dB_s^H = 0$. This property seems to be important in the modeling problem by stochastic differential equations with fractional Gaussian noise as the driving random process. Consider the following type of differential equation:

$$dX_t = b(X_t)dt + \sigma(X_t)dB_t^H. \tag{1.2}$$

It is natural to consider that $b(X_t)$ is the mean rate of change of the system state X_t at time t and $\sigma(X_t)dB_t^H$ is the random perturbation. So the term $\sigma(X_t)dB_t^H$ should not contribute to the mean rate of change. The term $b(X_t)$ is used to represent the average or deterministic part of the problem and $\sigma(X_t)$ is used to represent the intensity of the random part of the problem. Therefore it is important to extend the classical interpretation of b and σ to the differential equation (1.2).

To introduce the new integral $\int f \, dB^H$, the Wick product or Wick calculus is used. The use of the Wick product is not anomalous because in white noise analysis the usual product has been associated with integrals of Stratonovich type and the Wick product has been associated with integrals of Itô type (e.g., [8], [12]).

A brief outline of the paper is as follows. In section 2, some description and terminology for the fractional Brownian motions are given. In section 3, a derivative in special directions is defined and a stochastic integral of Itô type is defined using the Wick product. Furthermore, a stochastic integral of Stratonovich type is defined and the two types of stochastic integrals are related. In section 4, some change of variables formulas (Itô formulas) are given for the two types of stochastic integrals. In section 5, two applications of the Itô formula are given. In section 6, multiple integrals of Itô type and Stratonovich type for a fractional Brownian motion are defined and the Hu-Meyer formula is extended to these multiple integrals. The relation between these two types of multiple integrals is given. A square integrable functional of a fractional Brownian motion is represented as an infinite sum of orthogonal terms.

If the integrand of a stochastic integral of Stratonovich type for a fractional Brownian motion with $H \in (1/2, 1)$ is smooth, then the integral can be defined as a limit of a sequence of Riemann sums, where the integrand can be evaluated at any point between each pair of partition points. It is well known that this choice of evaluation of the integrand is not valid if the integrator is Brownian motion, that is, $H = 1/2$.

If f is smooth, then an application of an Itô formula is

$$f(B_t^H) = f(0) + \int_0^t f'(B_s^H) dB_s^H + H \int_0^t s^{2H-1} f''(B_s^H) ds,$$

where prime denotes differentiation and $H \in (1/2, 1)$. It is interesting to note that if $H = 1/2$ is formally substituted in the equation, then the well-known Itô formula for Brownian motion is obtained.

2. Fractional Brownian motion. Let $\Omega = C_0(\mathbb{R}_+, \mathbb{R})$ be the space of real-valued continuous functions on $\mathbb{R}_+$ with the initial value zero and the topology of local uniform convergence. There is a probability measure P^H on $(\Omega, \mathcal{F})$, where $\mathcal{F}$ is the Borel σ-algebra such that on the probability space $(\Omega, \mathcal{F}, P^H)$ the coordinate process $B^H : \Omega \to \mathbb{R}$ defined as

$$B_t^H(\omega) = \omega(t), \quad \omega \in \Omega,$$

is a Gaussian process satisfying (1.1). The process $(B_t^H, t \geq 0)$ is called the canonical (standard) fractional Brownian motion with Hurst parameter H. In this paper only this canonical process and its associated probability space are used. Throughout this paper it is assumed that $H \in (\frac{1}{2}, 1)$ is arbitrary but fixed. Clearly if $H = 1/2$, the fractional Brownian motion is the standard Brownian motion.

It is elementary to verify that a fractional Brownian motion for $H \neq 1/2$ is not a semimartingale. It is known [20] that if the (usual) stochastic integral $\int_a^b f_s \, dX_s$ is well defined for a large family of integrands with respect to a process $(X_t, t \geq 0)$, then this process $(X_t, t \geq 0)$ is a semimartingale. Thus the well-developed classical theory for semimartingales cannot be applied here, and the stochastic integral with respect to fractional Brownian motions needs to be developed.

Let $\phi : \mathbb{R}_+ \times \mathbb{R}_+ \to \mathbb{R}_+$ be given by

$$\phi(s,t) = H(2H-1)|s-t|^{2H-2}. \tag{2.1}$$

Many results of this paper can be extended to a more general $\phi(s,t)$ that is symmetric and positive definite, so ϕ in (2.1) is given as a function of two variables and not their difference. Let $f : \mathbb{R}_+ \to \mathbb{R}$ be a Borel measurable (deterministic) function. The function f belongs to the Hilbert space $L^2_\phi(\mathbb{R}_+)$ if

$$|f|_\phi^2 := \int_0^\infty \int_0^\infty f(s) f(t) \phi(s,t) ds \, dt < \infty. \tag{2.2}$$

The inner product on the Hilbert space L^2_ϕ is denoted by $\langle \cdot, \cdot \rangle_\phi$.

The stochastic (Wiener) integral with respect to fractional Brownian motions for deterministic kernels is easily defined.

LEMMA 2.1. *If $f, g \in L^2_\phi(\mathbb{R}_+)$, then $\int_0^\infty f_s \, dB_s^H$ and $\int_0^\infty g_s \, dB_s^H$ are well defined zero mean, Gaussian random variables with variances $|f|_\phi^2$ and $|g|_\phi^2$, respectively, and*

$$\mathbb{E}\left(\int_0^\infty f_s \, dB_s^H \int_0^\infty g_s \, dB_t^H \right) = \int_0^\infty \int_0^\infty f(s) g(t) \phi(s,t) ds dt = \langle f, g \rangle_\phi. \tag{2.3}$$

This lemma is verified in [7]. It can be proved directly by verifying it for simple functions $\sum_{i=1}^n a_i \chi_{[t_i, t_{i+1}]}(s)$ and then proceeding with a passage to the limit.

3. Stochastic integration for fractional Brownian motions. Let $(\Omega,\mathcal{F},P^H)$ be the probability space from section 2 where a fractional Brownian motion with Hurst parameter H is well defined. The probability measure P^H depends on H. Throughout this paper the Hurst parameter H is fixed such that $1/2 < H < 1$. Since H is fixed, the probability measure is denoted by P.

Let $L^p(\Omega,\mathcal{F},P) = L^p$ be the space of all random variables $F:\Omega\to\mathbb{R}$ such that

$$\|F\|_p := (\mathbb{E}|F|^p)^{1/p} < \infty$$

and let $L^2_\phi(\mathbb{R}_+) = \{f | f:\mathbb{R}_+\to\mathbb{R},\ |f|^2_\phi := \int_0^\infty\int_0^\infty f_s f_t \phi(s,t)ds\,dt < \infty\}$. Often for notational simplicity $L^2_\phi(\mathbb{R}_+)$ is denoted by L^2_ϕ. For any $f\in L^2_\phi$, define $\varepsilon : L^2_\phi \to L^1(\Omega,\mathcal{F},P)$ as

$$\begin{aligned}\varepsilon(f) :&= \exp\left\{\int_0^\infty f_t\,dB^H_t - \frac{1}{2}\int_0^\infty\int_0^\infty f_s f_t\phi(s,t)ds\,dt\right\}\\ &= \exp\left[\int_0^\infty f_t\,dB^H_t - \frac{1}{2}|f|^2_\phi\right].\end{aligned}\tag{3.1}$$

If $f\in L^2_\phi$, then $\varepsilon(f)\in L^p(\Omega,\mathcal{F},P)$ for each $p\ge 1$ and $\varepsilon(f)$ is called an exponential function (e.g., [21]). The Hilbert space L^2_ϕ is naturally associated with the Gaussian process, fractional Brownian motion, from the formulation as an abstract Wiener space. The Hilbert space plays a basic role for questions of absolute continuity [6] and the exponential function (3.1) is a Radon–Nikodym derivative for a translate of the fractional Brownian motion.

Let $\mathcal{E}$ be the linear span of the exponentials, that is,

$$\mathcal{E} = \left\{\sum_{k=1}^n a_k\varepsilon(f_k),\ n\in\mathbb{N}, a_k\in\mathbb{R},\ f_k\in L^2_\phi(\mathbb{R}_+) \text{ for } k\in\{1,\dots,n\}\right\}.\tag{3.2}$$

THEOREM 3.1. *$\mathcal{E}$ is a dense set of $L^p(\Omega,\mathcal{F},P)$ for each $p\ge 1$. In particular, $\mathcal{E}$ is a dense set of $L^2(\Omega,\mathcal{F},P)$.*

Proof. A random variable $F:\Omega\to\mathbb{R}$ is said to be a polynomial of the fractional Brownian motion if there is a polynomial $p(x_1,x_2,\dots,x_n)$ such that

$$F = p\left(B^H_{t_1}, B^H_{t_2},\dots,B^H_{t_n}\right)$$

for some $0\le t_1 < t_2 < \cdots < t_n$. Since $(B^H_t, t\ge 0)$ is a Gaussian process it is well known that the set of all polynomial fractional Brownian functionals is dense in $L^p(\Omega,\mathcal{F},P)$ for $p\ge 1$. In this case, the denseness of the polynomials follows from the continuity of the process and the Stone–Weierstrass theorem. To prove the theorem it is only necessary to prove that any polynomial can be approximated by the elements in $\mathcal{E}$. Since the Wick product of exponentials is still an exponential it is easy to see that it is only necessary to show that for any $t>0$, B^H_t can be approximated by elements in $\mathcal{E}$.

Let $f_\delta(s) = \chi_{[0,t]}(s)\delta$. It is clear that for $\delta>0$, f_δ is in L^2_ϕ, and $\varepsilon(f_\delta) = c(\delta)e^{\delta B^H_t}$ for some positive constant $c(\delta)$. It is easy to see that

$$F_\delta = \frac{\varepsilon(f_\delta) - c(\delta)}{c(\delta)\delta} = \frac{e^{\delta B^H_t} - 1}{\delta}$$

is in $\mathcal{E}$. If $\delta \to 0$, then $F_\delta \to B_t^H$ in $L^p(\Omega, \mathcal{F}, P)$ for each $p \geq 1$. This completes the proof. □

THEOREM 3.2. *If $f_1, f_2, \dots, f_n$ are elements in L_ϕ^2 such that $|f_i - f_j|_\phi \neq 0$ for $i \neq j$, then $\varepsilon(f_1)$, $\varepsilon(f_2), \dots, \varepsilon(f_n)$ are linearly independent in L_ϕ^2.*

Proof. This theorem is known to be true if the fractional Brownian motion is replaced by a standard Brownian motion (e.g., [21]).

Let $f_1, f_2, \dots, f_k$ be distinct elements in L_ϕ^2. Let $\lambda_1, \lambda_2, \dots, \lambda_k$ be real numbers such that

$$|\lambda_1 \varepsilon(f_1) + \lambda_2 \varepsilon(f_2) + \cdots + \lambda_k \varepsilon(f_k)|_\phi = 0.$$

Thus for any $g \in L_\phi^2$,

$$\mathbb{E}[\{\lambda_1 \varepsilon(f_1) + \lambda_2 \varepsilon(f_2) + \cdots + \lambda_k \varepsilon(f_k)\}\varepsilon(g)] = 0.$$

By an elementary computation for Gaussian random variables it follows that

$$\lambda_1 e^{\langle f_1, g\rangle_\phi} + \lambda_2 e^{\langle f_2, g\rangle_\phi} + \cdots + \lambda_k e^{\langle f_k, g\rangle_\phi} = 0.$$

Replace g by δg for $\delta \in \mathbb{R}$ to obtain

$$\lambda_1 e^{\delta\langle f_1, g\rangle_\phi} + \lambda_2 e^{\delta\langle f_2, g\rangle_\phi} + \cdots + \lambda_k e^{\delta\langle f_k, g\rangle_\phi} = 0.$$

Expand the above identity in the powers of δ and compare the coefficients of δ^p for $p \in \{0, 1, \cdots, k-1\}$ to obtain the family of equations

$$\lambda_1 \langle f_1, g\rangle_\phi^p + \lambda_2 \langle f_2, g\rangle_\phi^p + \cdots + \lambda_k \langle f_k, g\rangle_\phi^p = 0$$

for $p = 0, 1, \cdots, k-1$. This is a linear system of k equations and k unknowns. By the Vandermonde formula, the determinant of this linear system is

$$\det\left(\langle f_i, g\rangle_\phi^p\right) = \prod_{i<j} \langle f_i - f_j, g\rangle_\phi^p.$$

For every pair (i, j) with $i \neq j$, the set $\{g \in L_\phi^2 : \langle f_i - f_j, g\rangle_\phi \neq 0\}$ is the complement of a hyperplane in L_ϕ^2. Since the intersection of finitely many complements of hyperplanes in L_ϕ^2 is not empty, there is a $g \in L_\phi^2$ such that $\langle f_i - f_j, g\rangle_\phi \neq 0$ for all pairs i and j such that $i \neq j$. Thus $\lambda_1 = \lambda_2 = \cdots = \lambda_k = 0$. This proves the theorem. □

The above two theorems reduce many verifications for functions in $L^2(\Omega, \mathcal{F}, P)$ to verifications of exponentials in $\mathcal{E}$.

The following result is an absolute continuity of measures for some translates of fractional Brownian motion.

THEOREM 3.3. *If $F : \Omega \to \mathbb{R}$ is a random variable such that $F \in L^p(\Omega, \mathcal{F}, P)$ for some $p \geq 1$, then*

(3.3)
$$\mathbb{E}\left\{F\left(B_\cdot^H + \int_0^\cdot (\Phi g)(s)ds\right)\right\} = \mathbb{E}\left\{F(B_\cdot^H) e^{\int_0^\infty g_s\, dB_s^H - \frac{1}{2}\int_0^\infty \int_0^\infty \phi(u,v) g_u g_v\, du\, dv}\right\},$$

where Φ is given by

$$(\Phi g)(t) = \int_0^\infty \phi(t, u) g_u\, du$$

and $g \in L^2_\phi$.

Proof. The term $F(B^H_\cdot)$ denotes $F(\omega)$. Let $k \in L^2_\phi$ and

$$F(B^H_\cdot) = \varepsilon(k) = e^{\int_0^\infty k_s dB^H_s - \frac{1}{2}\int_0^\infty \int_0^\infty \phi(u,v)k_u k_v \, du\, dv}.$$

Then

$$F\left(B^H_\cdot + \int_0^\cdot (\Phi g)_s \, ds\right) = F(B^H_\cdot) e^{\int_0^\infty k_u (\Phi g)_u \, du}.$$

So

$$\mathbb{E}\left\{F\left(B^H_\cdot + \int_0^\cdot (\Phi g)_s \, ds\right)\right\} = e^{\int_0^\infty k_u (\Phi g)_u \, du}.$$

Furthermore, it follows that

$$\mathbb{E}\left\{F\big(B^H_\cdot\big)\varepsilon(g)\right\} = e^{\int_0^\infty \int_0^\infty \phi(u,v)k_u g_v \, du\, dv} = e^{\int_0^\infty k_u (\Phi g)_u \, du}.$$

Thus the theorem is true if F is an exponential function $\varepsilon(f) \in \mathcal{E}$. A limiting argument completes the proof. □

Let a Radon–Nikodym derivative $\frac{d\tilde{P}}{dP}$ on $(\Omega, \mathcal{F}, P)$ be given by

$$\frac{d\tilde{P}}{dP} = e^{\int_0^\infty g_s \, dB^H_s - \frac{1}{2}\int_0^\infty \int_0^\infty \phi(u,v) g_u g_v \, du\, dv},$$

and denote the expectation with respect to $\tilde{P}$ by $\tilde{\mathbb{E}}$; then (3.3) is given by

$$\mathbb{E}F\left(B^H_\cdot + \int_0^\cdot (\Phi g)_s \, ds\right) = \tilde{\mathbb{E}}(F(B^H_\cdot)).$$

For a random variable F in $L^p(\Omega, \mathcal{F}, P)$ $(p \geq 1)$ and a function $g \in L^2_\phi$, the random variable $F(B^H_\cdot + \int_0^\cdot (\Phi g)(v) dv)$ is well defined. An analogue of the Malliavin derivative [23] is introduced.

Definition 3.4. *The ϕ-derivative of a random variable $F \in L^p$ in the direction of Φg where $g \in L^2_\phi$ is defined as*

$$D_{\Phi g} F(\omega) = \lim_{\delta \to 0} \frac{1}{\delta}\left\{F(\omega + \delta \int_0^\cdot (\Phi g)(u) du) - F(\omega)\right\} \tag{3.4}$$

if the limit exists in $L^p(\Omega, \mathcal{F}, P)$. Furthermore, if there is a process $(D^\phi F_s, s \geq 0)$ such that

$$D_{\Phi g} F = \int_0^\infty D^\phi F_s g_s \, ds \quad \textit{almost surely (a.s.)}$$

for all $g \in L^2_\phi$, then F is said to be ϕ-differentiable.

The higher order derivatives can be defined in a similar manner.

Definition 3.5. *Let $F : [0,T] \times \Omega \to \mathbb{R}$ be a stochastic process. The process F is said to be ϕ-differentiable if for each $t \in [0,T]$, $F(t,\cdot)$ is ϕ-differentiable and $D^\phi_s F_t$ is jointly measurable.*

It is easy to verify an elementary version of a chain rule, that is, if $f : \mathbb{R} \to \mathbb{R}$ is a smooth function and $F : \Omega \to \mathbb{R}$ is ϕ-differentiable, then $f(F)$ is also ϕ-differentiable and

$$D_{\Phi g} f(F) = f'(F) D_{\Phi g} F$$

and

$$D_s^\phi f(F) = f'(F) D_s^\phi F$$

and the iterated directional derivatives

$$D_{\Phi g_1} D_{\Phi g_2} f(F) = f'(F) D_{\Phi g_1} D_{\Phi g_2} F + f''(F) D_{\Phi g_1} F D_{\Phi g_2} F.$$

The following rules for differentiation, which can be verified as in the proof of Proposition 3.6, are useful later:

$$D_{\Phi g} \int_0^\infty f_s \, dB_s^H = \int_0^\infty \int_0^\infty \phi(u,v) f_u g_v \, du \, dv = \langle f, g \rangle_\phi; \tag{3.5}$$

$$D_s^\phi \int_0^\infty f_u \, dB_u^H = \int_0^\infty \phi(u,s) f_u \, du = (\Phi f)(s); \tag{3.6}$$

$$D_{\Phi g} \varepsilon(f) = \varepsilon(f) \int_0^\infty \int_0^\infty \phi(u,v) f_u g_v \, du \, dv = \varepsilon(f) \langle f, g \rangle_\phi; \tag{3.7}$$

$$D_s^\phi \varepsilon(f) = \varepsilon(f) \int_0^\infty \phi(u,s) f_u \, du = \varepsilon(f)(\Phi f)(s), \tag{3.8}$$

where $f, g \in L_\phi^2$.

Now the Wick product $\diamond$ of two functionals is introduced. To extend the theory of stochastic calculus for Brownian motions to the fractional Brownian motions, the Wick calculus for Gaussian processes (or Gaussian measures) is used. The Wick product of two exponentials $\varepsilon(f)$ and $\varepsilon(g)$ is defined as

$$\varepsilon(f) \diamond \varepsilon(g) = \varepsilon(f + g). \tag{3.9}$$

Since for distinct $f_1, f_2, \ldots, f_n$ in L_ϕ^2, $\varepsilon(f_1), \varepsilon(f_2), \ldots, \varepsilon(f_n)$ are linearly independent, this definition can be extended to define the Wick product $F \diamond G$ of two functionals F and G in $\mathcal{E}$.

Note that $\int_0^\infty g_s \, dB_s^H$ is not an element in $\mathcal{E}$. The Wick product is extended to more general functionals, including the functionals of the form $\int_0^\infty g_s \, dB_s^H$, where $g \in L_\phi^2$.

PROPOSITION 3.6. *If $g \in L_\phi^2$, $F \in L^2(\Omega, \mathcal{F}, P)$, and $D_{\Phi g} F \in L^2(\Omega, \mathcal{F}, P)$, then*

$$F \diamond \int_0^\infty g_s \, dB_s^H = F \int_0^\infty g_s \, dB_s^H - D_{\Phi g} F. \tag{3.10}$$

Proof. By the definition (3.9),

$$\varepsilon(f) \diamond \varepsilon(\delta g) = \varepsilon(f + \delta g). \tag{3.11}$$

Differentiate the above identity with respect to δ, and evaluate at $\delta = 0$ to obtain

$$\begin{aligned}\varepsilon(f) \diamond \int_0^\infty g_s \, dB_s^H &= \varepsilon(f) \left[\int_0^\infty g_s \, dB_s^H - \langle f, g \rangle_\phi \right] \\ &= \varepsilon(f) \int_0^\infty g_s \, dB_s^H - \varepsilon(f) \langle f, g \rangle_\phi. \end{aligned} \tag{3.12}$$

By (3.7), it follows that the last term of the above expression is $D_{\Phi g} \varepsilon(f)$. Thus, the following equality is satisfied:

$$\varepsilon(f) \diamond \int_0^\infty g_s \, dB_s^H = \varepsilon(f) \int_0^\infty g_s \, dB_s^H - D_{\Phi g} \varepsilon(f). \tag{3.13}$$

If $F \in \mathcal{E}$ is a finite linear combination of $\varepsilon(f_1)$, $\varepsilon(f_2), \ldots, \varepsilon(f_n)$, then extend (3.13) by linearity

$$\begin{aligned} F \diamond \int_0^\infty g_s \, dB_s^H &= F \int_0^\infty g_s \, dB_s^H - D_{\Phi g} F \\ &= F \int_0^\infty g_s \, dB_s^H - \int_0^\infty D_s^\phi F g_s \, ds. \end{aligned} \tag{3.14}$$

The proof of the proposition is completed by Theorem 3.1. □

Now the second moment of (3.10) is computed. Note that by a simple computation for Gaussian random variables, it follows that

$$\mathbb{E}(\varepsilon(f)\varepsilon(g)) = \exp\{\langle f, g \rangle_\phi\}.$$

Thus

$$\begin{aligned} \mathbb{E}\{(\varepsilon(f) \diamond \varepsilon(\gamma g))(\varepsilon(h) \diamond \varepsilon(\delta g))\} &= \mathbb{E}\{\varepsilon(f + \gamma g)\varepsilon(h + \delta g)\} \\ &= \exp\{\langle f + \gamma g, h + \delta g \rangle_\phi\}. \end{aligned}$$

Both sides of this equality are functions of γ and δ. Taking the partial derivative $\frac{\partial^2}{\partial \gamma \partial \delta}$, evaluated at $\gamma = \delta = 0$, it follows that

$$\begin{aligned} \mathbb{E}\left\{ \left(\varepsilon(f) \diamond \int_0^\infty g_s \, dB_s^H \right) \left(\varepsilon(h) \diamond \int_0^\infty g_s \, dB_s^H \right) \right\} \\ = \exp(\langle f, h \rangle_\phi) \{ \langle f, g \rangle_\phi \langle h, g \rangle_\phi + \langle g, g \rangle_\phi \} \\ = \mathbb{E}\{ D_{\Phi g} \varepsilon(f) D_{\Phi g} \varepsilon(h) + \varepsilon(f)\varepsilon(h) \langle g, g \rangle_\phi \}. \end{aligned}$$

Thus

$$\begin{aligned} \mathbb{E}\left\{ \left(\varepsilon(f) \diamond \int_0^\infty g_s \, dB_s^H \right) \left(\varepsilon(h) \diamond \int_0^\infty g_s \, dB_s^H \right) \right\} \\ = \mathbb{E}(D_{\Phi g} \varepsilon(f) D_{\Phi g} \varepsilon(h) + \varepsilon(f)\varepsilon(h) \langle g, g \rangle_\phi). \end{aligned}$$

By bilinearity, for any F and G in $\mathcal{E}$, the following equality is satisfied:

$$\mathbb{E}\left\{ \left(F \diamond \int_0^\infty g_s \, dB_s^H \right) \left(G \diamond \int_0^\infty g_s \, dB_s^H \right) \right\} = \mathbb{E}\{ D_{\Phi g} F D_{\Phi g} G + FG \langle g, g \rangle_\phi \}.$$

Let F be equal to G. Then

$$\mathbb{E}\left(F \diamond \int_0^\infty g_s \, dB_s^H\right)^2 = \mathbb{E}\big[(D_{\Phi g}F)^2 + F^2|g|_\phi^2\big].$$

This result is stated in the following theorem.

THEOREM 3.7. *Let $g \in L_\phi^2$ and let $\mathcal{E}_g$ be the completion of $\mathcal{E}$ under the norm*

$$\|F\|_g^2 = \mathbb{E}\Big\{(D_{\Phi g}F)^2 + F^2\Big\},$$

where F is a random variable. Then for any element $F \in \mathcal{E}_g$, $F \diamond \int_0^\infty g_s \, dB_s^H$ is well defined and

$$\mathbb{E}\left(F \diamond \int_0^\infty g_s \, dB_s^H\right)^2 = \mathbb{E}\Big\{(D_{\Phi g}F)^2 + F^2|g|_\phi^2\Big\}. \tag{3.15}$$

By the polarization technique [21], there is the following corollary:

COROLLARY 3.8. *Let $g, h \in L_\phi^2$ and $F, G \in \mathcal{E}$. Then*

$$\mathbb{E}\left(F \diamond \int_0^\infty g_s \, dB_s^H \; G \diamond \int_0^\infty h_s \, dB_s^H\right) = \mathbb{E}[D_{\Phi g}F D_{\Phi h}G + FG\langle g, h\rangle_\phi].$$

This equality is the starting point for the definition of the stochastic integral with respect to the fractional Brownian motions. Let $F \in \mathcal{E}$. The stochastic integral $\int_0^T F_s \, dB_s^H$ is defined, and some properties associated with this stochastic integral are studied.

Consider an arbitrary partition of $[0, T]$, $\pi : 0 = t_0 < t_1 < t_1 < \cdots < t_n = T$. First, the following Riemann sum is given using the Wick product introduced above:

$$S(F, \pi) = \sum_{i=0}^{n-1} F_{t_i} \diamond (B_{t_{i+1}}^H - B_{t_i}^H).$$

From (3.9), it easily follows that for any F and G in $\mathcal{E}$, $\mathbb{E}(F \diamond G) = \mathbb{E}(F)\mathbb{E}(G)$. This identity extends to more general F and G such that $F \diamond G$ is well defined (e.g. [8], p. 83). Thus for any partition π,

$$\begin{aligned}\mathbb{E}\left(\sum_{i=0}^{n-1} F_{t_i} \diamond (B_{t_{i+1}}^H - B_{t_i}^H)\right) &= \sum_{i=0}^{n-1} \mathbb{E}\left(F_{t_i} \diamond (B_{t_{i+1}}^H - B_{t_i}^H)\right) \\ &= \sum_{i=0}^{n-1} \mathbb{E}(F_{t_i})\mathbb{E}\Big(B_{t_{i+1}}^H - B_{t_i}^H\Big) = 0.\end{aligned}$$

To compute the L^2 norm of $S(F, \pi)$, denote

$$\sigma_{ij} = \mathbb{E}\left\{\left(F_{t_i} \diamond (B_{t_{i+1}}^H - B_{t_i}^H)\right)\left(F_{t_j} \diamond (B_{t_{j+1}}^H - B_{t_j}^H)\right)\right\}.$$

By Corollary 3.8, it follows that

$$\sigma_{ij} = \mathbb{E}\left\{\int_{t_i}^{t_{i+1}} D_s^\phi F_{t_i} \, ds \int_{t_j}^{t_{j+1}} D_t^\phi F_{t_j} \, dt + F_{t_i} F_{t_j} \int_{t_i}^{t_{i+1}} \int_{t_j}^{t_{j+1}} \phi(u, v) du dv\right\}.$$

Thus

$$\mathbb{E}S(F,\pi)^2 = \sum_{i,j=0}^{n-1} \mathbb{E}\Big\{ \int_{t_i}^{t_{i+1}} D_s^\phi F_{t_i}\, ds \int_{t_j}^{t_{j+1}} D_t^\phi F_{t_j}\, dt + F_{t_i}F_{t_j} \int_{t_i}^{t_{i+1}} \int_{t_j}^{t_{j+1}} \phi(u,v) du dv \Big\}.$$

Denote $|\pi| := \max_i (t_{i+1} - t_i)$ and $F_t^\pi = F_{t_i}$ if $t_i \le t < t_{i+1}$. Assume that as $|\pi| \to 0$, $\mathbb{E}|F^\pi - F|_\phi^2 \to 0$ and

$$\sum_{i=0}^{n-1} \mathbb{E}\left\{ \int_{t_i}^{t_{i+1}} |D_s^\phi F_{t_i} - D_s^\phi F_s| ds \right\}^2$$

converges to 0. Then from the above it is easy to see that if $(\pi_n, n \in \mathbb{N})$ is a sequence of partitions such that $|\pi_n| \to 0$ as $n \to \infty$, then $(S(F,\pi_n), n \in \mathbb{N})$ is a Cauchy sequence in $L^2(\Omega, \mathcal{F}, P)$. The limit of this sequence in $L^2(\Omega, \mathcal{F}, P)$ is defined as $\int_0^T F_s\, dB_s^H$: that is, define

$$\int_0^T F_s\, dB_s^H = \lim_{|\pi| \to 0} \sum_{i=0}^{n-1} F_{t_i}^\pi \diamond (B_{t_{i+1}}^H - B_{t_i}^H) \tag{3.16}$$

so that

$$\mathbb{E}\left| \int_0^T F_s\, dB_s^H \right|^2 = \mathbb{E}\left\{ \left(\int_0^T D_s^\phi F_s\, ds \right)^2 + |F|_\phi^2 \right\}.$$

Let $\mathcal{L}(0,T)$ be the family of stochastic processes on $[0,T]$ such that $F \in \mathcal{L}(0,T)$ if $\mathbb{E}|F|_\phi^2 < \infty$, F is ϕ-differentiable, the trace of $(D_s^\phi F_t, 0 \le s \le T, 0 \le t \le T)$ exists, $\mathbb{E}\int_0^T (D_s^\phi F_s)^2\, ds < \infty$, and for each sequence of partitions $(\pi_n, n \in \mathbb{N})$ such that $|\pi_n| \to 0$ as $n \to \infty$, the quantities

$$\sum_{i=0}^{n-1} \mathbb{E}\left\{ \int_{t_i^{(n)}}^{t_{i+1}^{(n)}} |D_s^\phi F_{t_i^{(n)}}^\pi - D_s^\phi F_s| ds \right\}^2$$

and

$$\mathbb{E}|F^\pi - F|_\phi^2$$

tend to 0 as $n \to \infty$, where $\pi_n : 0 = t_0^{(n)} < t_1^{(n)} < \cdots < t_{n-1}^{(n)} < t_n^{(n)} = T$.

The following result summarizes the above construction of a stochastic integral.

THEOREM 3.9. *Let $(F_t, t \in [0,T])$ be a stochastic process such that $F \in \mathcal{L}(0,T)$. The limit* (3.16) *exists and this limit is defined as $\int_0^T F_s\, dB_s^H$. Moreover, this integral satisfies $\mathbb{E}\int_0^T F_s\, dB_s^H = 0$ and*

$$\mathbb{E}\left| \int_0^T F_s\, dB_s^H \right|^2 = \mathbb{E}\left\{ \left(\int_0^T D_s^\phi F_s\, ds \right)^2 + |1_{[0,T]} F|_\phi^2 \right\}. \tag{3.17}$$

The following properties follow directly from the above theorem.

(1) If $F, G \in \mathcal{L}(0,T)$, then

$$\int_0^t \big(aF_s + bG_s\big)dB_s^H = a\int_0^t F_s\,dB_s^H + b\int_0^t G_s\,dB_s^H \quad \text{a.s.}$$

for any constants a and b and $t \in (0,T]$.

(2) If $F \in \mathcal{L}(0,T)$, $\mathbb{E}[\sup_{0\le s\le T} F_s]^2 < \infty$, and $\sup_{0\le s\le T}\mathbb{E}|D_s^\phi F_s|^2 < \infty$, then $(\int_0^t F_s\,dB_s^H, 0 \le t \le T)$ has a continuous version.

Property (1) is obvious. To show (2) let $Y_t = \int_0^t F_s\,dB_s^H, 0 \le t \le T$. By the equality (3.17), it follows that

$$\begin{aligned}
\mathbb{E}|Y_t - Y_s|^2 &= \mathbb{E}\left|\int_s^t F_u\,dB_u^H\right|^2 \\
&\le \mathbb{E}\left\{\left(\int_s^t D_u^\phi F_u\,du\right)^2 + \int_s^t\int_s^t F_uF_v\phi(u,v)dudv\right\} \\
&\le (t-s)\int_s^t \mathbb{E}|D_u^\phi F_u|^2\,du + \mathbb{E}[\sup_{0\le s\le T} F_s]^2\int_s^t\int_s^t \phi(u,v)dudv \\
&\le (t-s)^2 + C(t-s)^{2H}.
\end{aligned}$$

By the Kolmogorov lemma [22], property (2) is satisfied.

In Theorem 3.9, it is not assumed that the stochastic process $(F_s, s \in [0,T])$ is adapted to the fractional Brownian motion. Now assume that $D_s^\phi F_s = 0$ for all $s \in [0,T]$. Thus in this case,

$$\mathbb{E}\left|\int_0^T F_s\,dB_s^H\right|^2 = \mathbb{E}\left\{\int_0^T\int_0^T F_uF_v\phi(u,v)du\,dv\right\}.$$

This fact is stated in the following theorem.

THEOREM 3.10. *If $F \in \mathcal{L}(0,T)$ and if F satisfies $\mathbb{E}\int_0^T |D_s^\phi F_s|ds = 0$, then*

$$\mathbb{E}\left|\int_0^T F_s\,dB_s^H\right|^2 = \mathbb{E}\{|1_{[0,T]}F|_\phi^2\}.$$

An analogue of the stochastic integral of Stratonovich type $\int_0^t F_s\delta B_s^H$ is also introduced. This type of integral is related to the integrals introduced by Lin [17] and Dai and Heyde [2].

DEFINITION 3.11. *Let $(\pi_n, n \in \mathbb{N})$ be a sequence of partitions of $[0,t]$ such that $|\pi_n| \to 0$ as $n \to \infty$. If $\sum_{i=0}^{n-1} f(t_i^{(n)})(B^H(t_{i+1}^{(n)}) - B^H(t_i^{(n)}))$ converges in $L^2(\Omega,\mathcal{F},P)$ to the same limit for all such sequences $(\pi_n, n \in \mathbb{N})$, then this limit is called the stochastic integral of Stratonovich type and the limit is denoted by $\int_0^t f(s)\delta B^H(s)$.*

THEOREM 3.12. *If $F \in \mathcal{L}(0,t)$, then the stochastic integral of Stratonovich type $\int_0^t F_s\delta B_s^H$ exists and the following equality is satisfied:*

$$\int_0^t F_s\delta B_s^H = \int_0^t F_s\,dB_s^H + \int_0^t D_s^\phi F_s\,ds \quad a.s.$$

Proof. By Proposition 3.6,

$$\begin{aligned}
&\sum_{i=0}^{n-1} F_{t_i^{(n)}}\left(B^H\left(t_{i+1}^{(n)}\right) - B^H\left(t_i^{(n)}\right)\right) \\
&\quad = \sum_{i=0}^{n-1} F_{t_i^{(n)}} \diamond \left(B^H\left(t_{i+1}^{(n)}\right) - B^H\left(t_i^{(n)}\right)\right) + \sum_{i=0}^{n-1} D_{\Phi\chi_{\left[t_i^{(n)}, t_{i+1}^{(n)}\right]}} F_{t_i^{(n)}} \\
&\quad = \sum_{i=0}^{n-1} \left[F_{t_i^{(n)}} \diamond \left(B^H\left(t_{i+1}^{(n)}\right) - B^H\left(t_i^{(n)}\right)\right) + \int_{t_i^{(n)}}^{t_{i+1}^{(n)}} D_s^{\phi} F_{t_i^{(n)}}\, ds\right].
\end{aligned}$$

This equality proves the Theorem. □

These two types of stochastic integrals are both interesting:

(1) The expectation of $\int_0^t F_s\, dB_s^H$ is 0, but the chain rule for this type of integral is more complicated than for the integral of Stratonovich type.

(2) The chain rule for the integral of Stratonovich type is simple, but $\mathbb{E}\int_0^t F_s \delta B_s^H \neq 0$ in general.

An example is provided that shows that $\mathbb{E}\{\int_0^t F_s \delta B_s^H\}$ is not 0.

It is well known that if X is a standard normal random variable, $X \sim N(0,1)$, then

$$\mathbb{E}X^n = \begin{cases} \frac{n!}{(\sqrt{2})^n (n/2)!} & \text{if } n \text{ is even,} \\ 0 & \text{if } n \text{ is odd.} \end{cases}$$

Let $f(x) = x^n$. If n is odd, then

$$\begin{aligned}
\mathbb{E}\int_0^t f(B_s^H)\delta B_s^H &= \mathbb{E}\int_0^t D_s^{\phi} f(B_s^H) ds \\
&= \mathbb{E}\int_0^t f'(B_s^H) D_s^{\phi} B_s^H\, ds \\
&= \mathbb{E}\int_0^t f'(B_s^H) \int_0^s \phi(u,s) du\, ds \\
&= H\int_0^t s^{2H-1}\mathbb{E} f'(B_s^H) ds \\
&= nH\int_0^t s^{2H-1}\mathbb{E}((B_s^H)^{n-1}) ds \\
&= nH\int_0^t s^{2H-1}\mathbb{E}\left(\frac{B_s^H}{s^H}\right)^{n-1} s^{nH-H} ds \\
&= \frac{n! H t^{(n+1)H}}{2^{\frac{n-1}{2}}(n+1)H\left(\frac{n-1}{2}\right)!},
\end{aligned}$$

which is not 0. If n is even, then by the same computation,

$$\mathbb{E}\int_0^t (B_s^H)^n \delta B_s^H = 0.$$

Now another interesting phenomenon is shown. Let π be a partition of the interval $[0,T] : 0 = t_0 < t_1 < t_2 < \cdots < t_n = T$. Let $(f(s), s \geq 0)$ be a smooth stochastic

process on the probability space $(\Omega, \mathcal{F}, P)$. For the Brownian motion $(B_t, t \geq 0)$, the Itô integral can be defined as the limit of the Riemann sums $\sum_{i=0}^{n-1} f_{t_i}(B_{t_{i+1}} - B_{t_i})$ as the partition $|\pi| \to 0$. The Stratonovich integral is defined as the limit of the Riemann sums $\sum_{i=0}^{n-1} \frac{f_{t_i}+f_{t_{i+1}}}{2}(B_{t_{i+1}} - B_{t_i})$ as the partition $|\pi| \to 0$. It may seem to be more natural to define the Stratonovich integral for a fractional Brownian motion $(B_t^H, t \geq 0)$ in a similar way. It is shown that the above two limits are the same for a large class of stochastic processes.

Initially the following lemma is given.

LEMMA 3.13. *Let p be a positive even integer. Then*

$$\mathbb{E}(B_t^H - B_s^H)^p = \frac{p!}{2^{p/2}(p/2)!}|t-s|^{pH}. \tag{3.18}$$

Proof. By (1.1) it follows that

$$\begin{aligned}\mathbb{E}|B_t^H - B_s^H|^2 &= \mathbb{E}(B_t^H)^2 + \mathbb{E}(B_s^H)^2 - 2\mathbb{E}B_t^H B_s^H \\ &= t^{2H} + s^{2H} - (t^{2H} + s^{2H} - |t-s|^{2H}) = |t-s|^{2H}.\end{aligned}$$

Thus $\frac{B_t^H - B_s^H}{|t-s|^H}$ is a standard Gaussian random variable and

$$\begin{aligned}\mathbb{E}|B_t^H - B_s^H|^p &= |t-s|^{pH}\mathbb{E}\left(\frac{B_t^H - B_s^H}{|t-s|^H}\right)^p \\ &= \frac{p!}{2^{p/2}(p/2)!}|t-s|^{pH}.\end{aligned}$$

COROLLARY 3.14. *For each $\alpha > 1$, there is a $C_\alpha < \infty$ such that*

$$\mathbb{E}|B_t^H - B_s^H|^\alpha \leq C_\alpha |t-s|^{\alpha H}. \tag{3.19}$$

DEFINITION 3.15. *The process $(f_s, 0 \leq s \leq T)$ is said to be a bounded quadratic variation process if there are constants $p \geq 1$ and $0 < C_p < \infty$ such that for any partition $\pi : 0 = t_0 < t_1 < t_2 < \cdots < t_n = T$,*

$$\sum_{i=0}^{n-1} (\mathbb{E}|f_{t_{i+1}} - f_{t_i}|^{2p})^{1/p} \leq C_p.$$

Example. Let $f : \mathbb{R} \to \mathbb{R}$ be continuously differentiable with the first derivative bounded by C. Then $f(B_s^H)$ is a bounded quadratic variation process. In fact, for any $p \geq 1$ and partition π,

$$\begin{aligned}&\sum_{i=0}^{n-1} \left\{\mathbb{E}|f(B_{t_{i+1}}^H) - f(B_{t_i}^H)|^{2p}\right\}^{1/p} \\ &\quad= \sum_{i=0}^{n-1} \left\{\mathbb{E}\left(\int_0^1 f'\left(B_{t_i}^H + \theta(B_{t_{i+1}}^H - B_{t_i}^H)\right) d\theta (B_{t_{i+1}}^H - B_{t_i}^H)\right)^{2p}\right\}^{1/p} \\ &\quad\leq C\sum_{i=0}^{n-1} \mathbb{E}\left(|B_{t_{i+1}}^H - B_{t_i}^H|^{2p}\right)^{1/p} \\ &\quad\leq C\sum_{i=0}^{n-1} |t_{i+1} - t_i|^{2H} \leq CT.\end{aligned}$$

THEOREM 3.16. *Let $(f(t), 0 \le t \le T)$ be a bounded quadratic variation process. Let $(\pi_n, n \in \mathbb{N})$ be a sequence of partitions of $[0,T]$ such that $|\pi_n| \to 0$ as $n \to \infty$ and*

$$\left(\sum_{i=0}^{n-1} f\left(t_i^{(n)}\right) \left(B^H\left(t_{i+1}^{(n)}\right) - B^H\left(t_i^{(n)}\right)\right),\ n \in \mathbb{N} \right)$$

converges to a random variable G in $L^2(\Omega, \mathcal{F}, P)$, where $\pi_n = \left\{t_0^{(n)}, \dots, t_n^{(n)}\right\}$. Then

$$\left(\sum_{i=0}^{n-1} f\left(t_{i+1}^{(n)}\right) \left(B^H\left(t_{i+1}^{(n)}\right) - B^H\left(t_i^{(n)}\right)\right),\ n \in \mathbb{N} \right)$$

also converges to G in $L^2(\Omega, \mathcal{F}, P)$.

Proof. It suffices to show that $\sum_{i=0}^{n-1}(f_{t_{i+1}} - f_{t_i})(B^H_{t_{i+1}} - B^H_{t_i})$ converges to 0 in $L^2(\Omega, \mathcal{F}, P)$. Let p be a number as indicated in the definition of bounded quadratic variation for $(f_t, 0 \le t \le T)$:

$$\begin{aligned}
&\left(\mathbb{E}\left\{ \sum_{i=0}^{n-1} (f_{t_{i+1}} - f_{t_i}) \left(B^H_{t_{i+1}} - B^H_{t_i}\right) \right\}^2 \right)^{1/2} \\
&\qquad \le \sum_{i=0}^{n-1} \left(\mathbb{E}(f_{t_{i+1}} - f_{t_i})^2 \mathbb{E}\left(B^H_{t_{i+1}} - B^H_{t_i}\right)^2 \right)^{1/2} \\
&\qquad \le \sum_{i=0}^{n-1} \left(\mathbb{E}(f_{t_{i+1}} - f_{t_i})^{2p} \right)^{1/2p} \left(\mathbb{E}\left(B^H_{t_{i+1}} - B^H_{t_i}\right)^{2q} \right)^{1/2q} \\
&\qquad \le \left\{ \sum_{i=0}^{n-1} \left(\mathbb{E}(f_{t_{i+1}} - f_{t_i})^{2p} \right)^{1/p} \right\}^{1/2} \left\{ \sum_{i=0}^{n-1} \left(\mathbb{E}|B^H_{t_{i+1}} - B^H_{t_i}|^{2q} \right)^{1/q} \right\}^{1/2} \\
&\qquad \le C \left\{ \sum_{i=0}^{n-1} |t_{i+1} - t_i|^{2H} \right\}^{1/2} \\
&\qquad \le C \max_{0 \le i \le n-1} (t_{i+1} - t_i)^{H-1/2} \left\{ \sum_{i=0}^{n-1} |t_{i+1} - t_i| \right\}^{1/2} \\
&\qquad \le C\sqrt{T} \max_{0 \le i \le n-1} (t_{i+1} - t_i)^{H-1/2} \to 0 \quad (\ as\ |\pi| \to 0),
\end{aligned}$$

where $1/p + 1/q = 1$. ∎

It can also be shown with a slightly more lengthy argument that if $(f_s, s \ge 0)$ is a process with bounded quadratic variation and ξ_i is any point in $[t_i, t_{i+1}]$, then a sequence of the Riemann sums $\sum_{i=0}^{n-1} f_{\xi_i}(B^H_{t_{i+1}} - B^H_{t_i})$ converges in $L^2(\Omega, \mathcal{F}, P)$ to $\int_0^T f_s \delta B^H_s$, if it is true for any particular choice of such a family of points ξ_i.

4. An Itô formula. Now an analogue of the Itô formula is established, that is, a chain rule for the integral introduced in the last section. Let $f : \mathbb{R} \to \mathbb{R}$ be a twice continuously differentiable function with bounded second derivative. Then for

a partition $\{t_0, t_1, \dots, t_n\}$ of $[0,T]$, it follows by Taylor's formula that

$$\begin{aligned} f\left(B_T^H\right) - f(0) &= \sum_{i=0}^{n-1} \left[f\left(B_{t_{i+1}}^H\right) - f\left(B_{t_i}^H\right) \right] \\ &= \sum_{i=0}^{n-1} f'\left(B_{t_i}^H\right) \left[B_{t_{i+1}}^H - B_{t_i}^H \right] + \frac{1}{2} \sum_{i=0}^{n-1} f''(\xi_i) \left[B_{t_{i+1}}^H - B_{t_i}^H \right]^2 \\ &= \sum_{i=0}^{n-1} f'\left(B_{t_i}^H\right) \diamond \left[B_{t_{i+1}}^H - B_{t_i}^H \right] + \sum_{i=0}^{n-1} \int_{t_i}^{t_{i+1}} D_s^\phi f'\left(B_{t_i}^H\right) ds \\ &\quad + \frac{1}{2} \sum_{i=0}^{n-1} f''(\xi_i) \left[B_{t_{i+1}}^H - B_{t_i}^H \right]^2 \\ &= I_1 + I_2 + I_3, \end{aligned}$$

where $\xi_i \in (B_{t_i}^H, B_{t_{i+1}}^H)$. Since it is assumed that $H > 1/2$, it follows that $I_3 \to 0$ in $L^2(\Omega, \mathcal{F}, P)$. By the definition of the stochastic integral introduced in the preceding section, the first term I_1 converges to $\int_0^T f'(B_s^H) dB_s^H$ in L^2. By a version of the chain rule for the ϕ-differentiation operator, it follows that, for $s \in [t_i, t_{i+1})$,

$$\begin{aligned} D_s^\phi f'\left(B_{t_i}^H\right) &= f''\left(B_{t_i}^H\right) D_s^\phi B_{t_i}^H \\ &= f''\left(B_{t_i}^H\right) \int_0^{t_i} \phi(u,s) du \\ &= H f''\left(B_{t_i}^H\right) [s^{2H-1} - (s-t_i)^{2H-1}]. \end{aligned}$$

Thus the second sum in the three sums from Taylor's formula converges to $H \int_0^T s^{2H-1} f''(B_s^H) ds$ in L^2. The following chain rule formula is obtained.

THEOREM 4.1. *If $f : \mathbb{R} \to \mathbb{R}$ is a twice continuously differentiable function with bounded derivatives to order two, then*

$$f(B_T^H) - f(B_0^H) = \int_0^T f'(B_s^H) dB_s^H + H \int_0^T s^{2H-1} f''(B_s^H) ds \quad a.s. \tag{4.1}$$

It is interesting to note that this formula implies the usual Itô formula for Brownian motion when $H = 1/2$ is formally substituted in (4.1).

The following theorem shows how to compute the ϕ-derivative of a stochastic integral of Itô type. It can be verified from the product rule and the Riemann sum approximations to the stochastic integral.

THEOREM 4.2. *Let $(F_t, t \in [0,T])$ be a stochastic process in $\mathcal{L}(0,T)$ and $\sup_{0 \le s \le T} \mathbb{E}|D_s^\phi F_s|^2 < \infty$, and let $\eta_t = \int_0^t F_u \, dB_u^H$ for $t \in [0,T]$. Then, for $s, t \in [0,T]$,*

$$D_s^\phi \eta_t = \int_0^t D_s^\phi F_u \, dB_u^H + \int_0^t F_u \phi(s,u) du \quad a.s.$$

Now a more general Itô formula is given.

THEOREM 4.3. *Let $\eta_t = \int_0^t F_u \, dB_u^H$, where $(F_u, 0 \le u \le T)$ is a stochastic process in $\mathcal{L}(0,T)$. Assume that there is an $\alpha > 1 - H$ such that*

$$\mathbb{E}|F_u - F_v|^2 \le C|u-v|^{2\alpha},$$

where $|u-v| \le \delta$ *for some* $\delta > 0$ *and*

$$\lim_{0\le u,v\le t,|u-v|\to 0} \mathbb{E}|D_u^\phi(F_u - F_v)|^2 = 0.$$

Let $f : \mathbb{R}_+ \times \mathbb{R} \to \mathbb{R}$ *be a function having the first continuous derivative in its first variable and the second continuous derivative in its second variable. Assume that these derivatives are bounded. Moreover, it is assumed that* $\mathbb{E}\int_0^T |F_s D_s^\phi \eta_s| ds < \infty$ *and* $(f'(s,\eta_s)F_s, s \in [0,T])$ *is in* $\mathcal{L}(0,T)$*. Then, for* $0 \le t \le T$*,*

$$\begin{aligned} f(t,\eta_t) = f(0,0) &+ \int_0^t \frac{\partial f}{\partial s}(s,\eta_s)ds + \int_0^t \frac{\partial f}{\partial x}(s,\eta_s)F_s\, dB_s^H \\ &+ \int_0^t \frac{\partial^2 f}{\partial x^2}(s,\eta_s)F_s D_s^\phi \eta_s\, ds \quad a.s. \end{aligned} \tag{4.2}$$

Proof. Let π be a partition defined as above by replacing T by t. Then

$$\begin{aligned} f(t,\eta_t) - f(0,0) &= \sum_{k=0}^{n-1} [f(t_{k+1},\eta_{t_{k+1}}) - f(t_k,\eta_{t_k})] \\ &= \sum_{k=0}^{n-1} [f(t_{k+1},\eta_{t_{k+1}}) - f(t_k,\eta_{t_{k+1}})] \\ &\quad + \sum_{k=0}^{n-1} [f(t_k,\eta_{t_{k+1}}) - f(t_k,\eta_{t_k})]. \end{aligned}$$

By the mean value theorem, it is easy to see that the first sum converges to

$$\int_0^t \frac{\partial f}{\partial s}(s,\eta_s)ds$$

in L^2. Now consider the second sum. Using Taylor's formula, it follows that

$$f(t_k,\eta_{t_{k+1}}) - f(t_k,\eta_{t_k}) = \frac{\partial f}{\partial x}(t_k,\eta_{t_k})(\eta_{t_{k+1}} - \eta_{t_k}) + \frac{1}{2}\frac{\partial^2 f}{\partial x^2}(t_k,\tilde{\eta}_{t_k})\left(\eta_{t_{k+1}} - \eta_{t_k}\right)^2,$$

where $\tilde{\eta}_{t_k} \in (\eta_{t_k}, \eta_{t_{k+1}})$. An upper bound is obtained for $\mathbb{E}\left(\eta_{t_{k+1}} - \eta_{t_k}\right)^2$ as follows:

$$\begin{aligned} \mathbb{E}\left(\eta_{t_{k+1}} - \eta_{t_k}\right)^2 &= \mathbb{E}\left(\int_{t_k}^{t_{k+1}} D_s^\phi F_s\, ds\right)^2 + \mathbb{E}\int_{t_k}^{t_{k+1}}\int_{t_k}^{t_{k+1}} F_u F_v \phi(u,v)dudv \\ &\le (t_{k+1} - t_k)\int_{t_k}^{t_{k+1}} \mathbb{E}(D_s^\phi F_s)^2\, ds \\ &\quad + \int_{t_k}^{t_{k+1}}\int_{t_k}^{t_{k+1}} (\mathbb{E}F_u^2)^{1/2}(\mathbb{E}F_v^2)^{1/2}\phi(u,v)dudv \\ &\le C\left[(t_{k+1} - t_k)^2 + \int_{t_k}^{t_{k+1}}\int_{t_k}^{t_{k+1}} \phi(u,v)dudv\right] \\ &\le C(t_{k+1} - t_k)^2 + C(t_{k+1} - t_k)^{2H} \le C(t_{k+1} - t_k)^{2H}, \end{aligned}$$

where $t_{i+1} - t_i < 1$ and C is a constant independent of the partition π that may differ for different applications.

Thus

$$\mathbb{E}\sum_{k=0}^{n-1}\frac{\partial^2 f}{\partial x^2}(t_k,\tilde{\eta}_{t_k})\left(\eta_{t_{k+1}}-\eta_{t_k}\right)^2 \le C\sum_{k=0}^{n-1}\mathbb{E}\left(\eta_{t_{k+1}}-\eta_{t_k}\right)^2$$
$$\le C\sum_{k=0}^{n-1}(t_{k+1}-t_k)^{2H}\to 0 \quad \text{as } |\pi|\to 0.$$

On the other hand,

$$\frac{\partial f}{\partial x}(t_k,\eta_{t_k})\left(\eta_{t_{k+1}}-\eta_{t_k}\right)=\frac{\partial f}{\partial x}(t_k,\eta_{t_k})\left(F_{t_k}\diamond\left(B^H_{t_{k+1}}-B^H_{t_k}\right)\right)$$
$$+\frac{\partial f}{\partial x}(t_k,\eta_{t_k})\left(\int_{t_k}^{t_{k+1}}(F_s-F_{t_k})dB^H_s\right).$$

The first term on the right-hand side can be expressed as

$$\frac{\partial f}{\partial x}(t_k,\eta_{t_k})\left(F_{t_k}\diamond\left(B^H_{t_{k+1}}-B^H_{t_k}\right)\right)$$
$$=\frac{\partial f}{\partial x}(t_k,\eta_{t_k})\left(F_{t_k}\left(B^H_{t_{k+1}}-B^H_{t_k}\right)-\int_{t_k}^{t_{k+1}}D^\phi_s F_{t_k}\,ds\right)$$
$$=\frac{\partial f}{\partial x}(t_k,\eta_{t_k})F_{t_k}\left(B^H_{t_{k+1}}-B^H_{t_k}\right)-\frac{\partial f}{\partial x}(t_k,\eta_{t_k})\int_{t_k}^{t_{k+1}}D^\phi_s F_{t_k}\,ds$$
$$=\left[\frac{\partial f}{\partial x}(t_k,\eta_{t_k})F_{t_k}\right]\diamond\left(B^H_{t_{k+1}}-B^H_{t_k}\right)+\int_{t_k}^{t_{k+1}}D^\phi_s\left(\frac{\partial f}{\partial x}(t_k,\eta_{t_k})F_{t_k}\right)ds$$
$$-\frac{\partial f}{\partial x}(t_k,\eta_{t_k})\int_{t_k}^{t_{k+1}}D^\phi_s F_{t_k}\,ds$$
$$=\left[\frac{\partial f}{\partial x}(t_k,\eta_{t_k})F_{t_k}\right]\diamond\left(B^H_{t_{k+1}}-B^H_{t_k}\right)+\int_{t_k}^{t_{k+1}}F_{t_k}D^\phi_s\frac{\partial f}{\partial x}(t_k,\eta_{t_k})ds.$$

Thus,

$$\sum_{k=0}^{n-1}\frac{\partial f}{\partial x}(t_k,\eta_{t_k})\left(\eta_{t_{k+1}}-\eta_{t_k}\right)=\sum_{k=0}^{n-1}\left[\frac{\partial f}{\partial x}(t_k,\eta_{t_k})F_{t_k}\right]\diamond\left(B^H_{t_{k+1}}-B^H_{t_k}\right)$$
$$+\sum_{k=0}^{n-1}\int_{t_k}^{t_{k+1}}F_{t_k}D^\phi_s\frac{\partial f}{\partial x}(t_k,\eta_{t_k})ds.$$

As $|\pi|\to 0$, the first term converges to

$$\int_0^t F_s\frac{\partial f}{\partial x}(s,\eta_s)dB^H_s$$

in L^2, and the second term converges to

$$\int_0^t\frac{\partial^2 f}{\partial x^2}(s,\eta_s)D^\phi_s\eta_s F_s\,ds$$

in L^2. To prove the theorem, it is only necessary to show that as $|\pi|\to 0$,

$$\sum_{k=0}^{n-1}\mathbb{E}|\frac{\partial f}{\partial x}(t_k,\eta_{t_k})\int_{t_k}^{t_{k+1}}(F_s-F_{t_k})dB^H_s|\to 0.$$

Since f has a bounded second derivative, it follows that

$$\left|\frac{\partial f}{\partial x}(t_k,\eta_{t_k})\right| \le C(1+|\eta_{t_k}|).$$

Thus

$$\mathbb{E}|\frac{\partial f}{\partial x}(t_k,\eta_{t_k})|^2 \le C.$$

Furthermore,

$$\begin{aligned}
&\sum_{k=0}^{n-1}\mathbb{E}\left|\frac{\partial f}{\partial x}(t_k,\eta_{t_k})\int_{t_k}^{t_{k+1}}(F_s-F_{t_k})\,dB_s^H\right| \\
&\quad\le C\sum_{k=0}^{n-1}\left\{\mathbb{E}\left|\int_{t_k}^{t_{k+1}}(F_s-F_{t_k})\,dB_s^H\right|^2\right\}^{1/2} \\
&\quad= C\sum_{k=0}^{n-1}\left\{\mathbb{E}\left(\int_{t_k}^{t_{k+1}}\big(D_s^\phi(F_s-F_{t_k})\big)ds\right)^2\right. \\
&\qquad\qquad\left.+\,\mathbb{E}\int_{t_k}^{t_{k+1}}\int_{t_k}^{t_{k+1}}(F_u-F_{t_k})(F_v-F_{t_k})\phi(u,v)dudv\right\}^{1/2} \\
&\quad\le C\sum_{k=0}^{n-1}\left\{(t_{k+1}-t_k)\int_{t_k}^{t_{k+1}}\mathbb{E}\big(D_s^\phi(F_s-F_{t_k})\big)^2\,ds\right. \\
&\qquad\qquad\left.+\int_{t_k}^{t_{k+1}}\int_{t_k}^{t_{k+1}}\{\mathbb{E}(F_u-F_{t_k})^2\}^{1/2}\{\mathbb{E}(F_v-F_{t_k})^2\}^{1/2}\phi(u,v)dudv\right\}^{1/2} \\
&\quad\le C\sum_{k=0}^{n-1}\left[\sup_{t_k\le s\le t_{k+1}}\mathbb{E}|D_s^\phi(F_s-F_{t_k})|^2(t_{k+1}-t_k)^2\right. \\
&\qquad\qquad\left.+(t_{k+1}-t_k)^{2H}\left\{\sup_{t_k\le s\le t_{k+1}}\mathbb{E}(F_s-F_{t_k})^2\right\}\right]^{1/2} \\
&\quad\le C\left\{\sup_{t_k\le s\le t_{k+1}}\mathbb{E}|D_s^\phi(F_s-F_{t_k})|^2\right\}^{1/2}+C|\pi|^{H+\alpha-1}\to 0
\end{aligned}$$

as $|\pi|\to 0$. This proves the theorem. □

The equality (4.2) can be formally expressed as

$$df(t,\eta_t)=\frac{\partial f}{\partial t}(t,\eta_t)dt+\frac{\partial f}{\partial x}(t,\eta_t)F_t\,dB_t^H+\frac{\partial^2 f}{\partial x^2}(t,\eta_t)F_tD_t^\phi\eta_t\,dt.$$

If $F(s)=a(s)$ is a deterministic function, then (4.1) simplifies as follows.

COROLLARY 4.4. *Let $\eta_t=\int_0^t a_u\,dB_u^H$, where $a\in L_\phi^2$ and $f:\mathbb{R}_+\times\mathbb{R}\to\mathbb{R}$ satisfies the conditions in Theorem* 4.3. *Let $(\frac{\partial f}{\partial x}(s,\eta_s)a_s, s\in[0,T])$ be in $\mathcal{L}(0,T)$. Then*

$$\begin{aligned}
f(t,\eta_t)=f(0,0)&+\int_0^t\frac{\partial f}{\partial s}(s,\eta_s)ds+\int_0^t\frac{\partial f}{\partial x}(s,\eta_s)a_s\,dB_s^H \\
&+\int_0^t\frac{\partial^2 f}{\partial x^2}(s,\eta_s)\int_0^s\phi(s,v)a_v\,dvds\quad a.s.,
\end{aligned}$$

or formally,

$$df(t,\eta_t) = \frac{\partial f}{\partial t}(t,\eta_t)dt + \frac{\partial f}{\partial x}(t,\eta_t)a_t\,dB_t^H + \frac{\partial^2 f}{\partial x^2}(t,\eta_t)\int_0^t \phi(t,v)a_v\,dvdt.$$

If $a_s \equiv 1$, then Theorem 4.1 is obtained.

In the classical stochastic analysis, the stochastic integral can be defined for general semimartingales and an Itô formula can be given. By the Doob–Meyer decomposition [4], a semimartingale can be expressed as the sum of a martingale and a bounded variation process. A semimartingale $(X_t, t \geq 0)$ with respect to a Brownian motion can often be expressed as $X_t = X_0 + \int_0^t f_s\,dB_s + \int_0^t g_s\,ds$. An Itô formula in the analogous form with respect to fractional Brownian motions is given. This generalization of the Itô formula is useful in applications.

THEOREM 4.5. *Let $(F_u, u \in [0,T])$ satisfy the conditions of Theorem* 4.3*, and let $\mathbb{E}\sup_{0\leq s\leq T}|G_s| < \infty$. Denote $\eta_t = \xi + \int_0^t G_u\,du + \int_0^t F_u\,dB_u^H$, $\xi \in \mathbb{R}$ for $t \in [0,T]$. Let $(\frac{\partial f}{\partial x}(s,\eta_s)F_s, s \in [0,T]) \in \mathcal{L}(0,T)$. Then, for $t \in [0,T]$,*

$$\begin{aligned} f(t,\eta_t) = f(0,\xi) &+ \int_0^t \frac{\partial f}{\partial s}(s,\eta_s)ds + \int_0^t \frac{\partial f}{\partial x}(s,\eta_s)G_s\,ds \\ &+ \int_0^t \frac{\partial f}{\partial x}(s,\eta_s)F_s\,dB_s^H + \int_0^t \frac{\partial^2 f}{\partial x^2}(s,\eta_s)F_s D_s^\phi \eta_s\,ds \quad a.s. \end{aligned}$$

The proof is the same as for the above theorem.

Now the Itô formula for $\mathbb{R}^n$-valued processes is given.

THEOREM 4.6. *Let $(F_s^i, i = 1,\dots,n, s \in [0,T])$ satisfy the conditions of Theorem* 4.3 *for F. Let $\xi_t^k = \int_0^t F_s^k\,dB_s^H$, $k = 1,2,\dots,n$ for $t \in [0,T]$. For $k = 1,2,\dots,n$ let $(f_{x_k}(s,\xi_s^1,\dots,\xi_s^n)F_s^k,\ s \in [0,T])$ be in $\mathcal{L}(0,T)$. Let $f : \mathbb{R}_+ \times \mathbb{R}^n \to \mathbb{R}$ be twice continuously differentiable with bounded derivatives to second order. Then*

$$\begin{aligned} f(t,\xi_t^1,\dots,\xi_t^n) = f(0,0,\dots,0) &+ \int_0^t f_s(s,\xi_s^1,\dots,\xi_s^n)ds \\ &+ \sum_{k=1}^n \int_0^t f_{x_k}(s,\xi_s^1,\dots,\xi_s^n)F_s^k\,dB_s^H \\ &+ \sum_{k,l=1}^n \int_0^t f_{x_k x_l}(s,\xi_s^1,\dots,\xi_s^n)F_s^k\ D_s^\phi \xi_s^l\,ds \quad a.s. \end{aligned}$$

Proof. The theorem is verified for $n = 2$. To simplify the notation, let $F^1 = F$ and $F^2 = G$. Let $\pi : 0 = t_0 < t_1 < t_2 < \cdots < t_n = t$ be a partition of the interval $[0,t]$. Then

$$\begin{aligned} f(t,\xi_t,\eta_t) - f(0,0,0) = &\sum_{k=0}^{n-1} \big(f(t_{k+1},\xi_{t_{k+1}},\eta_{t_{k+1}}) - f(t_k,\xi_{t_{k+1}},\eta_{t_{k+1}})\big) \\ &+ \sum_{k=0}^{n-1} \big(f(t_k,\xi_{t_{k+1}},\eta_{t_{k+1}}) - f(t_k,\xi_{t_k},\eta_{t_k})\big), \end{aligned} \tag{4.3}$$

where $\xi_t = \int_0^t F\,dB^H$ and $\eta_t = \int_0^t G\,dB^H$. It is easy to see that the first sum converges to $\int_0^t \frac{\partial f}{\partial t}ds$ in L^2 as $|\pi| \to 0$. To determine the limit of the second sum, consider each

term in the sum:

$$\begin{aligned}
&f(t_k,\xi_{t_{k+1}},\eta_{t_{k+1}}) - f(t_k,\xi_{t_k},\eta_{t_k}) \\
&\quad = f_x(t_k,\xi_{t_k},\eta_{t_k})\big(\xi_{t_{k+1}}-\xi_{t_k}\big) + f_y(t_k,\xi_{t_k},\eta_{t_k})\big(\eta_{t_{k+1}}-\eta_{t_k}\big) \\
&\qquad + \frac{1}{2}f_{xx}(t_k,\tilde{\xi}_{t_k},\tilde{\eta}_{t_k})\big(\xi_{t_{k+1}}-\xi_{t_k}\big)^2 + \frac{1}{2}f_{yy}(t_k,\tilde{\xi}_{t_k},\tilde{\eta}_{t_k})\left(\eta_{t_{k+1}}-\eta_{t_k}\right)^2 \\
&\qquad + f_{xy}(t_k,\tilde{\xi}_{t_k},\tilde{\eta}_{t_k})\left(\xi_{t_{k+1}}-\xi_{t_k}\right)\left(\eta_{t_{k+1}}-\eta_{t_k}\right),
\end{aligned}$$

where $\tilde{\xi}_{t_k} \in (\xi_{t_k},\xi_{t_{k+1}})$ and $\tilde{\eta}_{t_k} \in (\eta_{t_k},\eta_{t_{k+1}})$. Thus

$$\begin{aligned}
&\sum_{k=0}^{n-1}\big(f(t_k,\xi_{t_{k+1}},\eta_{t_{k+1}}) - f(t_k,\xi_{t_k},\eta_{t_k})\big) \\
&\quad = \sum_{k=0}^{n-1} f_x(t_k,\xi_{t_k},\eta_{t_k})\left(\xi_{t_{k+1}}-\xi_{t_k}\right) + \sum_{k=0}^{n-1} f_y(t_k,\xi_{t_k},\eta_{t_k})\big(\eta_{t_{k+1}}-\eta_{t_k}\big) \\
&\qquad + \frac{1}{2}\sum_{k=0}^{n-1} f_{xx}(t_k,\tilde{\xi}_{t_k},\tilde{\eta}_{t_k})\left(\xi_{t_{k+1}}-\xi_{t_k}\right)^2 \\
&\qquad + \frac{1}{2}\sum_{k=0}^{n-1} f_{yy}(t_k,\tilde{\xi}_{t_k},\tilde{\eta}_{t_k})\left(\eta_{t_{k+1}}-\eta_{t_k}\right)^2 \\
&\qquad + \sum_{k=0}^{n-1} f_{xy}(t_k,\tilde{\xi}_{t_k},\tilde{\eta}_{t_k})\left(\xi_{t_{k+1}}-\xi_{t_k}\right)\left(\eta_{t_{k+1}}-\eta_{t_k}\right) \\
&\quad = I_1^\pi + I_2^\pi + I_3^\pi + I_4^\pi + I_5^\pi.
\end{aligned}$$

In a similar way as the proof of Theorem 4.3, it can be shown that as $|\pi| \to 0$, I_k^π converges to 0 in L^2 for $k = 3,4,5$:

$$\begin{aligned}
f_x(t_k,\xi_{t_k},\eta_{t_k})\left(\xi_{t_{k+1}}-\xi_{t_k}\right) &= f_x(t_k,\xi_{t_k},\eta_{t_k})\left(F_{t_k} \diamond \left(B^H_{t_{k+1}} - B^H_{t_k}\right)\right) \\
&\quad + f_x(t_k,\xi_{t_k},\eta_{t_k})\left(\int_{t_k}^{t_{k+1}} (F_s - F_{t_k}) dB^H_s\right).
\end{aligned}$$

In a similar way to the proof of Theorem 4.3, the second sum does not contribute in the limit as $|\pi| \to 0$.

By the definition of the Wick product, it follows that

$$\begin{aligned}
&f_x(t_k,\xi_{t_k},\eta_{t_k})\left(F_{t_k} \diamond \left(B^H_{t_{k+1}} - B^H_{t_k}\right)\right) \\
&\quad = f_x(t_k,\xi_{t_k},\eta_{t_k})\left(F_{t_k}\left(B^H_{t_{k+1}} - B^H_{t_k}\right) - \int_{t_k}^{t_{k+1}} D^\phi_s F_{t_k}\, ds\right) \\
&\quad = f_x(t_k,\xi_{t_k},\eta_{t_k}) F_{t_k}\left(B^H_{t_{k+1}} - B^H_{t_k}\right) - f_x(t_k,\xi_{t_k},\eta_{t_k})\int_{t_k}^{t_{k+1}} D^\phi_s F_{t_k}\, ds \\
&\quad = (f_x(t_k,\xi_{t_k},\eta_{t_k})F_{t_k}) \diamond \left(B^H_{t_{k+1}} - B^H_{t_k}\right) + \int_{t_k}^{t_{k+1}} D^\phi_s (f_x(t_k,\xi_{t_k},\eta_{t_k})F_{t_k}) ds \\
&\qquad - f_x(t_k,\xi_{t_k},\eta_{t_k})\int_{t_k}^{t_{k+1}} D^\phi_s F_{t_k}\, ds \\
&\quad = (f_x(t_k,\xi_{t_k},\eta_{t_k})F_{t_k}) \diamond \left(B^H_{t_{k+1}} - B^H_{t_k}\right) + \int_{t_k}^{t_{k+1}} f_{xx}(t_k,\xi_{t_k},\eta_{t_k}) D^\phi_s \xi_{t_k} F_{t_k}\, ds
\end{aligned}$$

$$
\begin{aligned}
&+\int_{t_k}^{t_{k+1}} f_{xy}(t_k,\xi_{t_k},\eta_{t_k})D_s^\phi \eta_{t_k}F_{t_k}\,ds+\int_{t_k}^{t_{k+1}} f_x(t_k,\xi_{t_k},\eta_{t_k})D_s^\phi F_{t_k}\,ds\\
&-f_x(t_k,\xi_{t_k},\eta_{t_k})\int_{t_k}^{t_{k+1}} D_s^\phi F_{t_k}\,ds\\
&=\left(f_x\big(t_k,\xi_{t_k},\eta_{t_k}\big)F_{t_k}\right)\diamond\left(B_{t_{k+1}}^H-B_{t_k}^H\right)+\int_{t_k}^{t_{k+1}} f_{xx}(t_k,\xi_{t_k},\eta_{t_k})D_s^\phi \xi_{t_k}F_{t_k}\,ds\\
&+\int_{t_k}^{t_{k+1}} f_{xy}(t_k,\xi_{t_k},\eta_{t_k})D_s^\phi \eta_{t_k}F_{t_k}\,ds.
\end{aligned}
$$

In a similar way to the proof of Theorem 4.3, it can be shown that as $|\pi|\to 0$,

$$
\begin{aligned}
I_1^\pi \to &\int_0^t \frac{\partial f}{\partial x}(s,\xi_s,\eta_s)F_s\,dB_s^H+\int_0^t \frac{\partial^2 f}{\partial x^2}(s,\xi_s,\eta_s)D_s^\phi\xi_s F_s\,ds\\
&+\int_0^t \frac{\partial^2 f}{\partial x\partial y}(s,\xi_s,\eta_s)D_s^\phi\eta_s F_s\,ds
\end{aligned}
$$

and

$$
\begin{aligned}
I_2^\pi \to &\int_0^t \frac{\partial f}{\partial y}(s,\xi_s,\eta_s)G_s\,dB_s^H+\int_0^t \frac{\partial^2 f}{\partial y^2}(s,\xi_s,\eta_s)D_s^\phi\eta_s G_s\,ds\\
&+\int_0^t \frac{\partial^2 f}{\partial x\partial y}(s,\xi_s,\eta_s)D_s^\phi\xi_s G_s\,ds
\end{aligned}
$$

in L^2, proving the theorem. □

The Itô formula for the integrals of Stratonovich type is simpler.

THEOREM 4.7. *Let $(F_t, t\in[0,T])$ be a process such that the assumptions of Theorem 4.3 are satisfied. Let $\xi_t=\int_0^t F_s\delta B_s^H$. Let $g:\mathbb{R}_+\times\mathbb{R}\to\mathbb{R}$ be a twice continuously differentiable function with bounded derivatives to second order. Let $(\frac{\partial g}{\partial x}(s,\xi_s)F_s, s\in[0,T])$ be in $\mathcal{L}(0,T)$. Then, for $t\in[0,T]$,*

$$
g(t,\xi_t)=g(0,0)+\int_0^t \frac{\partial g}{\partial s}(s,\xi_s)ds+\int_0^t \frac{\partial g}{\partial x}(s,\xi_s)F_s\delta B_s^H \quad a.s. \tag{4.4}
$$

Proof. Note that $\tilde{\xi}_t=\int_0^t F_s\,dB_s^H$ also exists and

$$
\xi_t=\tilde{\xi}_t+\int_0^t D_s^\phi F_s\,ds.
$$

Using the Itô formula (4.2),

$$
\begin{aligned}
g(t,\xi_t)&=g\left(t,\tilde{\xi}_t+\int_0^t D_s^\phi F_s\,ds\right)\\
&=g(0,0)+\int_0^t \frac{\partial g}{\partial s}(s,\xi_s)ds+\int_0^t \frac{\partial g}{\partial x}(s,\xi_s)D_s^\phi F_s\,ds\\
&\quad+\int_0^t \frac{\partial g}{\partial x}(s,\xi_s)F_s\,dB_s^H+\int_0^t \frac{\partial^2 g}{\partial x^2}(s,\xi_s)D_s^\phi\xi_s F_s\,ds. \qquad (4.5)
\end{aligned}
$$

Now,

$$\begin{aligned}\int_0^t \frac{\partial g}{\partial x}(s,\xi_s)F_s\,dB_s^H &= \int_0^t g_x(s,\xi_s)F_s\delta B_s^H - \int_0^t D_s^\phi(g_x(s,\xi_s)F_s)ds\\ &= \int_0^t g_x(s,\xi_s)F_s\delta B_s^H - \int_0^t g_{xx}(s,\xi_s)D_s^\phi\xi_s F_s\,ds\\ &\quad - \int_0^t g_x(s,\xi_s)D_s^\phi F_s\,ds. \qquad (4.6)\end{aligned}$$

Combining the above two equalities, it follows that

$$g(t,\xi_t) = g(0,0) + \int_0^t \frac{\partial g}{\partial s}(s,\xi_s)ds + \int_0^t \frac{\partial g}{\partial x}(s,\xi_s)F_s\delta B_s^H,$$

proving the theorem. □

REMARK 1. *The equation* (4.4) *can be expressed formally as*

$$\delta g(t,\xi_t) = g_t(t,\xi_t)dt + g_x(t,\xi_t)\delta\xi_t$$

or more generally as

$$\begin{aligned}\delta g(t,\xi_t^1,\xi_t^2,\dots,\xi_t^n) &= \frac{\partial g}{\partial t}(t,\xi_t^1,\xi_t^2,\dots,\xi_t^n)dt + \frac{\partial g}{\partial x_1}(t,\xi_t^1,\xi_t^2,\dots,\xi_t^n)\delta\xi_t^1\\ &\quad + \cdots + \frac{\partial g}{\partial x_n}(t,\xi_t^1,\xi_t^2,\dots,\xi_t^n)\delta\xi_t^n.\end{aligned}$$

5. Two applications of the Itô formula. Two applications of the Itô formula for fractional Brownian motion are given. First, the so-called homogeneous chaos is extended to a fractional Brownian motion. Second, an L^p estimate of the (Itô type) stochastic integral for a fractional Brownian motion is given.

Let $H_n(x)$ be the Hermite polynomial of degree n, that is,

$$e^{tx-\frac{1}{2}t^2} = \sum_{n=0}^{\infty} t^n H_n(x).$$

Let

$$|f|_{\phi,t} = \left\{\int_0^t\int_0^t \phi(u,v)f_u f_v\,dudv\right\}^{1/2}.$$

Define

$$\tilde{f}_t = |f|_{\phi,t}^{-1}\int_0^t f_s\,dB_s^H$$

and

$$H_n^{\phi,f}(t) = |f|_{\phi,t}^n H_n\left(\tilde{f}_t\right).$$

THEOREM 5.1. *If* $f1_{[0,T]} \in L_\phi^2$, *then the following equality is satisfied:*

$$dH_n^{\phi,f}(t) = nH_{n-1}^{\phi,f}(t)f_t\,dB_t^H,$$

where d is the Itô type differential given in Theorem 4.3 *and* $t \in [0,T]$.

Proof. Fix n and denote $X_t = H_n^{\phi,f}(t)$ for $t \in [0,T]$. Using the Itô formula (Theorem 4.3) and prime for differentiation, it follows that

$$\begin{aligned}
dX_t &= n|f|_{\phi,t}^{n-2} f_t \int_0^t \phi(u,t) f_u \, du H_n\Big(\tilde{f}_t\Big) dt \\
&\quad - |f|_{\phi,t}^{n} f_t \int_0^t \phi(u,t) f_u \, du H_n'\Big(\tilde{f}_t\Big) |f|_{\phi,t}^{-3} \left(\int_0^t f_s \, dB_s^H \right) dt \\
&\quad + |f|_{\phi,t}^{n} H_n'\Big(\tilde{f}_t\Big) |f|_{\phi,t}^{-1} f_t \, dB_t^H \\
&\quad + |f|_{\phi,t}^{n} f_t \int_0^t \phi(u,t) f_u \, du H_n''\Big(\tilde{f}_t\Big) |f|_{\phi,t}^{-2} \, dt \\
&= n|f|_{\phi,t}^{n-1} H_{n-1}\Big(\tilde{f}_t\Big) f_t \, dB_t^H \\
&\quad + |f|_{\phi,t}^{n-2} f_t \int_0^t \phi(u,t) f_u du \cdot \Big\{ nH_n(\tilde{f}_t) - \tilde{f}_t H_n'(\tilde{f}_t) + H_n''(\tilde{f}_t) \Big\} \, dt.
\end{aligned}$$

It is well known that for each $n \in \mathbb{N}$ the Hermite polynomial satisfies

$$nH_n(x) - xH_n'(x) + H_n''(x) = 0$$

for each $x \in \mathbb{R}$. Thus the sum of the terms in the above { } equals 0. The first term is

$$nH_{n-1}^{\phi,f}(t) f_t \, dB_t^H.$$

Thus

$$dH_n^{\phi,f}(t) = nH_{n-1}^{\phi,f}(t) f_t \, dB_t^H,$$

proving the theorem. □

The following estimate for the L^p norm of a stochastic integral is useful in some applications.

THEOREM 5.2. *Let $(g_s, s \in [0,t])$ be a stochastic process satisfying the assumptions of Theorem* 4.3 *for F. Let $F_t := \int_0^t g_s \, dB_s^H$. If $\mathbb{E}\int_0^t |g_s|^p \, ds < \infty$, $\int_0^t \mathbb{E}|D_s^\phi F_s|^p \, ds < \infty$, and $F^{p-1} g \in \mathcal{L}(0,t)$, then*

$$\mathbb{E}F_t^p \le p^p \left\{ \int_0^t \Big(\mathbb{E}|g_s D_s^\phi F_s|^{p/2} \Big)^{2/p} ds \right\}^{p/2}. \tag{5.1}$$

Proof. Applying the Itô formula (Theorem 4.3) to F_t^p (by the assumption that $F^{p-1} g \in \mathcal{L}(0,t)$, the restriction on the boundedness of f to its second derivative in Theorem 4.3 can be removed), it follows that

$$F_t^p = p \int_0^t F_s^{p-1} g_s \, dB_s^H + p(p-1) \int_0^t F_s^{p-2} g_s D_s^\phi F_s \, ds.$$

Thus

$$\begin{aligned}
\mathbb{E}F_t^p &= p(p-1) \int_0^t \mathbb{E}(F_s^{p-2} g_s D_s^\phi F_s) ds. \\
\mathbb{E}F_t^p &\le p(p-1) \int_0^t \mathbb{E}|F_s^{p-2} g_s D_s^\phi F_s| ds \\
&\le p^2 \int_0^t (\mathbb{E}F_s^p)^{\frac{p-2}{p}} \Big(\mathbb{E}|g_s D_s^\phi F_s|^{\frac{p}{2}} \Big)^{\frac{2}{p}} ds.
\end{aligned}$$

By an inequality of Langenhop, (e.g. [1]), there is the inequality

$$\mathbb{E}F_t^p \le p^p \left\{ \int_0^t \left(\mathbb{E}|g_s D_s^\phi F_s|^{\frac{p}{2}} \right)^{\frac{2}{p}} ds \right\}^{\frac{p}{2}}.$$

This completes the proof of the theorem. □

COROLLARY 5.3. *Let the conditions of Theorem* 5.2 *be satisfied, and let* $p \ge 2$. *Then*

$$\mathbb{E}F_t^p \le p^p \left\{ \int_0^t (\mathbb{E}|g_s|^p)^{\frac{2}{p}} ds + \int_0^t (\mathbb{E}|D_s^\phi F_s|^p)^{\frac{2}{p}} ds \right\}^{\frac{p}{2}}.$$

Proof. From $|ab| \le a^2 + b^2$, it follows that

$$\mathbb{E}|g_s D_s^\phi F_s|^{\frac{p}{2}} \le \mathbb{E}|g_s|^p + \mathbb{E}|D_s^\phi F_s|^p.$$

Thus

$$\begin{aligned} \left(\mathbb{E}|g_s D_s^\phi F_s|^{\frac{p}{2}} \right)^{\frac{2}{p}} &\le (\mathbb{E}|g_s|^p + \mathbb{E}|D_s^\phi F_s|^p)^{\frac{2}{p}} \\ &\le (\mathbb{E}|g_s|^p)^{\frac{2}{p}} + (\mathbb{E}|D_s^\phi F_s|^p)^{\frac{2}{p}}. \end{aligned}$$

This verifies the corollary. □

6. Iterated integrals and multiple integrals. Let $f \in L_\phi^2(\mathbb{R}_+)$ be such that $|f|_\phi = 1$. Similar to [8] define $(\int_0^\infty f_s\, dB_s^H)^{\diamond n}$ as the nth Wick power of $\int_0^\infty f_s\, dB_s^H$, that is, denote formally

$$\begin{aligned} \left(\int_0^\infty f_s\, dB_s^H \right)^{\diamond n} &:= \left(\int_0^\infty f_s\, dB_s^H \right)^{\diamond(n-1)} \diamond \int_0^\infty f_s\, dB_s^H, \quad n = 2, 3, \dots \\ \exp^\diamond \left(\int_0^\infty f_s\, dB_s^H \right) &:= \sum_{n=0}^\infty \frac{1}{n!} \left(\int_0^\infty f_s\, dB_s^H \right)^{\diamond n}; \\ \log^\diamond \left(1 + \int_0^\infty f_s\, dB_s^H \right) &:= \sum_{n=1}^\infty \frac{(-1)^{n-1}}{n} \left(\int_0^\infty f_s\, dB_s^H \right)^{\diamond n}. \end{aligned}$$

LEMMA 6.1. *If* $|f|_\phi = 1$, *then* $(\int_0^\infty f_s\, dB_s^H)^{\diamond n}$ *is well defined for each* $n \in \mathbb{N}$ *and*

$$\left(\int_0^\infty f_s\, dB_s^H \right)^{\diamond n} = H_n \left(\int_0^\infty f_s\, dB_s^H \right), \tag{6.1}$$

where H_n *denotes the Hermite polynomial of degree* n.

Proof. The equality (6.1) is verified by induction.

It is easy to see that (6.1) is true for $n = 1$. Let (6.1) be true for $1, 2, \dots, n-1$.

Then,

$$\begin{aligned}\left(\int_0^\infty f_s\,dB_s^H\right)^{\diamond n} &= H_{n-1}\left(\int_0^\infty f_s\,dB_s^H\right)\diamond\int_0^\infty f_s\,dB_s^H\\ &= H_{n-1}\left(\int_0^\infty f_s\,dB_s^H\right)\int_0^\infty f_s\,dB_s^H - D_{\Phi f}\left\{H_{n-1}\left(\int_0^\infty f_s\,dB_s^H\right)\right\}\\ &= H_{n-1}\left(\int_0^\infty f_s\,dB_s^H\right)\int_0^\infty f_s dB_s^H - H'_{n-1}\left(\int_0^\infty f_s\,dB_s^H\right)|f|_\phi^2\\ &= H_{n-1}\left(\int_0^\infty f_s\,dB_s^H\right)\int_0^\infty f_s\,dB_s^H - H'_{n-1}\left(\int_0^\infty f_s\,dB_s^H\right)\\ &= H_n\left(\int_0^\infty f_s\,dB_s^H\right)\end{aligned}$$

by an identity for Hermite polynomials. This verifies (6.1). ∎

For an arbitrary, nonzero $f\in L_\phi^2(\mathbb{R}_+)$, the product defined in (6.1) is extended as

$$\left(\int_0^\infty f_s\,dB_s^H\right)^{\diamond n} = |f|_\phi^n\left(\frac{\int_0^\infty f_s\,dB_s^H}{|f|_\phi}\right)^{\diamond n} = |f|_\phi^n H_n\left(\int_0^\infty \frac{\int_0^\infty f_s\,dB_s^H}{|f|_\phi}\right).$$

LEMMA 6.2. *If $f\in L_\phi^2(\mathbb{R}_+)$, then $(\int_0^\infty f_s\,dB_s^H)^{\diamond n}$ is well defined for each $n\in\mathbb{N}$ and*

$$\left(\int_0^t f_s\,dB_s^H\right)^{\diamond n} = H_n^{\phi,f}(t).$$

Since $\int_0^\infty f_s\,dB_s^H$ is a Gaussian random variable, it is easy to estimate its moments and to show that the series defining $\exp^\diamond(\int_0^\infty f_s\,dB_s^H)$ is convergent in $L^2(\Omega,\mathcal{F},P)$. Moreover there is the following corollary.

COROLLARY 6.3. *If $f\in L_\phi^2(\mathbb{R}_+)$, then*

$$\exp^\diamond\left(\int_0^\infty f_s\,dB_s^H\right) = \varepsilon(f) = \exp\left(\int_0^\infty f_s\,dB_s^H - \frac{1}{2}|f|_\phi^2\right).$$

Proof. It follows that

$$\begin{aligned}\exp^\diamond\left(\int_0^\infty f_s\,dB_s^H\right) &= \sum_{n=0}^\infty \frac{1}{n!}\left(\int_0^\infty f_s\,dB_s^H\right)^{\diamond n}\\ &= \sum_{n=0}^\infty \frac{1}{n!}|f|_\phi^n H_n\left(\frac{\int_0^\infty f_s\,dB_s^H}{|f|_\phi}\right)\\ &= \exp\left(|f|_\phi\frac{\int_0^\infty f_s\,dB_s^H}{|f|_\phi} - \frac{1}{2}|f|_\phi^2\right)\\ &= \exp\left(\int_0^\infty f_s\,dB_s^H - \frac{1}{2}|f|_\phi^2\right).\end{aligned}$$

This completes the proof of the lemma. ∎

The following lemma is also easy to prove.

LEMMA 6.4. *For any two functions f and g in $L^2_\phi(\mathbb{R}_+)$ with $\langle f,g\rangle_\phi = 0$, the following equality is satisfied:*

$$\left(\int_0^\infty f_s\,dB_s^H\right)^{\diamond n}\diamond\left(\int_0^\infty g_s\,dB_s^H\right)^{\diamond m} = \left(\int_0^\infty f_s\,dB_s^H\right)^{\diamond n}\left(\int_0^\infty g_s\,dB_s^H\right)^{\diamond m} = H_n^{\phi,f}(\infty)H_m^{\phi,g}(\infty).$$

Since $\int_0^\infty f_s\,dB_s^H/|f|_\phi$ and $\int_0^\infty g_s\,dB_s^H/|g|_\phi$ are Gaussian random variables with mean 0 and variance 1, their covariance is

$$\mathbb{E}\left\{\left(\int_0^\infty f_s\,dB_s^H/|f|_\phi\right)\left(\int_0^\infty g_s\,dB_s^H/|g|_\phi\right)\right\} = \langle f/|f|_\phi, g/|g|_\phi\rangle_\phi.$$

It follows that

$$\begin{aligned}\mathbb{E}&\left\{\left(\int_0^\infty f_s\,dB_s^H\right)^{\diamond n}\left(\int_0^\infty g_s\,dB_s^H\right)^{\diamond m}\right\}\\ &= \mathbb{E}\left\{|f|_\phi^n|g|_\phi^m H_n\left(\int_0^\infty f_s\,dB_s^H/|f|_\phi\right)H_m\left(\int_0^\infty g_s\,dB_s^H/|g|_\phi\right)\right\}\\ &= \begin{cases}0 & \text{if } m\neq n,\\ |f|_\phi^n|g|_\phi^n\langle f/|f|_\phi, g/|g|_\phi\rangle_\phi^n & \text{if } m=n\end{cases}\\ &= \begin{cases}0 & \text{if } m\neq n,\\ \langle f,g\rangle_\phi^n & \text{if } m=n.\end{cases}\end{aligned}$$

By a polarization technique [21] it is easy to verify the following lemma.

LEMMA 6.5. *Let $f^1,\dots,f^n$, $g^1,\dots,g^m\in L^2_\phi(\mathbb{R}_+)$. The following equality is satisfied:*

$$\mathbb{E}\left\{\left(\int_0^\infty f_s^1\,dB_s^H\diamond\cdots\diamond\int_0^\infty f_s^n\,dB_s^H\right)\left(\int_0^\infty g_s^1\,dB_s^H\diamond\cdots\diamond\int_0^\infty g_s^m\,dB_s^H\right)\right\} = \begin{cases}0 & \text{if } n\neq m,\\ \frac{1}{n!}\sum_\sigma\langle f^1,g^{\sigma(1)}\rangle_\phi\langle f^2,g^{\sigma(2)}\rangle_\phi\cdots\langle f^n,g^{\sigma(n)}\rangle_\phi & \text{if } n=m,\end{cases}$$

where $\sum_\sigma$ denotes the sum over all permutations σ of $\{1,2,\dots,n\}$.

Let $e_1,e_2,\dots,e_n,\dots$ be a complete orthonormal basis of $L^2_\phi(\mathbb{R}_+)$. Consider the nth symmetric tensor product of $L^2_\phi(\mathbb{R}_+)$: $L^s_\phi(\mathbb{R}_+^n) := L^2_\phi(\mathbb{R}_+)\otimes\cdots\otimes L^2_\phi(\mathbb{R}_+)$. It is the completion of all functions of the following form:

$$f(s_1,\dots,s_n) = \sum_{1\le k_1,\dots,k_n\le k} a_{k_1\cdots k_n}e_{k_1}(s_1)e_{k_2}(s_2)\cdots e_{k_n}(s_n), \tag{6.2}$$

where f is a symmetric function of its variables $s_1,\dots,s_n$ and k is a positive integer. The set of all of the above finite sums is denoted $\mathcal{L}_n$. For an element of the form (6.2), its multiple integral is defined by

$$I_n(f) = \sum_{1\le k_1,\dots,k_n\le k} a_{k_1\cdots k_n}\int_0^\infty e_{k_1}(s)dB_s^H\diamond\int_0^\infty e_{k_2}(s)dB_s^H\diamond\cdots\diamond\int_0^\infty e_{k_n}(s)dB_s^H. \tag{6.3}$$

By Lemma 6.5, the norm of (6.3) is given by

$$\mathbb{E}|I_n(f)|^2 = \int_{\mathbb{R}_+^{2n}} \phi(u_1,v_1)\phi(u_2,v_2)\cdots\phi(u_n,v_n)f(u_1,u_2,\dots,u_n) \tag{6.4}$$
$$f(v_1,v_2,\dots,v_n)du_1\,du_2\cdots du_n\,dv_1\,dv_2\cdots dv_n.$$

Thus for any element f in

$$L_\phi^2(\mathbb{R}_+^n) = \{f:\mathbb{R}_+^n\to\mathbb{R}; f \text{ is symmetric with respect to its arguments},$$
$$|f|_\phi^2 := \langle f,f\rangle_\phi < \infty\},$$

where

$$\langle f,g\rangle_\phi = \int_{\mathbb{R}_+^{2n}} \phi(u_1,v_1)\phi(u_2,v_2)\cdots\phi(u_n,v_n)f(u_1,u_2,\dots,u_n)$$
$$g(v_1,v_2,\dots,v_n)du_1\,du_2\cdots du_n\,dv_1\,dv_2\cdots dv_n.$$

The multiple integral $I_n(f)$ can be defined by a limit from elements in $\mathcal{L}_n$, and it follows that

$$\mathbb{E}(|I_n(f)|^2) = |f|_\phi^2.$$

The following lemma can also be shown by the polarization technique.

LEMMA 6.6. *If $f\in L_\phi^2(\mathbb{R}_+^n)$ and $g\in L_\phi^2(\mathbb{R}_+^m)$, then*

$$\mathbb{E}(I_n(f)I_m(g)) = \begin{cases}\langle f,g\rangle_\phi & \text{if } n=m,\\ 0 & \text{if } n\neq m.\end{cases} \tag{6.5}$$

Let $f\in L_\phi^2(\mathbb{R}_+^n)$. The iterated integral can be defined by the recursive formula

$$\int_{0\le s_1<s_2<\cdots<s_n\le t} f(s_1,s_2,\dots,s_n)dB_{s_1}^H\,dB_{s_2}^H\cdots dB_{s_n}^H$$
$$= \int_0^t \left(\int_{0\le s_1<s_2<\cdots\le s_n} f(s_1,s_2,\dots,s_{n-1},s_n)dB_{s_1}^H\,dB_{s_2}^H\cdots dB_{s_{n-1}}^H\right) dB_{s_n}^H. \tag{6.6}$$

THEOREM 6.7. *If $f\in L_\phi^2(\mathbb{R}_+^n)$, then the iterated integral* (6.6) *exists and*

$$I_n(f) = n!\int_{0\le s_1<s_2<\cdots<s_n\le t} f(s_1,s_2,\dots,s_n)dB_{s_1}^H\,dB_{s_2}^H\cdots dB_{s_n}^H. \tag{6.7}$$

Proof. First let f have the special form $f=g^{\otimes n}$, that is, $f(s_1,s_2,\dots,s_n) = g(s_1)g_2(s_2)\cdots g(s_n)$. Then

$$I_n(f) = H_n^{\phi,g}(t),$$

and

$$\begin{aligned} dI_n(f) &= dH_n^{\phi,g}(t)\\ &= nH_{n-1}^{\phi,g}(t)g(t)dB_t^H\\ &= nI_{n-1}(g^{\otimes(n-1)})g(t)dB_t^H. \end{aligned}$$

This verifies (6.7) for the case where $f = g^{\otimes n}$. By the polarization technique [21], the theorem follows easily. □

REMARK 2. *For Brownian motion, a multiple integral was originally introduced by Wiener* [24]*; Wiener's original multiple integral is in fact a multiple integral of Stratonovich type. The multiple integral of Itô type was introduced in* [16].

For Brownian motion, the multiple Stratonovich integrals also have been widely used in the applications. Since the work of [9] and [10], it is known that the definition of multiple Stratonovich integrals is related to the definition of "trace." There has been much work on this topic. The reader is referred to [13] and the references therein.

A class of *traces* and *multiple Stratonovich integrals* are defined, and the Hu–Meyer formula is extended to the fractional Brownian motions.

As in [11], introduce the *ϕ-trace* Tr_ϕ for simple functions. This new type of trace extends the classical one and plays an important role in this section.

Let $f_1, f_2, \dots, f_m \in L^2_\phi(\mathbb{R}_+)$. Consider the simple functions in $L^2_\phi(\mathbb{R}^n_+)$ of the following form:

$$f(t_1, t_2, \dots, t_n) = \sum_{1 \le i_1, i_2, \dots, i_n \le m} a_{i_1, i_2, \dots, i_n} f_{i_1}(t_1) f_{i_2}(t_2) \cdots f_{i_n}(t_n). \tag{6.8}$$

If f is given by (6.8), then for $k \in \{1, 2, \dots, [\frac{n}{2}]\}$, define

$$\mathrm{Tr}^k_\phi f(t_1, \dots, t_{n-2k}) = \int_0^\infty \cdots \int_0^\infty f(s_1, s_2, \dots, s_{2k-1}, s_{2k}, t_1, \dots, t_{n-2k}) \phi(s_1, s_2)\phi(s_3, s_4) \cdots \phi(s_{2k-1}, s_{2k}) ds_1 \cdots ds_{2k}.$$

To define the trace for general functions, as in [11], let $\gamma_\varepsilon(s, t)$ be an approximation of the Dirac function, that is,

$$\lim_{\varepsilon \to 0} \int \gamma_\varepsilon(s, t) f(s) ds = f(t)$$

in some sense. Assume that

$$\int_0^\infty \int_0^\infty \gamma_\varepsilon(s, t)^2 ds\, dt < \infty.$$

For any function $f \in L^2_\phi(\mathbb{R}^n_+)$, the approximation

$$f^\varepsilon(t_1, t_2, \dots, t_n) = \int_0^\infty \cdots \int_0^\infty f(s_1, s_2, \dots, s_n) \gamma_\varepsilon(s_1, t_1) \gamma_\varepsilon(s_2, t_2) \cdots \gamma_\varepsilon(s_n, t_n) ds_1\, ds_2 \cdots ds_n,$$

is a simple function of type (6.8), and if f is symmetric, then f^ε is also symmetric. Let

$$\rho_\varepsilon(s, t) = \int_0^\infty \gamma_\varepsilon(s, u) \gamma_\varepsilon(t, u) du.$$

According to the definition of f^ε,

$$\mathrm{Tr}^k_\phi f^\varepsilon(t_1, \dots, t_{n-2k}) = \int_0^\infty \cdots \int_0^\infty f(s_1, s_2, \dots, s_n) \rho(s_1, s_2) \cdots \rho(s_{2k-1}, s_{2k}) \gamma_\varepsilon(s_{2k+1}, t_1) \cdots \gamma_\varepsilon(s_n, t_{n-2k}) ds_1\, ds_2 \cdots ds_n.$$

DEFINITION 6.8. *Let $f \in L^2_\phi(\mathbb{R}^n_+)$. The kth trace of f is said to exist if $\mathrm{Tr}^k_\phi f^\varepsilon(t_1,\dots,t_{n-2k})$ converges to a function in $L^2_\phi(\mathbb{R}^{n-2k}_+)$ as $\varepsilon \to 0$. The limiting function is called the kth trace of f, that is,*

$$\mathrm{Tr}^k_\phi f(t_1,\dots,t_{n-2k}) = \lim_{\varepsilon\to 0} \mathrm{Tr}^k_\phi f^\varepsilon(t_1,\dots,t_{n-2k}).$$

Now introduce the multiple Stratonovich integrals for fractional Brownian motions. Define $(B^H_t)^\varepsilon = \int_0^\infty \gamma_\varepsilon(t,s)dB^H_s$. Then $(B^H_t)^\varepsilon$ is differentiable. Let $f \in L^2_\phi(\mathbb{R}^n_+)$. Consider

$$S^\varepsilon_n(f) := \int_{\mathbb{R}^n_+} f(s_1,s_2,\dots,s_n)(\dot B^H_{s_1})^\varepsilon(\dot B^H_{s_2})^\varepsilon\cdots(\dot B^H_{s_n})^\varepsilon ds_1\,ds_2\cdots ds_n.$$

DEFINITION 6.9. *If $S^\varepsilon_n(f)$ converges in $L^2(\Omega,\mathcal{F},P)$ as $\varepsilon\to 0$, then the multiple Stratonovich integral is said to exist and is denoted by*

$$S_n(f) = \int_{\mathbb{R}^n_+} f(s_1,s_2,\dots,s_n)\delta B^H_{s_1}\delta B^H_{s_2}\cdots\delta B^H_{s_n}.$$

The remaining part of this section is devoted to giving conditions such that $S^\varepsilon_n(f)$ is convergent in $L^2(\Omega,\mathcal{F},P)$ as $\varepsilon\to 0$.

By the identity $x^n = \sum_{k\le[\frac{n}{2}]} \frac{n!}{2^k k!(n-2k)!} H_{n-2k}(x)$, it follows that

$$\begin{aligned}
\left(\int_0^\infty f_s dB^H_s\right)^n &= |f|^n_\phi \left(\frac{\int_0^\infty f_s\,dB^H_s}{|f|_\phi}\right)^n \\
&= |f|^n_\phi \sum_{k\le[\frac{n}{2}]} \frac{n!}{2^k k!(n-2k)!} H_{n-2k}\left(\frac{\int_0^\infty f_s\,dB^H_s}{|f|_\phi}\right) \\
&= \sum_{k\le[\frac{n}{2}]} \frac{n!}{2^k k!(n-2k)!} |f|^n_\phi \left(\frac{\int_0^\infty f_s\,dB^H_s}{|f|_\phi}\right)^{\diamond(n-2k)} \\
&= \sum_{k\le[\frac{n}{2}]} \frac{n!}{2^k k!(n-2k)!} |f|^{2k}_\phi \left(\int_0^\infty f_s\,dB^H_s\right)^{\diamond(n-2k)} \\
&= \sum_{k\le[\frac{n}{2}]} \frac{n!}{2^k k!(n-2k)!} |f|^{2k}_\phi I_{n-2k}(f^{\otimes(n-2k)}) \\
&= \sum_{k\le[\frac{n}{2}]} \frac{n!}{2^k k!(n-2k)!} I_{n-2k}\left(\mathrm{Tr}^k_\phi(f^{\otimes n})\right),
\end{aligned}$$

where $f^{\otimes n}$ is the symmetric tensor product of f, that is, $f^{\otimes n}(s_1,s_2,\dots,s_n) = f(s_1)f(s_2)\cdots f(s_n)$.

Let $f_1,f_2,\dots,f_n \in L^2_\phi(\mathbb{R}_+)$, and let f be the symmetrization of $f_1f_2\cdots f_n$. Then by a polarization technique,

$$\int_0^\infty f_1(s)dB^H_s\int_0^\infty f_2(s)dB^H_s\cdots\int_0^\infty f_n(s)dB^H_s = \sum_{k\le[\frac{n}{2}]} \frac{n!}{2^k k!(n-2k)!} I_{n-2k}\left(\mathrm{Tr}^k_\phi(f)\right).$$

Using this formula, it follows that

$$S_n^\varepsilon(f) = \sum_{k\le[\frac{n}{2}]} \frac{n!}{2^k k!(n-2k)!} \int_{\mathbb{R}_+^n} f(s_1,s_2,\dots,s_n)\rho_\varepsilon(s_1,s_2)\rho_\varepsilon(s_3,s_4)\cdots\rho_\varepsilon(s_{2k-1},s_{2k})$$
$$\gamma_\varepsilon(s_{2k+1},t_1)\cdots\gamma_\varepsilon(s_n,t_{n-2k})ds_1\cdots ds_n\, dB_{t_1}^H\cdots dB_{t_{n-2k}}^H.$$

It is easy to verify the following result.

THEOREM 6.10. *Let $f \in L_\phi^2(\mathbb{R}_+^n)$ be such that all of the traces exist in the following sense: For $1 \le k \le [\frac{n}{2}]$,*

$$\int_{\mathbb{R}_+^n} f(s_1,s_2,\dots,s_n)\rho_\varepsilon(s_1,s_2)\rho_\varepsilon(s_3,s_4)\cdots\rho_\varepsilon(s_{2k-1},s_{2k})\gamma_\varepsilon(s_{2k+1},t_1)$$
$$\cdots\gamma_\varepsilon(s_n,t_{n-2k})ds_1\cdots ds_n \tag{6.9}$$

converges to a function $\operatorname{Tr}_\phi^k f$ in $L_\phi^2(\mathbb{R}_+^{n-2k})$ as $\varepsilon\to 0$. Then the sequence $(S_n^\varepsilon(f), n\in\mathbb{N})$ converges in $L^2(\Omega,\mathcal{F},P)$, and the limit is given by the extended Hu-Meyer formula

$$S_n(f) = \sum_{k\le[\frac{n}{2}]} \frac{n!}{2^k k!(n-2k)!} I_{n-2k}(\operatorname{Tr}_\phi^k f). \tag{6.10}$$

REMARK 3. *It should be noted that the analogue of this theorem and in particular the formula* (6.10) *has been discussed extensively for the Brownian motion.*

As a consequence of Theorem 6.10, (6.10), a chaos expansion theorem is described. It is well known that the family of all polynomials in the random variables $B_{t_1}^H,\dots,B_{t_k}^H$, for $0\le t_1<\cdots<t_k$ and $k\in\mathbb{N}$, is dense in $L^2(\Omega,\mathcal{F},P)$. Since each of these polynomials is a finite sum of the monomials of the form $\int_0^\infty f_1(s_1)\delta B_{s_1}^H\cdots\int_0^\infty f_n(s_n)\delta B_{s_n}^H$, where $f_1,\dots,f_n\in L_\phi^2$, this product of integrals of Wiener type can be expressed as a multiple Stratonovich integral: $\int_0^\infty\cdots\int_0^\infty (f_1\otimes\cdots\otimes f_n)(s_1,\cdots,s_n)\times\delta B_{s_1}^H\cdots\delta B_{s_n}^H$. By the equality (6.10), this multiple integral of Stratonovich type can be expressed as a finite, linear combination of multiple integrals of Itô type. Thus the family of all linear combinations of multiple integrals of the form $\int_0^\infty\cdots\int_0^\infty (f_1\otimes\cdots\otimes f_n)(s_1,\dots,s_n)dB_{s_1}^H\cdots dB_{s_n}^H$ is dense in $L^2(\Omega,\mathcal{F},P)$. Thus

$$L^2(\Omega,\mathcal{F},P) = \Big\{F : F = F_0 + \sum_{n=1}^\infty \int_0^\infty\cdots\int_0^\infty f_n(s_1,\dots,s_n)dB_{s_1}^H\cdots dB_{s_n}^H,$$
$$F_0\in\mathbb{R}, f_n\in L^2(\mathbb{R}_+^n) \text{ and } \sum_{n=1}^\infty |f_n|_\phi^2<\infty\Big\}. \tag{6.11}$$

The equality (6.11) is described in the following theorem.

THEOREM 6.11. *If $F\in L^2(\Omega,\mathcal{F},P)$, then there is a sequence $(f_n\in L_\phi^2(\mathbb{R}_+^n), n\in\mathbb{N})$ such that $\sum_{n=1}^\infty |f_n|_\phi^2<\infty$ and*

$$F = \mathbb{E}(F) + \sum_{n=1}^\infty \int_{\mathbb{R}_+^n} f_n(s_1,\dots,s_n)dB_{s_1}^H\cdots dB_{s_n}^H. \tag{6.12}$$

REMARK 4. *The expansion* (6.12) *is an analogue of the Itô–Wiener chaos expansion which is extended to fractional Brownian motion. Replacing the multiple integrals by the iterated multiple integrals and summing the infinite series gives a stochastic integral representation for $F-\mathbb{E}F$. Note that the terms on the right-hand side of* (6.12) *are orthogonal.*

Acknowledgments. An earlier version of this paper was given to G. Kallianpur and he subsequently provided us with the preprint [3], where the multiple integrals of Stratonovich type are defined and their first and second moments are computed. After the submission of this paper, the authors became aware of [5], where some related work on fractional Brownian motion is done.

REFERENCES

[1] E. F. Beckenbach and R. Bellman, *Inequalities*, Springer, New York, 1965.
[2] W. Dai and C. C. Heyde, *Itô formula with respect to fractional Brownian motion and its application*, J. Appl. Math. Stochastic Anal., 9 (1996), pp. 439–448.
[3] A. Dasgupta and G. Kallianpur, *Multiple fractional integrals*, Appl. Math. Optim., to appear.
[4] C. Dellacherie and P. A. Meyer, *Probability and Potentials B*, North–Holland, Amsterdam, 1982.
[5] L. Decreusefond and A. S. Üstünel, *Stochastic analysis of the fractional Brownian motion*, Potential Anal., 10 (1999), pp. 177–214.
[6] T. E. Duncan, *Absolute continuity for abstract Wiener spaces*, Pacific J. Math., 52 (1974), pp. 359–367.
[7] G. Gripenberg and I. Norros, *On the prediction of fractional Brownian motion*, J. Appl. Probab., 33 (1996), pp. 400–410.
[8] H. Holden, B. Øksendal, J. Ubøe, and T. S. Zhang, *Stochastic Partial Differential Equations, a Modeling, White Noise Functional Analysis*, Birkhäuser, Cambridge, MA, 1996.
[9] Y. Z. Hu and P. A. Meyer, *Chaos de Wiener et intégrales de Feynman*, in Séminaire de Probabilités 22, J. Azema, P. A. Meyer, and M. Yor, eds., Lecture Notes in Math. 1321, Springer-Verlag, New York, NY, 1988, pp. 51–71.
[10] Y. Z. Hu and P. A. Meyer, *Sur les intégrales multiples de Stratonovich*, in Séminaire de Probabilités 26, J. Azema, P. A. Meyer, and M. Yor, eds., Lecture Notes in Math. 1321, Springer-Verlag, New York, 1988, pp. 72–81.
[11] Y. Z. Hu and P. A. Meyer, *On the approximation of Stratonovich multiple integrals*, in Stochastic Processes, a festschrift in honor of G. Kallianpur, S. Cambanis, et al., eds., Springer, New York, 1993, pp. 141–147.
[12] Y. Z. Hu and B. Øksendal, *Wick approximation of anticipating linear stochastic differential equations*, in Stochastic Analysis and Related Topics, Progr. Probab. 38, Birkhäuser, Boston, MA, 1996, pp. 203–231.
[13] C. Houdré, V. Pérez-Abreu, and A. S. Üstünel, *Multiple Wiener-Itô integrals: An introductory survey*, in Chaos Expansions, Multiple Wiener-Itô Integrals and Their Applications, C. Houdré et al., eds., Probab. Stochastics Ser., CRC, Boca Raton, FL, 1994, pp. 1–33.
[14] H. E. Hurst, *Long-term storage capacity in reservoirs*, Trans. Amer. Soc. Civil Eng., 116 (1951), pp. 400–410.
[15] H. E. Hurst, *Methods of using long-term storage in reservoirs*, Proc. Inst. Civil Engineers Part 1, 1956, pp. 519–590.
[16] K. Itô, *Multiple Wiener integrals*, J. Math. Soc. Japan, 3 (1951), pp. 157–164.
[17] S. J. Lin, *Stochastic analysis of fractional Brownian motions*, Stochastics Stochastics Rep. 55 (1995), pp. 121–140.
[18] B. B. Mandelbrot, *The Fractal Geometry of Nature*, Freeman, San Francisco, CA, 1983.
[19] B. B. Mandelbrot and J. W. Van Ness, *Fractional Brownian motions, fractional noises and applications*, SIAM Rev., 10 (1968), pp. 422–437.
[20] M. Métivier and J. Pellaumail, *Stochastic Integration*, Academic Press, New York, London, Toronto, 1980.
[21] P. A. Meyer, *Quantum Probability for Probabilists*, Lecture Notes in Math. 1538, Springer, New York, NY, 1993.
[22] D. W. Stroock and S. R. S. Varadhan, *Multidimensional Diffusion Processes*, Springer, New York, 1979.
[23] S. Watanabe, *Stochastic Differential Equation and Malliavin Calculus*, Tata Institute of Fundamental Research, Springer, New York, 1984.
[24] N. Wiener, *The homogeneous chaos*, Amer. J. Math., 60 (1941), pp. 897–936.

SIAM J. CONTROL OPTIM.
Vol. 38, No. 2, pp. 613–626

NONSMOOTH DUALITY, SANDWICH, AND SQUEEZE THEOREMS*

A. S. LEWIS† AND R. E. LUCCHETTI‡

Abstract. Given a nonlinear function h separating a convex and a concave function, we provide various conditions under which there exists an affine separating function whose graph is somewhere almost parallel to the graph of h. Such results blend Fenchel duality with a variational principle and are closely related to the Clarke–Ledyaev mean value inequality.

Key words. sandwich theorem, squeeze theorem, Fenchel duality, variational principle, mean value inequality, Clarke subdifferential

AMS subject classifications. 49J52, 90C46, 26D07

PII. S0363012998334213

1. Introduction. The central theorems in this paper blend two completely distinct types of result, both fundamental in optimization theory: Fenchel duality and variational principles. The simplest version of Fenchel duality states that for any convex functions f and g on $\mathbf{R}^n$ satisfying $f \geq -g$, a regularity condition implies the set

$$L \stackrel{\text{def}}{=} \{y \in \mathbf{R}^n : f^*(y) + g^*(-y) \leq 0\}$$

is nonempty (where f^* is the Fenchel conjugate of f). Geometrically, this means there exists an affine function sandwiched between f and $-g$. On the other hand, one of the easiest examples of a variational principle states that if h is a locally Lipschitz function bounded below on $\mathbf{R}^n$, then h has arbitrarily small Clarke subgradients:

$$0 \in \mathrm{cl}\,(\mathrm{Im}\ \partial h).$$

Geometrically, there are points where the graph of h is almost horizontal (in a certain nonsmooth sense).

The theorems we discuss here combine the features of both results above. We consider functions f, g, and h as before, now satisfying $f \geq h \geq -g$, and under various regularity conditions we prove

$$L \cap \mathrm{cl}\,(\mathrm{Im}\ \partial h) \neq \emptyset.$$

Geometrically, there are affine functions between f and $-g$ whose graphs are somewhere almost parallel to the graph of h.

As we show by means of various examples, the existence of a suitable affine separating function depends on both local and asymptotic properties of the three functions. Hence the regularity conditions we need to impose combine assumptions on the domains of the primal functions, f and g, and of their conjugates, f^* and g^*, as well as local and global growth conditions on h.

*Received by the editors February 18, 1998; accepted for publication (in revised form) May 27, 1999; published electronically February 9, 2000. This research was partially supported by the Natural Sciences and Engineering Research Council of Canada.

http://www.siam.org/journals/sicon/38-2/33421.html

†Department of Combinatorics and Optimization, University of Waterloo, Waterloo, Ontario, Canada N2L 3G1 (aslewis@orion.uwaterloo.ca; http://orion.uwaterloo.ca/∼aslewis).

‡Politecnico di Milano, Facoltá di Ingegneria di Como, P.le Gerbetto 6, 22100 Como, Italy (rel@komodo.ing.unico.it).

The key tool for our results is a recent, somewhat surprising mean value inequality of Clarke–Ledyaev [3], rephrased as a hybrid sandwich theorem in [6]. We illustrate the application of this type of result with two apparently simple but rather remarkable consequences. First, any convex function f and locally Lipschitz function $h \leq f$ satisfy

$$\operatorname{dom} f^* \cap \operatorname{cl}(\operatorname{Im} \partial h) \neq \emptyset.$$

Second (a "squeeze theorem"), any locally Lipschitz functions $p \geq h \geq q$ with $p(0) = h(0) = q(0)$ satisfy

$$\partial p(0) \cap \partial h(0) \cap \partial q(0) \neq \emptyset.$$

We have not been able to find simple proofs or references for either of these two results.[1]

With the exception of this last squeeze theorem, our results do not appear to be substantially easier with the assumption that h is smooth (in which case ∂h reduces to the singleton ∇h). We believe they provide further evidence of the depth, applicability, and fundamental nature of the Clarke–Ledyaev inequality in optimization theory.

2. Notation and preliminary results. We begin by reviewing some basic ideas from convex analysis (see [7]). Given a convex set $A \subset \mathbf{R}^n$, we denote by $\operatorname{aff} A$ the smallest affine space containing A and by $\operatorname{ri} A$ the set of the internal points of $A \subset \operatorname{aff} A$ (with the induced topology). Observe that $\operatorname{ri} A$ is a nonempty convex set. Given a function $f : \mathbf{R}^n \to [-\infty, \infty]$, we denote its *effective domain* by

$$\operatorname{dom} f \stackrel{\text{def}}{=} \{x \in \mathbf{R}^n : f(x) < \infty\}$$

and by $\operatorname{epi} f$ its epigraph, the set

$$\operatorname{epi} f \stackrel{\text{def}}{=} \{(x, r) \in \mathbf{R}^n \times \mathbf{R} : r \geq f(x)\},$$

a convex set if and only if f is convex. The hypograph of the function $-g$ is instead

$$\operatorname{hyp}(-g) \stackrel{\text{def}}{=} \{(x, r) \in \mathbf{R}^n \times \mathbf{R} : r \leq -g(x)\},$$

again a convex set if and only if g is convex. The epigraph of f is closed if and only if f is lower semicontinuous in the usual sense.

We shall write $f \in \Gamma_0$ to mean that $\operatorname{epi} f$ is nonempty, closed, and convex and does not contain vertical lines.

For a set A, let I_A be the indicator function of the set A,

$$I_A(x) = \begin{cases} 0 & \text{if } x \in A, \\ \infty & \text{otherwise.} \end{cases}$$

In particular, I_{kB} denotes the indicator function of the ball centered at $0 \in \mathbf{R}^n$ and with radius k.

[1] Subsequent investigations revealed alternative approaches to the last theorem independent of the Clarke–Ledyaev result [1]. Nonetheless, the original approach we present here remains attractive for its transparency.

The *Fenchel conjugate* of a function $f : \mathbf{R}^n \to [-\infty, \infty]$ is the function $f^* : \mathbf{R}^n \to [-\infty, \infty]$ defined by

$$f^*(y) \stackrel{\text{def}}{=} \sup_{x \in \mathbf{R}^n} \{\langle y, x\rangle - f(x)\},$$

a convex lower semicontinuous function (even if f is not) which belongs to Γ_0 if f does. Furthermore, $f = f^{**}$, providing $f \in \Gamma_0$.

The function f is said to be *cofinite* if its conjugate f^* satisfies dom $f^* = \mathbf{R}^n$. It is easy to see this is equivalent to saying that $\lim_{\|x\| \to \infty} \frac{f(x)}{\|x\|} = \infty$ [5, Chapter 10, Proposition 1.3.8].

The *subdifferential* of a convex function f at a point $x \in$ dom f is the closed convex set

$$\partial f(x) \stackrel{\text{def}}{=} \{y \in \mathbf{R}^n : f(z) \geq f(x) + \langle y, z - x\rangle \text{ for all } z \in \mathbf{R}^n\}.$$

The fundamental connection between the subdifferentials of a function and of its Fenchel conjugate is shown by the following *Fenchel identity*:

$$y \in \partial f(x) \Longleftrightarrow f(x) + f^*(y) = \langle y, x\rangle.$$

It follows, in particular, that $y \in \partial f(x)$ if and only if $x \in \partial f^*(y)$, providing $f \in \Gamma_0$.

Given two functions $p, q : \mathbf{R}^n \to [-\infty, \infty]$, we define the *infimal convolution* between them by

$$(p \square q)(x) = \inf_{y \in \mathbf{R}^n} \{p(y) + q(x - y)\},$$

a convex function if p and q are, possibly assuming the value $-\infty$, and that may fail to be lower semicontinuous. Finally, for a function $p \in \Gamma_0$, we denote by

$$p_k \stackrel{\text{def}}{=} p \square k\|\cdot\|,$$

the infimal convolution between p and $k\|\cdot\|$: this function is the largest k-Lipschitz function minorizing p. For more about convex functions, the interested reader is invited to consult [4], [5], [7].

Next we briefly consider the subdifferential of a locally Lipschitz function $h : U \to \mathbf{R}$, where U is an open subset of $\mathbf{R}^n$. This notion is not uniquely defined in the literature, and here we make the choice of the *Clarke subdifferential* (see [2]) which is more suited for our scopes, as an example in the final section will show. To define it, first let us introduce the notion of *generalized directional derivative* of h at the point x in the direction v:

$$h^\circ(x, v) \stackrel{\text{def}}{=} \limsup_{\substack{z \to x \\ t \searrow 0}} \frac{h(z + tv) - h(z)}{t}.$$

The function $v \mapsto h^\circ(x, v)$ is everywhere finite, subadditive, and positively homogeneous; hence, in particular, it is continuous and convex. Then the *subdifferential* of h at x is defined as

$$\partial h(x) \stackrel{\text{def}}{=} \{y \in \mathbf{R}^n : \langle y, v\rangle \leq h^\circ(x, v) \text{ for all } v \in \mathbf{R}^n\},$$

a nonempty closed convex set. Moreover, if k is a Lipschitz constant for h, the subdifferential is norm-bounded by k. In particular, the multifunction ∂h is bounded on

bounded sets. Observe that the Clarke subdifferential of h at x is the same set as the (convex) subdifferential of the function $v \mapsto h^\circ(x, v)$ at $v = 0$, a simple but useful property we shall use throughout this paper.

For more about nonsmooth analysis for locally Lipschitz functions, the interested reader is invited to consult [2].

In this paper we shall deal with two convex functions $f, g \in \Gamma_0$ and a locally Lipschitz function h such that $f \geq h \geq -g$. For a moment, let us focus on the problem of nonemptyness of the set

$$L \stackrel{\text{def}}{=} \{y \in \mathbf{R}^n : f^*(y) + g^*(-y) \leq 0\}.$$

This set can be characterized in a more geometric way, as the following easy proposition states.

PROPOSITION 2.1. *Let $f, g \in \Gamma_0$. Then, for $y \in \mathbf{R}^n$,*

$$f^*(y) + g^*(-y) \leq 0$$

if and only if there exists $a \in \mathbf{R}$ such that

$$f(x) \geq a + \langle y, x \rangle \geq -g(x) \quad \text{for all } x \in \mathbf{R}^n.$$

Thus the problem of nonemptyness of L is equivalent to finding an affine separator lying below f and above g. This can be stated in terms of a separation problem for the sets $\operatorname{epi} f$ and $\operatorname{hyp}(-g)$. The assumption $f \geq -g$ ensures that

$$\operatorname{ri} \operatorname{epi} f \cap \operatorname{ri} \operatorname{hyp}(-g) = \emptyset,$$

(see [4, Chapter 4, Proposition 1.1.9]) and this in turn implies that $\operatorname{epi} f$ and $\operatorname{hyp}(-g)$ can be separated by a hyperplane [4, Chapter 3, Theorem 4.1.4]. However, it can happen that the only separating hyperplane is vertical, which unfortunately says nothing about nonemptyness of L.

The first result stating that L is nonempty is the following well-known Fenchel duality theorem [7, Theorem 31.1], which in our setting can be rephrased in the following way.

THEOREM 2.1. *Let $f, g \in \Gamma_0$ be such that $f \geq -g$ and suppose*

$$\operatorname{ri}(\operatorname{dom} f) \cap \operatorname{ri}(\operatorname{dom} g) \neq \emptyset.$$

Then there exists $y \in \mathbf{R}^n$ such that

$$f^*(y) + g^*(-y) \leq 0.$$

We illustrate the role of the assumption on the domains of f and g with the help of the following four examples, where the set L is always empty.

EXAMPLE 2.1.

$$f(x) = \begin{cases} -\sqrt{x} & \text{if } x \geq 0, \\ \infty & \text{otherwise,} \end{cases}$$

$$g(x) = \begin{cases} 0 & \text{if } x = 0, \\ \infty & \text{otherwise.} \end{cases}$$

Here $\operatorname{ri}(\operatorname{dom} f) \cap \operatorname{ri}(\operatorname{dom} g) = \emptyset$.

EXAMPLE 2.2.

$$f(u,v) = \begin{cases} -1 & \text{if } uv \geq 1, u \geq 0, \\ \infty & \text{otherwise,} \end{cases}$$

$$g(u,v) = \begin{cases} 0 & \text{if } u \geq 0, v = 0, \\ \infty & \text{otherwise.} \end{cases}$$

Here we have $\operatorname{dom} f \cap \operatorname{dom} g = \emptyset$.

EXAMPLE 2.3.

$$f(u,v) = \begin{cases} u & \text{if } v = -1, \\ \infty & \text{otherwise,} \end{cases}$$

$$g(u,v) = \begin{cases} 0 & \text{if } v = 0, \\ \infty & \text{otherwise.} \end{cases}$$

Here the distance between $\operatorname{dom} f$ *and* $\operatorname{dom} g$ *is* 1.

In the last two examples the domains of f and g do not intersect, while in the first example a crucial role is played by the fact that $\inf(f+g) = 0$. In the following example $\inf(f+g) > 0$, and yet there is no affine separator. Observe that such an example could not be provided in one dimension [4, Chapter 1, Remark 3.3.4].

EXAMPLE 2.4.

$$f(u,v) = \begin{cases} 1 - 2\sqrt{uv} & \text{if } u, v \geq 0, \\ \infty & \text{otherwise,} \end{cases}$$

$$g(u,v) = \begin{cases} 1 - 2\sqrt{-uv} & \text{if } u \leq 0, v \geq 0, \\ \infty & \text{otherwise.} \end{cases}$$

A straightforward calculation shows

$$f^*(u^*,v^*) = \begin{cases} -1 & \text{if } u^* \leq 0, u^*v^* \geq 1, \\ \infty & \text{otherwise,} \end{cases}$$

$$g^*(u^*,v^*) = \begin{cases} -1 & \text{if } u^* \geq 0, u^*v^* \leq -1, \\ \infty & \text{otherwise.} \end{cases}$$

Thus the set L is empty.

3. Sandwich theorems. We turn now to the case of three functions f, g, and h, such that f and g are convex, h is locally or globally Lipschitz, and $f \geq h \geq -g$. Examples 2.2, 2.3, and 2.4 show that the existence of a locally Lipschitz function h between f and $-g$ (take $h(x,y) = -xy$ in all cases) does not change the situation: there is no affine separator.

First let us recall now some known results.

THEOREM 3.1 (see [6, Theorem 2]). *Let C be a nonempty convex compact set in $\mathbf{R}^n$. Let $f, g \in \Gamma_0$ and with domains contained in C. Let $h : \mathbf{R}^n \to \mathbf{R}$ be Lipschitz on a neighborhood of C. Suppose moreover $f \geq h \geq -g$ on C.*

Then there exist $c \in C$ and $y \in \partial h(c)$ such that

$$f^*(y) + g^*(-y) \leq 0.$$

Boundedness of C can be relaxed at the expense of requiring more about the functions f and g and/or about the function h. Specifically, we have the following two results.

THEOREM 3.2 (see [6, Theorem 7]). *Let C be a nonempty closed convex set in $\mathbf{R}^n$. Let $f, g \in \Gamma_0$ be cofinite, with domains contained in C. Moreover suppose*

$$\operatorname{int}(\operatorname{dom} f) \cap \operatorname{int}(\operatorname{dom} g) \neq \emptyset.$$

Let $h : \mathbf{R}^n \to \mathbf{R}$ be locally Lipschitz on a neighborhood of C and suppose $f \geq h \geq -g$ on C.

Then there exist $c \in C$ and $y \in \partial h(c)$ such that

$$f^*(y) + g^*(-y) \leq 0.$$

THEOREM 3.3 (see [6, Theorem 8]). *Let C be a nonempty closed convex set in $\mathbf{R}^n$. Let $f, g \in \Gamma_0$ be cofinite, with domains contained in C. Let $h : \mathbf{R}^n \to \mathbf{R}$ be globally Lipschitz on a neighborhood of C and suppose $f \geq h \geq -g$ on C.*

Then there exist $c \in C$ and $y \in \partial h(c)$ such that

$$f^*(y) + g^*(-y) \leq 0.$$

Observe one does not need a qualification condition on the domains of f and g if h is globally Lipschitz. In the last two theorems, however, cofiniteness is required, which can be regarded as a (strong) qualification condition on the domains of the conjugates.

The first result we want to prove deals simply with the existence of the affine separator. To prove it, we need the following proposition about regularizing Fenchel problems.

PROPOSITION 3.1. *Suppose $p, q \in \Gamma_0$ and*

$$(\bullet) \qquad \operatorname{ri}(\operatorname{dom} p) \cap \operatorname{ri}(\operatorname{dom} q) \neq \emptyset.$$

Then, for all large k, we have

$$\inf(p+q) = \inf(p_k + q_k)$$

and

$$\operatorname{argmin}(p+q) = \operatorname{argmin}(p_k + q_k).$$

Proof. To prove the first equality, we need to prove only $\inf(p+q) \leq \inf(p_k + q_k)$. There is nothing to prove if $\inf(p+q) = -\infty$. Therefore, let us assume it is finite. (It cannot be ∞ because of $(\bullet)$.) By Fenchel duality, there is $y \in \mathbf{R}^n$ such that

$$-\inf(p+q) = p^*(y) + q^*(-y).$$

Take $k > \|y\|$. Then

$$-\inf(p+q) = p^*(y) + q^*(-y) = (p^* + I_{kB})(y) + (q^* + I_{kB})(-y)$$

$$= (p_k)^*(y) + (q_k)^*(-y) \geq \inf_{z \in \mathbf{R}^n}((p_k)^*(z) + (q_k)^*(-z))$$

$$= -\inf(p_k + q_k) \geq -\inf(p+q).$$

This shows the first equality and also that y as above is optimal for the problem of minimizing, on $\mathbf{R}^n$, $(p_k)^*(\cdot) + (q_k)^*(-\cdot)$.

Now, writing down optimality conditions, we obtain, using $k > \|y\|$,

$$\begin{aligned} x \in \operatorname{argmin}(p+q) &\Leftrightarrow p(x) + q(x) = -p^*(y) - q^*(-y), \\ &\Leftrightarrow x \in \partial p^*(y) \cap \partial q^*(-y), \\ &\Leftrightarrow x \in \partial(p^* + I_{kB})(y) \cap \partial(q^* + I_{kB})(-y), \\ &\Leftrightarrow x \in \partial(p_k)^*(y) \cap \partial(q_k)^*(-y), \\ &\Leftrightarrow x \in \operatorname{argmin}(p_k + q_k). \qquad \square \end{aligned}$$

We begin our sequence of main results by proving some variants of Fenchel duality, where the usual regularity condition is replaced by the existence of a Lipschitz separator.

THEOREM 3.4. *For $f, g \in \Gamma_0$, suppose*

$$\operatorname{ri}(\operatorname{dom} f^*) \cap \operatorname{ri}(-\operatorname{dom} g^*) \neq \emptyset.$$

Suppose further there exists a locally Lipschitz function h such that $f \geq h \geq -g$. Then there is $y \in \mathbf{R}^n$ such that

$$f^*(y) + g^*(-y) \leq 0.$$

(Moreover, if $\inf(f+g) = 0$, then such a y can be found in the range of ∂h.)

Proof. From Proposition 3.1, applied to $p = f^*$ and $q(\cdot) = g^*(-\cdot)$, we have

$$\inf((f^*)_k(\cdot) + (g^*)_k(-\cdot)) = \inf(f^*(\cdot) + g^*(-\cdot))$$

and

$$\operatorname{argmin}((f^*)_k(\cdot) + (g^*)_k(-\cdot)) = \operatorname{argmin}(f^*(\cdot) + g^*(-\cdot))$$

for all large k. Apply Theorem 3.1 to the functions $f + I_{kB}$, h, and $g + I_{kB}$, for large k, and the set $C = kB$, to find $y_k \in \operatorname{Im} \partial h$ such that

$$(f^*)_k(y_k) + (g^*)_k(-y_k) \leq 0.$$

If

$$y_k \in \operatorname{argmin}((f^*)_k(\cdot) + (g^*)_k(-\cdot))$$

(as in the case when $\inf(f+g) = 0$), then we deduce $f^*(y_k) + g^*(-y_k) \leq 0$ and we conclude. Otherwise, for all large k,

$$0 > \inf((f^*)_k(\cdot) + (g^*)_k(-\cdot)) = \inf(f^*(\cdot) + g^*(-\cdot)).$$

Thus there is $y \in \mathbf{R}^n$ such that $f^*(y) + g^*(-y) \leq 0$, as required. □

We provided here the result when $\inf(f+g) = 0$ for the sake of completeness. However observe that under the assumptions of Theorem 3.4 $\inf f + g$ is attained. In this circumstance the squeeze theorem in the next section will provide a more precise result.

With respect to the role of the assumptions in Theorem 3.4, Example 2.1 shows the set L can be empty if we do not assume the existence of a locally Lipschitz function sandwiched between f and $-g$, while Example 2.2 shows the necessity of the qualification condition

$$\text{ri}\,(\text{dom } f^*) \cap \text{ri}\,(-\text{dom } g^*) \neq \emptyset.$$

We turn now to the problem of providing conditions under which the slope of an affine separator can be found in the closure of the range of the Clarke subdifferential of the separating function h. To do this, we prove first the following proposition.

PROPOSITION 3.2. *Suppose $f, g \in \Gamma_0$ satisfy $f \geq -g$. For $k = 1, 2, \ldots$, define*

$$L_k \stackrel{\text{def}}{=} \{y \in \mathbf{R}^n : (f^*)_k(y) + (g^*)_k(-y) \leq 0\}$$

and

$$L \stackrel{\text{def}}{=} \{y \in \mathbf{R}^n : f^*(y) + g^*(-y) \leq 0\}.$$

Then L_k is a decreasing collection of closed convex sets containing L, and

$$y_k \in L_k, \quad y_k \to y \quad \textit{implies} \quad y \in L.$$

If moreover the condition

$$0 \in \text{int}\,(\text{dom } f - \text{dom } g)$$

holds, then the sets L_k for large k are all contained in a compact set.

Proof. Since $(f^*)_k \leq (f^*)_{k+1} \leq f^*$, clearly $L \subset L_{k+1} \subset L_k$, for all $k > 0$. Let us prove that, if y_k is such that $y_k \to y$ and

$$(f^*)_k(y_k) + (g^*)_k(-y_k) \leq 0 \quad \text{for all } k,$$

then

$$f^*(y) + g^*(-y) \leq 0.$$

From Proposition 2.1 there exists $a_k \in \mathbf{R}$ such that

$$f(x) \geq a_k + \langle y_k, x\rangle \geq -g(x) \quad \text{for all } x \in kB.$$

It is easy to show the sequence $\{a_k\}$ is bounded, so it has some cluster $a \in \mathbf{R}$. (Use the boundedness of $\{y_k\}$ and the existence of an element $x \in \text{dom } f \cap \text{dom } g$.) It follows that

$$f(x) \geq a + \langle y, x\rangle \geq -g(x) \quad \text{for all } x \in \mathbf{R}^n,$$

so $y \in L$. We have proved the first part of the claim. Now define a function

$$v(w) = \inf_{x \in \mathbf{R}^n} (f(x+w) + g(x))$$

and a sequence of functions decreasing pointwise to v,

$$v^k(w) = \inf_{x\in\mathbf{R}^n}((f + I_{kB})(x + w) + (g + I_{kB})(x)).$$

Observe that $(v^k)^*(y) = (f^*)_k(y) + (g^*)_k(-y)$ and $v^*(y) = f^*(y) + g^*(-y)$ and that $\operatorname{dom} v = \operatorname{dom} f - \operatorname{dom} g$, so that $0 \in \operatorname{int}(\operatorname{dom} v)$.

Since v is continuous at 0, there exist reals $r > 0$ and α and a cube C such that $rB \subset C \subset \operatorname{int}(\operatorname{dom} v)$ and $v \le \alpha - 1$ on C. Hence, for large k we have $v^k \le \alpha$ on each vertex of C and hence on rB, so $(v^k)^*(w) \ge r\|w\| - \alpha$ for all points w in $\mathbf{R}^n$, and therefore $L_k \subset (\alpha/r)B$. □

We are now ready for a new result.

THEOREM 3.5. *For $f, g \in \Gamma_0$, and locally Lipschitz $h : \mathbf{R}^n \to \mathbf{R}$ satisfying $f \ge h \ge -g$, suppose*

$$0 \in \operatorname{int}(\operatorname{dom} f - \operatorname{dom} g).$$

Then

$$\exists y \in \operatorname{cl}(\operatorname{Im} \partial h) : f^*(y) + g^*(-y) \le 0.$$

Proof. Apply Theorem 3.1 to the functions $f + I_{kB} \ge h \ge -(g + I_{kB})$, for large k. Then there exists

$$y_k \in \operatorname{Im}(\partial h) : (f^*)_k(y_k) + (g^*)_k(-y_k) \le 0.$$

By Proposition 3.2 the sequence (y_k) clusters and any cluster point satisfies the required property. □

We intend now to prove that the constraint qualification in Theorem 3.5 can be replaced by an assumption involving the growth of f and h at infinity. To do this, we need the following proposition.

PROPOSITION 3.3. *For $f \in \Gamma_0$ and locally Lipschitz h satisfying $f \ge h$, suppose*

$$\liminf_{\|x\|\to\infty} \frac{f(x)}{\|x\|} > \max\left\{\limsup_{\|x\|\to\infty} \frac{h(x)}{\|x\|}, 0\right\}.$$

Then, for all large k, $f_k \ge h$.

Proof. Let $0 < a < b$ and c be such that

$$\frac{f(x)}{\|x\|} \ge b, \quad \frac{h(x)}{\|x\|} \le a,$$

for all x such that $\|x\| \ge c$. Then there exists $rbf \in \mathbf{R}$ such that

$$f(x) \ge r + b\|x\| \quad \text{for all } x \in \mathbf{R}^n,$$

and f has bounded level sets. For the sake of contradiction, suppose there exists, for each $k \in \mathbf{N}$, x_k such that $f_k(x_k) < h(x_k)$. Two cases can occur.

(i) (x_k) is unbounded. Taking a subsequence, we can suppose $\|x_k\| \to \infty$. For $k > b$, we have $f_k(x_k) \ge r + b\|x_k\|$. It follows that

$$a\|x_k\| \ge h(x_k) > f_k(x_k) \ge r + b\|x_k\|,$$

a contradiction.

(ii) (x_k) is bounded. Again taking a subsequence, we can suppose $x_k \to x$. Pick $m > \|x\|$ and r so that h is r-Lipschitz on mB. Since f has compact level sets, for each k there is y_k such that $h(x_k) > f_k(x_k) = f(y_k) + k\|x_k - y_k\|$. As $h(x_k) \to h(x)$, for large k one has

$$f(y_k) \leq f(y_k) + k\|x_k - y_k\| \leq h(x) + 1.$$

Thus (y_k) is bounded and, taking another subsequence, we can suppose $y_k \to y$. Since

$$k\|x_k - y_k\| \leq h(x) + 1 - \inf f \quad \text{for all } k,$$

we deduce $y = x$. Thus, for large k, $x_k, y_k \in mB$, so

$$h(x_k) \leq h(y_k) + r\|x_k - y_k\| \leq f(y_k) + k\|x_k - y_k\| < h(x_k),$$

a contradiction. □

THEOREM 3.6. *For $f, g \in \Gamma_0$ and locally Lipschitz $h : \mathbf{R}^n \to \mathbf{R}$ satisfying $f \geq h \geq -g$, suppose*

$$\liminf_{x\to\infty} \frac{f(x)}{\|x\|} > \max\left\{\limsup_{x\to\infty} \frac{h(x)}{\|x\|}, 0\right\}.$$

Then

$$\exists y \in \text{ cl}\,(\text{Im } \partial h) : f^*(y) + g^*(-y) \leq 0.$$

Proof. From Proposition 3.3, $f_k \geq h \geq -g$ for large k. Since we know $\text{dom } f_k - \text{dom } g = \mathbf{R}^n$, Theorem 3.5 implies there exists $y \in \text{cl Im } \partial h$ with

$$f^*(y) + g^*(-y) \leq (f_k)^*(y) + g^*(-y) \leq 0. \qquad \square$$

The proof of the theorem above relies on the fact that we are able to construct a function $p \in \Gamma_0$ such that $f \geq p \geq h$ and whose domain contains internal points. However, to do this is not always possible, as the following example shows.

EXAMPLE 3.1. *Let*

$$f(u, v) = \begin{cases} 0 & \text{if } v = 0, \\ \infty & \text{otherwise} \end{cases}$$

and $h(u, v) = |uv|$. Suppose the convex function p satisfies $f \geq p \geq h$. Then clearly $p(u, 0) = 0$ for all u. For any real u and v and positive integer r,

$$\begin{aligned} \frac{1}{r}p(u, v) &= \frac{1}{r}p(u, v) + \frac{r-1}{r}p(u, 0) \\ &\geq p\left(u, \frac{v}{r}\right) \\ &= p\left(u, \frac{v}{r}\right) + p(r, 0) \\ &\geq 2p\left(\frac{u+r}{2}, \frac{v}{2r}\right) \\ &\geq \left|\frac{(u+r)v}{2r}\right| \\ &\to \frac{|v|}{2} \end{aligned}$$

as $r \to \infty$. Hence $p(u,v) = +\infty$ whenever $v \neq 0$, so $p = f$.

The next result deals with the case of h being globally Lipschitz.

THEOREM 3.7. *For $f, g \in \Gamma_0$ and globally Lipschitz $h : \mathbf{R}^n \to \mathbf{R}$, suppose $f \geq h \geq -g$. Then*

$$\exists y \in \ \mathrm{cl}\,(\mathrm{Im}\ \partial h) : f^*(y) + g^*(-y) \leq 0.$$

Proof. Apply Theorem 3.3 to the functions $f + I_{kB}$, h, and $g + I_{kB}$ to obtain the existence of $y_k \in \mathrm{Im}\ \partial h$ such that

$$(f + I_{kB})^*(y_k) + (g + I_{kB})^*(-y_k) \leq 0.$$

Since h is globally Lipschitz, (y_k) has a cluster point y. Then we conclude with the help of Proposition 3.2. □

The next example shows that in general no $y \in \mathrm{Im}\ \partial h$ satisfies $f^*(y)+g^*(-y) \leq 0$.

EXAMPLE 3.2. *Let $f(x) = |x|$,*

$$g(x) = \begin{cases} 1 + x^2 & \text{if } x \geq 0, \\ \infty & \text{otherwise,} \end{cases}$$

$$h(x) = \begin{cases} x - \exp(-x) & \text{if } x \geq 0, \\ 2x - 1 & \text{otherwise.} \end{cases}$$

Then all the assumptions of Theorem 3.5 are fulfilled, and moreover h is globally Lipschitz.

We end the section by proving a unilateral result which can be regarded as a generalized variational principle.

THEOREM 3.8. *For $f \in \Gamma_0$ and locally Lipschitz $h : \mathbf{R}^n \to \mathbf{R}$, suppose $f \geq h$. Then*

$$\mathrm{cl}\,(\mathrm{Im}\,\partial h) \cap \mathrm{dom}\, f^* \neq \emptyset.$$

Proof. Choose any point $z \in \mathrm{dom}\ f$ and real $k > \|z\|$. Define $g(\cdot) = I_{kB}(\cdot) - \inf_{kB} h$ and apply Theorem 3.5. □

The special case $f = 0$ gives the well-known variational result that a locally Lipschitz function h which is bounded above satisfies $0 \in \mathrm{cl}\,(\mathrm{Im}\,\partial h)$.

4. Squeeze theorems. In this section we specialize the situation studied before. We shall make the further assumption that there is a point where the three functions are equal. In this case, as we shall see, we are able to provide more precise results.

We shall start with the following easy proposition, that we state without proof.

PROPOSITION 4.1. *For $f, g \in \Gamma_0$ satisfying $f \geq -g$, suppose there exists x such that $f(x) = -g(x)$. Then*

$$\{y : f^*(y) + g^*(-y) \leq 0\} = \partial f(x) \cap -\partial g(x).$$

We prove now a "convex" squeeze theorem.

THEOREM 4.1. *For $f, g \in \Gamma_0$ and locally Lipschitz $h : \mathbf{R}^n \to \mathbf{R}$ satisfying $f \geq h \geq -g$, suppose there exists $x \in \mathbf{R}^n$ such that $f(x) = -g(x)$. Then*

$$\partial f(x) \cap \partial h(x) \cap -\partial g(x) \neq \emptyset.$$

Proof. Without loss of generality, we can suppose $x = 0$. For each positive integer r, as $f \geq h \geq -g$ on $\frac{1}{r}B$, we can apply Theorem 3.1 to find $x_r \in \frac{1}{r}B$ and $y_r \in \partial h(x_r)$ with

$$(f + I_{\frac{1}{r}B})^*(y_r) + (g + I_{\frac{1}{r}B})^*(-y_r) \leq 0.$$

By Proposition 4.1

$$y_r \in \partial(f + I_{\frac{1}{r}B})(0) \cap -\partial(g + I_{\frac{1}{r}B})(0) = \partial f(0) \cap -\partial g(0).$$

Since ∂h is locally bounded, there exists a subsequence (y_{r_k}) of (y_r) converging to some y, and since $x_{r_k} \to 0$ and ∂h is closed, $y \in \partial h(0)$. □

The next squeeze theorem deals instead with three locally Lipschitz functions. To prove it, we need the following proposition.

PROPOSITION 4.2. *Let $f : \mathbf{R}^n \to \mathbf{R}$ be locally Lipschitz and suppose $\delta > 0$. Then*

$$f(0) + f^\circ(0, x) + \delta\|x\| > f(x)$$

for all small nonzero x.

Proof. Suppose that $f(0) = 0$, that f is k-Lipschitz near 0, and that, for the sake of contradiction, there is a sequence (x_r) such that $x_r \neq 0$ for all r and $x_r \to 0$, with

$$f^\circ(0, x_r) + \delta\|x_r\| \leq f(x_r).$$

Thus

$$f^\circ\left(0, \frac{x_r}{\|x_r\|}\right) + \delta \leq \frac{f(x_r)}{\|x_r\|}.$$

Suppose, without loss of generality, $\frac{x_r}{\|x_r\|} \to d$. Then

$$\begin{aligned}
\limsup_{r\to\infty} f^\circ\left(0, \frac{x_r}{\|x_r\|}\right) + \delta &\leq \limsup_{r\to\infty} \frac{f(x_r)}{\|x_r\|} \\
&= \limsup_{r\to\infty} \frac{f(\|x_r\|d) + f(x_r) - f(\|x_r\|d)}{\|x_r\|} \\
&\leq \limsup_{r\to\infty} \frac{f(\|x_r\|d) + k\|x_r - \|x_r\|d\|)}{\|x_r\|} \\
&\leq f^\circ(0, d).
\end{aligned}$$

It follows that

$$f^\circ(0, d) + \delta \leq f^\circ(0, d),$$

which is impossible. □

THEOREM 4.2. *Suppose $f, h, g : \mathbf{R}^n \to \mathbf{R}$ are three locally Lipschitz functions such that $f \geq h \geq g$. Moreover, suppose $f(x) = g(x)$ for some x. Then*

$$\partial f(x) \cap \partial h(x) \cap \partial g(x) \neq \emptyset.$$

Proof. Suppose, without loss of generality, $f(0) = h(0) = g(0) = 0$. By Proposition 4.2, for each $r \in \mathbf{N}$, there exists $\varepsilon_r > 0$ such that

$$f^\circ(0, x) + \frac{\|x\|}{r} \geq h(x) \geq -\left((-g)^\circ(0, x) + \frac{\|x\|}{r}\right)$$

for all x such that $\|x\| \le \varepsilon_r$. Now, take $\varepsilon < \varepsilon_r$. Then

$$f^\circ(0,x) + \frac{\|x\|}{r} + I_{\varepsilon B}(x) \ge h(x) \ge -\left((-g)^\circ(0,x) + \frac{\|x\|}{r} + I_{\varepsilon B}(x)\right)$$

for all $x \in \mathbf{R}^n$. We can then apply Theorem 4.1 to get an element y_r such that

$$\begin{aligned} y_r &\in \partial\left(f^\circ(0,\cdot) + \frac{\|\cdot\|}{r} + I_{\varepsilon B}(\cdot)\right)(0) \cap \partial h(0) \\ &\quad \cap -\partial\left((-g)^\circ(0,\cdot) + \frac{\|\cdot\|}{r} + I_{\varepsilon B}(\cdot)\right)(0) \\ &= \left(\partial f(0) + \frac{1}{r}B\right) \cap \partial h(0) \cap \left(\partial g(0) + \frac{1}{r}B\right). \end{aligned}$$

Since $y_r \in \partial h(0)$, the sequence (y_r) is bounded and any of its cluster points does the job. □

COROLLARY 4.1. *Let $f_1 \ge f_2 \ge \cdots f_k : \mathbf{R}^n \to \mathbf{R}$ be locally Lipschitz. Suppose $f_1(0) = \cdots = f_k(0)$. Then*

$$\bigcap_{i=1}^{k} \partial f_i(0) \ne \emptyset,$$

provided at least one of the following conditions holds:

- *$k = 1, 2, 3$;*
- *at least one f_i is smooth;*
- *$n = 1, 2$.*

Proof. The cases $k = 2$ and the case when f_i is smooth follow from the sum rule applied to $f_1 - f_2$ and $f_j - f_i$, respectively. The case $k = 3$ is Theorem 4.2 and the cases $n = 1, 2$ are consequences of Theorem 4.2 and Helly's theorem. □

5. Final remarks. We have seen some sandwich and squeeze theorems, dealing with convex and locally Lipschitz functions. While the convex subdifferential is standard, there are several notions of subdifferential for locally Lipschitz functions. Here we use the Clarke subdifferential rather than, for instance, the approximate subdifferential, because the latter is not suitable for the results we seek. Consider the following simple example.

EXAMPLE 5.1. *Let*

$$f(x) = I_{[-1,1]}(x),$$

$$g(x) = |x| + I_{[-1,1]}(x),$$

and

$$h(x) = -|x|.$$

Then $f \ge h \ge -g$ and h is (globally) Lipschitz. However

$$L = \{y : f^*(y) + g^*(y) \le 0\} = \{0\},$$

while the approximate subdifferential of h is the set $\{-1, 1\}$.

Finally, here is a list of questions we leave to the interested reader.

Question 1. Does

$$\operatorname{ri}(\operatorname{dom} f) \cap \operatorname{ri}(\operatorname{dom} g) \neq \emptyset$$

imply

$$L \cap \operatorname{cl}(\operatorname{Im} \partial h) \neq \emptyset?$$

Question 2. Does

$$\operatorname{ri}(\operatorname{dom} f^*) \cap \operatorname{ri}(-\operatorname{dom} g^*) \neq \emptyset$$

imply

$$L \cap \operatorname{cl}(\operatorname{Im} \partial h) \neq \emptyset?$$

Question 3.[2] Does the nonsmooth squeeze result of Corollary 4.1 hold more generally for any n, k?

REFERENCES

[1] J. M. Borwein and S. P. Fitzpatrick, *Duality Inequalities and Sandwiched Functions*, preprint, 1999.
[2] F. H. Clarke, *Optimization and Nonsmooth Analysis*, John Wiley and Sons, New York, 1983.
[3] F. H. Clarke and Yu. S. Ledyaev, *Mean value inequalities*, Proc. Amer. Math. Soc., 122 (1994), pp. 1075–1083.
[4] J. B. Hiriart-Urruty and C. Lemaréchal, *Convex Analysis and Minimization Algorithms* I, Grundlehren Math. Wiss., 305, Springer–Verlag, Berlin, 1993.
[5] J. B. Hiriart-Urruty and C. Lemaréchal, *Convex Analysis and Minimization Algorithms* II, Grundlehren Math. Wiss., 306, Springer–Verlag, Berlin, 1993.
[6] A. S. Lewis and D. Ralph, *A nonlinear duality result equivalent to the Clarke–Ldyaev mean value inequality*, Nonlinear Anal., 26 (1996), pp. 343–350.
[7] R. T. Rockafellar, *Convex Analysis*, Princeton University Press, Princeton, NJ, 1970.

[2] After this paper was submitted, this question was resolved in the affirmative in [1].

SIAM J. CONTROL OPTIM.
Vol. 38, No. 2, pp. 627–661

A BEHAVIORAL APPROACH TO THE POLE STRUCTURE OF ONE-DIMENSIONAL AND MULTIDIMENSIONAL LINEAR SYSTEMS*

J. WOOD†, U. OBERST‡, E. ROGERS†, AND D. H. OWENS§

Abstract. We use the tools of behavioral theory and commutative algebra to produce a new definition of a (finite) pole of a linear system. This definition agrees with the classical one and allows a direct dynamical interpretation. It also generalizes immediately to the case of a multidimensional (nD) system. We make a natural division of the poles into controllable and uncontrollable poles. When the behavior in question has latent variables, we make a further division into observable and unobservable poles. In the case of a one-dimensional (1D) state-space model, the uncontrollable and unobservable poles correspond, respectively, to the input and output decoupling zeros, whereas the observable controllable poles are the transmission poles.

Most of these definitions can be interpreted dynamically in both the 1D and nD cases, and some can be connected to properties of kernel representations. We also examine the connections between poles, transfer matrices, and their left and right matrix fraction descriptions (MFDs). We find behavioral results which correspond to the concepts that a controllable system is precisely one with no input decoupling zeros and an observable system is precisely one with no output decoupling zeros. We produce a decomposition of a behavior as the sum of subbehaviors associated with various poles. This is related to the integral representation theorem, which describes every system trajectory as a sum of integrals of polynomial exponential trajectories.

Key words. pole, decoupling zero, behavioral approach, multidimensional systems, characteristic variety, polynomial exponential function, associated primes, primary decomposition, integral representation

AMS subject classifications. 13C05, 13C12, 35B37, 93A99, 93B25, 93B55

PII. S036301299733213X

1. Introduction.

This paper is about the pole structure of one-dimensional (1D) and multidimensional (nD) linear systems. Although it is motivated by the need to establish a cohesive theory of poles for nD systems, many of the results in the paper give a new perspective to the classical 1D theory. Furthermore, this paper places all of the classical results in the new and growing behavioral framework [39, 40, 41]. We consider only linear systems with constant coefficients; our results apply to both continuous and discrete cases.

The systematic theory of poles (and zeros) of 1D systems originated in the work of Rosenbrock [33] and others. Since then, a number of papers have reviewed, extended, and refined the theory (e.g., [8, 23, 35]; see also the references at the end of this paper).

An nD system is one in which information propagates in two or more independent directions. Such systems arise as the solutions of partial differential or difference

*Received by the editors December 30, 1997; accepted for publication (in revised form) June 14, 1999; published electronically February 9, 2000. The research of the first author was supported by EPSRC grant GR/K 18504 and by the Royal Society. The research of the second author was supported in part by Austrian FWF-project P11431-MAT.

http://www.siam.org/journals/sicon/38-2/33213.html

†ISIS Group, Department of Electronics and Computer Science, University of Southampton, Southampton SO17 1BJ, UK (jjw@ecs.soton.ac.uk, etar@ecs.soton.ac.uk). Jeffrey Wood is a Royal Society University Research Fellow.

‡Institut für Mathematik, Naturwissenschaftliche Fakultat, Universität Innsbruck, Technikerstraße 25/7, A-6020 Innsbruck, Austria (ulrich.oberst@uibk.ac.at).

§Department of Automatic Control and Systems Engineering, University of Sheffield, Mappin Street, Sheffield S1 3JD, UK (D.H.Owens@sheffield.ac.uk).

equations. Shankar and Sule [36, 37] have defined poles for SISO nD systems, using the ring of stable causal rational functions. As yet, to our knowledge, no theory of pole structure or even a definition of a pole has been proposed for nD MIMO systems. The primary aim of this paper is to fill that gap. Our secondary aim is to extend the behavioral theory of both 1D and nD systems by introducing into it concepts concerned with pole structure. Thirdly, we intend to shed new light on the existing 1D theory, both through our use of behaviors and by identifying correspondences between system-theoretic concepts and algebraic ones.

In accordance with the behavioral paradigm, we define poles (decoupling zeros, etc.) to be objects associated with a system's behavior, rather than with any given representation of it. One advantage of such definitions is that they are all intrinsic, as is also the case with the recent definitions of Bourlès and Fliess [4] and of Pommaret and Quadrat [31]. For example, the input decoupling zeros are not defined as the zeros eliminated by a reduction procedure as in [33], but are instead defined directly.

The algebraic definitions of Bourlès and Fliess [4] have been extended to the case of nD systems with variable coefficients by Pommaret and Quadrat [31]. Our definitions are equivalent to these in the 1D/ nD constant coefficients case, the relationship being given by module duality, which translates, for example, the definition of a pole given in terms of the finitely generated system module into a corresponding one in terms of the system's behavior. Our definitions are not directly related to those of Shankar and Sule [36].

One of our principal tools in this paper is the duality theory in [24] between finitely generated modules over a polynomial ring and nD behaviors described by linear partial differential or difference equations with constant coefficients. Consequently, we use a considerable amount of commutative algebra; see, for example, [3, 5, 7, 10] for background in this field.

The paper is structured as follows. We begin with an overview of relevant concepts from behavioral theory, with special emphasis on module duality (see section 2). Section 3 collects some results on nD behavioral theory which we need later in the paper.

In section 4 we review the concepts of exponential trajectories and the characteristic variety. Exponential trajectories are central to any understanding of system poles, and the characteristic variety is a geometric object which captures the essential structure of a behavior as regards its exponential trajectories. Based on these ideas, we define the poles of an nD system in section 5. We also define the concepts of controllable and uncontrollable poles. The controllable poles have properties which relate strongly to those of the system's transfer matrix, whereas the uncontrollable poles play the role of input decoupling zeros.

Section 6 considers the case where the behavior in question is a latent variable representation (e.g., a state-input-output behavior). We divide the system poles into observable and unobservable poles, which have natural interpretations. This leads naturally to a further breakdown in terms of controllable observable poles, etc., all of which correspond naturally to known sets of poles in the classical 1D context.

Finally, in section 7 we consider a decomposition which is dual to the standard algebraic concept of the primary decomposition of a module. This "polar decomposition" expresses the zero-input behavior of a system as the finite sum of subbehaviors associated with the distinct poles of the system. Using a similar decomposition we can write the system behavior itself as the sum of the controllable part and a finite set of subbehaviors associated with the system's uncontrollable poles (input decou-

pling zeros). This latter decomposition is a refinement of the well-known controllable-autonomous decomposition. Using the little-known but important integral representation theorem, it is possible to write any zero-input trajectory as a sum of integrals of polynomial exponential trajectories. In the case of a finite-dimensional behavior, every zero-input trajectory can be written as a sum of polynomial-weighted exponential trajectories with frequencies corresponding to the system poles, as in the 1D case.

We leave a number of important issues untouched, for example, the definition of multiplicity of a pole, the definition of an infinite pole, and the whole question of zeros. However, the basic structures which we uncover in this work should prove a good foundation for further developments of this type.

2. Behaviors, modules, and duality. This section provides an overview of behavioral theory and the duality between finitely generated modules and behaviors given by partial differential/difference equations.

2.1. The behavioral approach. The behavioral approach to 1D systems is due to Willems [39, 40, 41], and centers on the concept of the system's behavior, which is the set of associated trajectories. Formally, we define a *system* as a triple $(\mathcal{A}, q, \mathcal{B})$, where $\mathcal{A}$ is a set called the *signal space*, $q \in \mathbb{Z}^+$ is the number of components, and $\mathcal{B} \subseteq \mathcal{A}^q$ is the *behavior*. The elements of $\mathcal{B}$ are called *trajectories*. We will make little distinction between a system and its behavior. Note that the standard behavioral definition of a system as a triple $(T, W, \mathcal{B})$ (where T is the signal domain, W the signal value set, and $\mathcal{B} \subseteq W^T$) does not cover the important case of distributions.

In all cases of interest (as listed below), $\mathcal{A}$ will have the structure of a module over a suitable ring of differential or difference operators; in particular, $\mathcal{A}$ is always a vector space over a field k, which is taken to be $\mathbb{R}$ or $\mathbb{C}$. Specifically, throughout this paper $\mathcal{A}$ denotes one of the following signal spaces, with a module structure as described.

1. The discrete signal space $\mathcal{A} = k^{\mathbb{N}^n}$, $k = \mathbb{R}, \mathbb{C}$, which is a module over the polynomial ring $\mathcal{R} = k[\underline{\mathbf{z}}] = k[z_1, \ldots, z_n]$, where the action of z_i on a trajectory $w \in \mathcal{A}$ is taken to be the shift operator σ_i, defined by

$$(\sigma_i w)(t_1, \ldots, t_n) \;:=\; w(t_1, \ldots, t_{i-1}, t_i + 1, t_{i+1}, \ldots, t_n). \tag{2.1}$$

By extension, any element of $\mathcal{R}$ has an action on $\mathcal{A}$

$$\text{for all } p \in k[\underline{\mathbf{z}}], w \in \mathcal{A}, \qquad p(z_1, \ldots, z_n)w \;:=\; p(\sigma_1, \ldots, \sigma_n)(w). \tag{2.2}$$

2. The discrete signal space $\mathcal{A} = k^{\mathbb{Z}^n}$, $k = \mathbb{R}, \mathbb{C}$, which is a module over the Laurent polynomial ring $\mathcal{R} = k[\underline{\mathbf{z}}, \underline{\mathbf{z}}^{-1}] = k[z_1, \ldots, z_n, z_1^{-1}, \ldots, z_n^{-1}]$, where z_i acts as the shift operator σ_i and z_i^{-1} as the inverse shift σ_i^{-1}.

3. The signal space $\mathcal{A} = \mathcal{C}^\infty(\mathbb{R}^n, k)$, $k = \mathbb{R}, \mathbb{C}$ of all k-valued $\mathcal{C}^\infty$ functions on $\mathbb{R}^n$, which is a module over $\mathcal{R} = k[\underline{\mathbf{z}}]$, where z_i acts as the partial derivative operator $\partial/\partial t_i$. This action is extended to $\mathcal{R}$ by

$$\text{for all } p \in k[\underline{\mathbf{z}}], w \in \mathcal{A}, \qquad p(z_1, \ldots, z_n)w \;:=\; p(\partial/\partial t_1, \ldots, \partial/\partial t_n)(w). \tag{2.3}$$

4. The signal space $\mathcal{A} = \mathcal{D}'(\mathbb{R}^n, k)$, $k = \mathbb{R}, \mathbb{C}$ of all k-valued distributions on $\mathbb{R}^n$, which is a module over $\mathcal{R} = k[\underline{\mathbf{z}}]$, where z_i again acts as the partial derivative operator $\partial/\partial t_i$.

Throughout the paper, we consider the behavior $\mathcal{B} \subseteq \mathcal{A}^q$ to be the solution space in $\mathcal{A}^q$ of a finite set of linear partial differential or difference equations in q dependent

variables with constant coefficients. Such behaviors will be called *differential behaviors* and *difference behaviors*, respectively. The term *autoregressive (AR) behavior* has formerly been used but is undesirable due to the inappropriate stochastic connotations.

The set of systems with differential/difference behaviors covers all linear time-invariant systems dealt with by the classical 1D state-space framework and its nD analogues. Differential and difference behaviors are, in particular, linear and shift-invariant and in the continuous case are also closed under partial differentiation. In each case any differential or difference behavior $\mathcal{B}$ is therefore a submodule of $\mathcal{A}^q$; the ring action is the componentwise application of (2.2) or (2.3).

It is convenient to describe differential/difference behaviors using polynomial matrices (Laurent polynomial matrices when $\mathcal{A} = k^{\mathbb{Z}^n}$). Thus let $E \in \mathcal{R}^{g,q}$; then the behavior described by E is the submodule of $\mathcal{A}^q$ consisting of all w satisfying $Ew = 0$, where the meaning of Ew is given by the action of the ring $\mathcal{R}$ on $\mathcal{A}$. For example, the three-dimensional (3D) difference behavior $\mathcal{B}$ in 2 variables described by the equations

$$w_1(t_1, t_2, t_3) - w_1(t_1, t_2 + 1, t_3 + 1) + 2w_2(t_1 + 1, t_2, t_3) = 0,$$
$$w_1(t_1 + 2, t_2, t_3 + 1) - w_2(t_1, t_2, t_3) + 3w_2(t_1, t_2 + 1, t_3) = 0$$

is given by the polynomial matrix

$$E \;=\; \begin{pmatrix} 1 - z_2 z_3 & 2z_1 \\ z_1^2 z_3 & -1 + 3z_2 \end{pmatrix}.$$

We say that E is a *kernel representation* of $\mathcal{B}$, and we write $\mathcal{B} = \ker_{\mathcal{A}} E$. For any matrix $E \in \mathcal{R}^{g,q}$, we also define the following modules:[1]

$$\operatorname{Ker}_{\mathcal{R}} E := \{v \in \mathcal{R}^{1,g} \;\mid\; vE = 0\}, \tag{2.4}$$
$$\operatorname{Im}_{\mathcal{R}} E := \{v \in \mathcal{R}^{1,q} \;\mid\; v = xE \text{ for some } x \in \mathcal{R}^{1,g}\}, \tag{2.5}$$
$$\operatorname{Coker}_{\mathcal{R}} E := \mathcal{R}^{1,q} / \operatorname{Im}_{\mathcal{R}} E, \tag{2.6}$$
$$\operatorname{im}_{\mathcal{A}} E := \{w \in \mathcal{A}^g \;\mid\; w = El \text{ for some } l \in \mathcal{A}^q\}. \tag{2.7}$$

Note the different subscripts used to denote different ring actions. Also, the modules $\operatorname{Ker}_{\mathcal{R}} E$, $\operatorname{Im}_{\mathcal{R}} E$, $\operatorname{Coker}_{\mathcal{R}} E$ are defined with respect to a left action of E, whereas $\ker_{\mathcal{A}} E$ and $\operatorname{im}_{\mathcal{A}} E$ are defined with respect to a right action.

2.2. Module duality. In this section we describe the duality between finitely generated $\mathcal{R}$-modules and differential/difference behaviors [24]. For the purposes of the next definition and the discussion immediately following only, the space $\mathcal{A}$ can be thought of as an arbitrary module.

DEFINITION 2.1. *Let M be a finitely generated $\mathcal{R}$-module and $\mathcal{A}$ an arbitrary $\mathcal{R}$-module. The* dual of M *with respect to $\mathcal{A}$, denoted $D(M)$, is defined by*

$$D(M) \;:=\; \operatorname{Hom}_{\mathcal{R}}(M, \mathcal{A}). \tag{2.8}$$

If $\phi : M \mapsto N$ is a morphism of finitely generated $\mathcal{R}$-modules, then the dual map $D(\phi) : D(N) \mapsto D(M)$ is given by

$$\text{for all } v \in D(N), \qquad (D(\phi))(v) \;:=\; v \circ \phi. \tag{2.9}$$

[1]Note that the modules $\operatorname{Ker}_{\mathcal{R}} E$, $\operatorname{Im}_{\mathcal{R}} E$, $\operatorname{Coker}_{\mathcal{R}} E$ are defined using row vectors and do not therefore require transposition of E, contrary to the first author's previous practice.

The action of $\mathcal{R}$ on the dual module $D(M)$ is given by

$$\text{for all } r \in \mathcal{R}, w \in D(M), \qquad (rw)(m) \;:=\; r \cdot w(m) \;=\; w(rm) \tag{2.10}$$

for all $m \in M$.

Note for any $q \in \mathbb{Z}^+$ that $D(\mathcal{R}^{1,q}) \cong \mathcal{A}^q$. Furthermore, it turns out that we have the following theorem.

THEOREM 2.2 (see [24, pp. 19–21]). *Differential/difference behaviors are precisely the dual modules of finitely generated $\mathcal{R}$-modules. Specifically, if $\mathcal{B} = \ker_{\mathcal{A}} E$, then $\mathcal{B} = D(M)$, where M is the finitely generated module* $\operatorname{Coker}_{\mathcal{R}} E$.

It is not hard to show [24, Cor. 2.57] that the dual action of a polynomial matrix map $E : \mathcal{R}^{1,g} \mapsto \mathcal{R}^{1,q}$, $v \mapsto vE$ is the map $E : \mathcal{A}^q \mapsto \mathcal{A}^g$, $w \mapsto Ew$ given by the ring action on $\mathcal{A}$ ((2.2) or (2.3) in our cases of interest).

DEFINITION 2.3. *A* complex *of modules is a set of maps $\phi_i : F_i \mapsto F_{i-1}$, $i \in \mathbb{Z}$, where $\phi_i \circ \phi_{i+1} = 0$ for all i. We represent a complex by a diagram of the form*

$$\cdots \xrightarrow{\phi_{i+2}} F_{i+1} \xrightarrow{\phi_{i+1}} F_i \xrightarrow{\phi_i} F_{i-1} \xrightarrow{\phi_{i-1}} \cdots.$$

The complex is said to be exact *and is called an* exact sequence*, if for every i we have that* $\ker \phi_i = \operatorname{im} \phi_{i+1}$.

Exact sequences are a compact way of expressing structural relationships. Exactness of the following two sequences express the injectivity and surjectivity, respectively, of the map ϕ:

$$0 \longrightarrow M \xrightarrow{\phi} N \qquad\qquad M \xrightarrow{\phi} N \longrightarrow 0.$$

The next exact sequence expresses that K is (isomorphic to) the kernel of ϕ and L is (isomorphic to) the cokernel:

$$0 \longrightarrow K \xrightarrow{\iota} M \xrightarrow{\phi} N \xrightarrow{\rho} L \longrightarrow 0.$$

Note that to any complex of modules, there is a corresponding dual complex, obtained by reversing the arrows and dualizing each module and each map, and conversely. The following crucial result shows that, in our cases of interest, the same can be said of exact sequences.

THEOREM 2.4 (see [24, Thm. 2.54, Cor. 2.43, 2.46, 2.47]). *Each module $\mathcal{A}$ listed in section* 2.1 *is an injective cogenerator of the category of $\mathcal{R}$-modules. This signifies that duality is contravariant and faithfully exact. In other words, given a complex of modules*

$$\cdots \xrightarrow{\phi_{i+2}} F_{i+1} \xrightarrow{\phi_{i+1}} F_i \xrightarrow{\phi_i} F_{i-1} \xrightarrow{\phi_{i-1}} \cdots \tag{2.11}$$

and its dual complex

$$\cdots \xrightarrow{D(\phi_{i-1})} D(F_{i-1}) \xrightarrow{D(\phi_i)} D(F_i) \xrightarrow{D(\phi_{i+1})} D(F_{i+1}) \xrightarrow{D(\phi_{i+2})} \cdots, \tag{2.12}$$

we have that (2.11) *is exact if and only if* (2.12) *is exact. In consequence, if $\mathcal{B} = \ker_{\mathcal{A}} E \subseteq \mathcal{A}^q$, then the* orthogonal module $\mathcal{B}^{\perp} \subseteq \mathcal{R}^{1,q}$ *(the module of all equations satisfied by the system $\mathcal{B}$) is equal to* $\operatorname{Im}_{\mathcal{R}} E$.

The module duality gives rise to a lattice duality $\mathcal{B} \mapsto \mathcal{B}^\perp$ between differential/difference behaviors and submodules of $\mathcal{R}^{1,q}$. This duality takes summation to intersection and vice versa [24, sect. 2.22], [42, Lem. 1]. Note also that $\mathcal{B} = D(M)$, where $M = \mathcal{R}^{1,q}/\mathcal{B}^\perp$. For any $\mathcal{R}$-matrix E it can easily be seen that $(\mathrm{im}_{\mathcal{A}}\, E)^\perp = \mathrm{Ker}_{\mathcal{R}}\, E$.

Now if $\mathcal{B} = \ker_{\mathcal{A}} E$ and $\mathcal{B}' = \ker_{\mathcal{A}} E$ are two behaviors with the same number of system variables, then $\mathcal{B}' \subseteq \mathcal{B}$ if and only if there exists an $\mathcal{R}$-matrix L with $E = LE'$ [24, sect. 2.61]. Another important consequence of Theorem 2.4 is that the dual of a submodule is a factor module of the dual, and vice versa. More precisely, given a finitely generated $\mathcal{R}$-module M and a submodule N, we have an exact sequence

$$0 \longrightarrow N \longrightarrow M \longrightarrow M/N \longrightarrow 0. \tag{2.13}$$

This dualizes to an exact sequence

$$0 \longrightarrow D(M/N) \longrightarrow D(M) \longrightarrow D(N) \longrightarrow 0, \tag{2.14}$$

which we can interpret as meaning that $D(M/N)$ is a submodule of $D(M)$, with factor module $D(N)$. Indeed, we identify $D(M/N)$ with the set of functions of $D(M)$ which are zero on N.

An immediate application of the duality theorem is the following [24, pp. 24–25].

LEMMA 2.5. *If $E \in \mathcal{R}^{g,q}$, then $\mathrm{im}_{\mathcal{A}}\, E = \mathcal{A}^g$ if and only if E has full row rank.*

Proof. The map $E : \mathcal{R}^{1,g} \mapsto \mathcal{R}^{1,q}$, $v \mapsto vE$ is monic if and only if the dual map $E : \mathcal{A}^q \mapsto \mathcal{A}^g$, $w \mapsto Ew$ is epic. ∎

We investigate some of the links between the theory of differential/difference behaviors and commutative algebra resulting from this duality in section 3.

Henceforth, $\mathcal{B} = \ker_{\mathcal{A}} E$ denotes a differential or difference behavior with kernel representation E, which is contained in $\mathcal{A}^q$ for one of the signal spaces $\mathcal{A}$ listed in section 2.1, and $M = \mathrm{Coker}_{\mathcal{R}}\, E$ denotes the finitely generated module to which $\mathcal{B}$ is dual:

$$E \in \mathcal{R}^{g,q}, \qquad M = \mathrm{Coker}_{\mathcal{R}}\, E, \qquad \mathcal{B} = D(M) = \ker_{\mathcal{A}} E \ \subseteq\ \mathcal{A}^q. \tag{2.15}$$

A similar behavioral and/or module-theoretic approach can be taken for linear systems with variable coefficients, as shown in the 1D case, for example, in [12, 14, 16, 34], using noncommutative techniques. The algebraic approach has been extended to nD variable coefficient systems by Pommaret and Quadrat [30, 31]. Recent work has indicated that it may be possible to use behavioral tools in the analysis of nonlinear systems [47].

Recall that the annihilator of a module M, denoted $\mathrm{ann}\, M$, is the ideal of all $r \in \mathcal{R}$ such that $rM = 0$. We denote the set of all elements of $\mathcal{B}$ annihilated by an ideal J (i.e., all elements w such that $rw = 0$ for all $r \in J$) by $[\mathcal{B} : J]$. The following properties of duality will be useful.

LEMMA 2.6. *Let J be an ideal of $\mathcal{R}$ and M a finitely generated $\mathcal{R}$-module with dual $\mathcal{B}$; then*

$$D(M/JM) \ =\ [\mathcal{B} : J]. \tag{2.16}$$

Also, $\mathrm{ann}\, M = \mathrm{ann}\, \mathcal{B}$.

Proof. Equation (2.16) is given in [24, Cor. 2.101]. Considering the case $J = (r)$ for some $r \in \mathcal{R}\backslash 0$, we now find that $D(M/rM) = [\mathcal{B} : (r)]$. If r is in the annihilator of M, then $M/rM = M$, so $[\mathcal{B} : (r)] = \mathcal{B}$, i.e., r is in the annihilator of $\mathcal{B}$. The converse is similar. ∎

3. Controllable, observable, and autonomous systems. In this section we recount relevant results from [24, 42, 46] and other sources, some of which are extended here from discrete to continuous systems. These results concern the characterization of controllable systems, observable systems, and autonomous systems using algebraic tools.

3.1. Free variables. The following definition is adapted from [15].

DEFINITION 3.1. *Let $\mathcal{B} \subseteq \mathcal{A}^q$. The set of variables $\{w_i \mid i \in \Phi\}$ for some subset Φ of $\{1, \ldots, q\}$ is said to be a* set of free variables *if the mapping $\rho : \mathcal{A}^q \mapsto \mathcal{A}^\Phi$, which projects a trajectory onto the components of Φ, is epic when restricted to $\mathcal{B}$.*

The maximum size of a set of free variables is called the number of free variables *of $\mathcal{B}$ and is denoted by $m(\mathcal{B})$.*

Thus a set of variables indexed by Φ is free if the corresponding components can take any value in $\mathcal{A}^\Phi$. Given a kernel representation $\mathcal{B} = \ker_{\mathcal{A}} E$ of $\mathcal{B}$, it can be shown that the number of free variables is equal to $q - \operatorname{rank} E$ [24, Thm. 2.69]. It can also be shown that the number of free variables is additive [42, remarks following Def. 6], i.e., given an exact sequence

$$0 \longrightarrow \mathcal{B}_1 \longrightarrow \mathcal{B}_2 \longrightarrow \mathcal{B}_3 \longrightarrow 0$$

of differential/difference behaviors, we have $m(\mathcal{B}_2) = m(\mathcal{B}_1) + m(\mathcal{B}_3)$. Consequently, if $\mathcal{B}_1, \mathcal{B}_2 \subseteq \mathcal{A}^q$ for some q, then we have

$$m(\mathcal{B}_1) + m(\mathcal{B}_2) \;=\; m(\mathcal{B}_1 + \mathcal{B}_2) + m(\mathcal{B}_1 \cap \mathcal{B}_2). \tag{3.1}$$

3.2. Autonomous systems. In the behavioral framework, an autonomous discrete system is defined as follows [15, 42].

DEFINITION 3.2. *A difference behavior $\mathcal{B}$ with signal domain T is called* autonomous *if there exists $T_1 \subseteq T$ such that any trajectory of $\mathcal{B}$ is determined by its values on T_1, and also $T \backslash T_1$ contains an n-dimensional cone.*

The main characterization of such behaviors [42, Thm. 2] follows; similar results or special cases have also appeared in [15, 46].

THEOREM 3.3 (see [15, 42, 46]). *Let $\mathcal{B} = D(M)$ be a difference behavior. The following are equivalent:*

1. *$\mathcal{B}$ is autonomous.*
2. *$\mathcal{B}$ has no free variables.*
3. *For any E with $\mathcal{B} = \ker_{\mathcal{A}} E$, E has full column rank.*
4. *$\operatorname{ann} \mathcal{B} = \operatorname{ann} M \neq 0$; equivalently, M is a torsion module.*

In the continuous case, Pillai and Shankar [29] have defined an autonomous system as one with a characteristic variety which is not equal to $\mathbb{C}^n$ [29, Def. 4.2]. We will discuss characteristic varieties in section 4; for now, let us use the following definition, which is easily seen to be equivalent to this condition and to the conditions of Theorem 3.3.

DEFINITION 3.4. *A differential behavior is called* autonomous *if* $\operatorname{ann} \mathcal{B} \neq 0$.

Recall that a behavior is said to be *finite-dimensional* or *strongly autonomous* if and only if it is finite-dimensional as a vector space over k [15, 29]. Note that this has nothing to do with the traditional use of the term "finite-dimensional" in classical systems theory.

3.3. Minimal left annihilators and generalized factor primeness. We will also need the concepts of a minimal left annihilator [32, p. 24] and generalized factor primeness [24, Thm. 7.21], [46].

DEFINITION 3.5. *Let $\mathcal{R} = k[\underline{\mathbf{z}}]$ or $k[\underline{\mathbf{z}}, \underline{\mathbf{z}}^{-1}]$, and suppose that $E \in \mathcal{R}^{g,q}$. Then E is called a* minimal left annihilator *(of F) if there exists a matrix $F \in \mathcal{R}^{q,h}$ for some h for which the following conditions hold:*

1. *$EF = 0$.*
2. *For any E' satisfying $E'F = 0$, there exists L with $E' = LE$.*

The matrix E is said to be factor left prime in the generalized sense (GFLP)*, if the existence of a $\mathcal{R}$-matrix factorization $E = LE_1$ (L not necessarily square) with* rank $E =$ rank E_1 *implies the existence of an $\mathcal{R}$-matrix L' with $E_1 = L'E$.*

It is shown in [24, Lem. 2.27], [42, Lem. 7] that E is a minimal left annihilator of F if and only if $\ker_{\mathcal{A}} E = \operatorname{im}_{\mathcal{A}} F$, or equivalently, $\operatorname{Im}_{\mathcal{R}} E = \operatorname{Ker}_{\mathcal{R}} F$. Every matrix over $\mathcal{R}$ has a minimal left annihilator [24, Lem. 2.27], [42, Lem. 7]. A matrix over $\mathcal{R}$ is a minimal left annihilator if and only if it is GFLP [42, Lem. 10].

3.4. Controllable systems. Any differential/difference behavior admits certain input/output structures. In this paper, as is traditional in control theory, the "inputs" are assumed to be a maximal set of free variables, and in consequence the outputs have no freedom once the inputs are determined. The more general situation, in which an input/output structure is an arbitrary partitioning of the system variables and which may be better suited to certain applications, is not considered in this paper. In [31], Pommaret and Quadrat extend some of the algebraic theory of this paper to this general case.

DEFINITION 3.6 (see [24, Thm. 2.69], [32, Def. IV.8], [42, Def. 12]). *An* input/output structure *on the behavior $\mathcal{B}$ is a partitioning of the system variables $w = (u, y)$, such that the set of variables u is free and the* zero-input behavior $\mathcal{B}_{0,y}$*, defined by*

$$\mathcal{B}_{0,y} \;=\; \{(u, y) \in \mathcal{B} \;\mid\; u = 0\} \tag{3.2}$$

is autonomous.

Equivalently, we can consider a partitioning $E = (-Q\ P)$ of any kernel representation E of $\mathcal{B}$ (to within a permutation), where the columns of Q correspond to the input variables u and the columns of P to the output variables y, and we have the condition

$$\text{rank } E \;=\; \text{rank } P \;=\; \text{number of columns of } P.$$

It is easy to show that the number of inputs is equal to m, the number of free variables. In particular, the number of inputs and number of outputs of a behavior is independent of the input/output structure.

If $\mathcal{B}$ has a kernel representation $(-Q\ P)$, where P is the submatrix corresponding to the output components, then $\mathcal{B}_{0,y}$ is (trivially) isomorphic to $\ker_{\mathcal{A}} P$. We usually make no distinction between these two behaviors.

For a given admissible input/output structure, any behavior $\mathcal{B}$ has a unique *transfer (function) matrix* $G \in k(\underline{\mathbf{z}})^{p,m}$ characterized by the equation $PG = Q$; see [24, Thm. 2.69] and also [32, p. 75] for the discrete case. Collecting together all behaviors with a given input/output structure and the same transfer matrix with respect to that input/output structure, we obtain a transfer class. The transfer class turns out to be independent of the input/output structure and resulting transfer matrix, and the transfer classes partition the set of behaviors. Furthermore, each transfer class has a unique element which is minimal with respect to set inclusion [24, Thm. 7.21], [32, p. 76]. Such a behavior will be called the *minimal realization* of the transfer matrix. The minimal realization has properties which are closely linked to those of the transfer matrix, as we will see later in the paper.

The concept of a minimal realization is related to controllability, which in the behavioral context is defined as follows.

DEFINITION 3.7 (see [29, 32, 42, 44]). *A differential behavior $\mathcal{B}$ is* controllable *if, for any two open sets $T_1, T_2 \subseteq \mathbb{R}^n$ with disjoint closures and for any two trajectories $w^{(1)}, w^{(2)} \in \mathcal{B}$, there exists a $w \in \mathcal{B}$ such that*

$$w(\underline{t}) \;=\; \begin{cases} w^{(1)}(\underline{t}) & \text{if } \underline{t} \in T_1, \\ w^{(2)}(\underline{t}) & \text{if } \underline{t} \in T_2. \end{cases} \tag{3.3}$$

A difference behavior $\mathcal{B}$ with signal domain $T = \mathbb{Z}^n$ or $T = \mathbb{N}^n$ is controllable *if there exists a real number $\rho > 0$ such that for any sets $T_1, T_2 \subseteq T$ with $d(T_1, T_2) > \rho$, for any $\underline{b}_1, \underline{b}_2 \in T$, and any two trajectories $w^{(1)}, w^{(2)} \in \mathcal{B}$, there exists a $w \in \mathcal{B}$ such that*

$$w(\underline{t}) \;=\; \begin{cases} w^{(1)}(\underline{t} - \underline{b}_1) & \text{if } \underline{t} \in T_1 \text{ and } \underline{t} - \underline{b}_1 \in T, \\ w^{(2)}(\underline{t} - \underline{b}_2) & \text{if } \underline{t} \in T_2 \text{ and } \underline{t} - \underline{b}_2 \in T. \end{cases} \tag{3.4}$$

In the discrete case, we can take $\underline{b}_1 = 0$ without loss of generality, and for $T = \mathbb{Z}^n$, we can also take $\underline{b}_2 = 0$. For both differential and difference behaviors, controllability expresses the idea of being able to join with a system trajectory any two system trajectories defined on regions which are sufficiently far apart.

The next result collects previous results in the literature on the characterization of nD behavioral controllability. In the 1D case, the equivalence of controllability and torsionfreeness is first due to Fliess [13], following an observation of Pommaret [30]. Special cases of the next result in the 1D and two-dimensional (2D) discrete cases are due, respectively, to Willems [40] and to Rocha [32].

THEOREM 3.8. *Let $\mathcal{B} = D(M) = \ker_{\mathcal{A}} E$. The following are equivalent:*

1. *$\mathcal{B}$ is controllable.*
2. *$\mathcal{B}$ is minimal in its transfer class.*
3. *$\mathcal{B}$ has an image representation.*
4. *E is GFLP.*
5. *$\mathcal{B}$ is a divisible module.*
6. *M is a torsionfree module.*
7. *$\mathcal{B}$ has no proper differential/difference subbehaviors with the same number of free variables.*

Proof. The equivalence of Parts 1 and 3 is given in [29, Prop. 3.4, Thm. 3.9] for differential behaviors, in [42, Thm. 5] for difference behaviors on $\mathbb{Z}^n$, and in [44] for general difference behaviors. The equivalence of Parts 3 and 4 is given in [42, Lem. 10] and also [46]. The equivalence of Parts 3 and 6 has been given in [30, Prop. VII.A.10], and the equivalence of Parts 2, 4, and 6 in [24, Thm. 7.21]. The equivalence of Parts 5 and 6 can be proved easily using the exactness of duality. Condition 7 is a direct interpretation of condition 4, as discussed following Definition 10 in [42]. □

Note that any differential/ difference behavior has a "controllable-autonomous decomposition" $\mathcal{B} = \mathcal{B}^c + \mathcal{B}^a$ [42, Thm. 7]; the argument given in that paper applies equally well to continuous systems. In this decomposition, the controllable behavior $\mathcal{B}^c \subseteq \mathcal{B}$ is uniquely determined as the minimal element of the transfer class of $\mathcal{B}$; henceforth we call this behavior the *controllable part* of $\mathcal{B}$. It is shown in [24, Thm. 7.21] that if $\mathcal{B} = D(M)$, then

$$\mathcal{B}^c \;=\; D(M/tM), \tag{3.5}$$

where tM denotes the torsion submodule of M. We will also find the behavior $\mathcal{B}/\mathcal{B}^c$ interesting. This behavior is dual to tM, as can be seen from (3.5) using basic duality principles (see (2.13) and (2.14)). Given a representation E^c for $\mathcal{B}^c$, we can see by considering the restriction of the map E^c to $\mathcal{B}$ that the behavior $E^c\mathcal{B}$ is naturally isomorphic to $\mathcal{B}/\mathcal{B}^c$.

The controllable-autonomous decomposition as defined above is in fact slightly more general (and less rich in structure) for a 1D system than as originally defined for 1D systems (see, e.g., [41]). This is because we do not require that the sum should be direct.

The controllable part of $\mathcal{B}$ has the same input/output structures as $\mathcal{B}$ itself (inevitably, since it can be defined with respect to any such input/output structure via the transfer matrix G). As we will see later in the paper, the zero-input behavior of the controllable part, $(\mathcal{B}^c)_{0,y}$, has an interesting structure. It has already been shown in [24, Cor. 7.29] that the corresponding orthogonal module has the following form:

$$((\mathcal{B}^c)_{0,y})^{\perp} \;=\; \{v \in \mathcal{R}^{1,p} \mid vG \in \mathcal{R}^{1,m}\}, \tag{3.6}$$

where we treat $(\mathcal{B}^c)_{0,y}$ as a subbehavior of $\mathcal{A}^p$ rather than of $\mathcal{A}^q$, p being the number of outputs.

The next result has not been explicitly proved in the literature before.

LEMMA 3.9. *Let $\mathcal{B}$ be a behavior with zero-input behavior $\mathcal{B}_{0,y}$ according to some input/output structure. Then $\mathcal{B}_{0,y}$ is an autonomous part of $\mathcal{B}$ in a controllable-autonomous decomposition.*

Proof. For any $(u, y) \in \mathcal{B}$, since the variables u are free in $\mathcal{B}^c$, there must exist a y^c with $(u, y^c) \in \mathcal{B}^c$. As $(0, y - y^c) \in \mathcal{B}_{0,y}$, we have the required decomposition $(u, y) = (u, y^c) + (0, y - y^c)$. □

DEFINITION 3.10. *A matrix $E \in \mathcal{R}^{g,q}$ is called* zero left prime *if its gth order minors generate $\mathcal{R}$. A behavior $\mathcal{B}$ is called* strongly controllable *if it has a zero left prime kernel representation.*

It is a straightforward consequence that $\mathcal{B} = D(M)$ is strongly controllable if and only if M is free. Definition 3.10 is not the definition originally given for strong controllability [32], which has not yet been usefully extended from the 2D to nD or to the continuous case.

3.5. Observable systems. We use the behavioral concept of observability [32, Defn. 19], [40]: given a behavior $\mathcal{B}_{l,w}$ involving two sets of variables l and w, the variables l are said to be *observable* from the variables w if for any two trajectories $(l_1, w_1), (l_2, w_2) \in \mathcal{B}_{l,w}$, it holds that $w_1 = w_2$ implies $l_1 = l_2$. When the variables l are latent variables in a latent variable description of the behavior $\mathcal{B}_w$ (see section 6.1), we say that the behavior $\mathcal{B}_{l,w}$ is *observable* if the latent variables l are observable from the manifest ones w. We have that $\mathcal{B}_{l,w}$ is observable if and only if $\mathcal{B}_{l,0} = 0$.

4. Exponential trajectories and the characteristic variety. We commence our main discussion by explaining the notion of the characteristic variety of a behavior, which is a geometric object describing the exponential trajectories contained in the behavior. To motivate this analysis, we first review the classical interpretation of a pole of a 1D continuous system (see, e.g., [8, 20]).

The point a is a pole of the system Σ if, when zero input $u(t)$ is fed to the system, there exists a nonzero initial condition $x(0)$ such that the resulting state trajectory has the form $x(t) = x(0)e^{at}$ [20]. This is sometimes rephrased by allowing $u(t)$ to be a finite sum of generalized delta functions which "kick the system" into the appropriate

initial condition $x(0)$. Note that as $y(t)$ is determined linearly by $x(t)$ ($u(t)$ being zero), $y(t)$ must also be of the form $y(0)e^{at}$.

This *trajectory interpretation* of a pole can be formalized in behavioral theory as follows. Let $\mathcal{B}_{x,u,y}$ denote the state-input-output behavior of our given state space representation of the system. Then a is a pole of the system (of $\mathcal{B}_{x,u,y}$) if there exists a nonzero trajectory of the form v_0e^{at} in $\mathcal{B}_{x,0,y}$, the set of zero-input trajectories. Exponential trajectories therefore play a crucial role in describing the poles of a system, indeed they are the motive for considering poles.

4.1. Polynomial exponential trajectories. The algebraic significance of a trajectory such as $w(t) = v_0e^{at}$ is that it is killed by the differential operator $d/dt - a$. In the module formalism we are using, $w(t)$ is annihilated by the maximal ideal $(z - a)$. The characterization of trajectories annihilated by maximal ideals, or more generally by powers of maximal ideals, has been done in full detail in [25, Eqs. (1.25–26), (5.26–29), (6.6), (6.10)]. Such trajectories turn out to be exponential trajectories, or more generally, polynomial exponential trajectories, in a sense appropriate to the underlying signal space. These are precisely the trajectories in which we are interested in our study of poles.

The description of exponential trajectories in [25] can be reformulated as in the next two results.

THEOREM AND DEFINITION 4.1 (see [25]). *Consider a scalar trajectory $w(\underline{t})$ in $\mathcal{A}$, where $\mathcal{A}$ is one of the signal spaces listed in section 2.1, which is a module over the ring $\mathcal{R} = k[\underline{\mathbf{z}}]$ or $k[\underline{\mathbf{z}}, \underline{\mathbf{z}}^{-1}]$. Let $\underline{a} \in \mathbb{C}^n$ or $(\mathbb{C}\backslash 0)^n$, respectively, according to $\mathcal{R}$, and let $I(\underline{a})$ denote the maximal ideal of all elements of $\mathcal{R}$ which vanish at $\underline{a}$. (Note that we may have $\underline{a} \notin k^n$, but $I(\underline{a})$ is still meaningful.)*

Then $w(\underline{t})$ is annihilated by $I(\underline{a})$ if and only if it is of the following form (dependent on the signal space):

1. *For $\mathcal{A} = \mathbb{C}^{\mathbb{N}^n}$ or $\mathcal{A} = \mathbb{C}^{\mathbb{Z}^n}$,*

$$w(\underline{t}) \;=\; \alpha a_1^{t_1} \cdots a_n^{t_n}$$

for some $\alpha \in \mathbb{C}$.

2. *For $\mathcal{A} = \mathbb{R}^{\mathbb{N}^n}$, write $a_i = r_i e^{\imath\theta_i}$, $\theta_i := 0$ for $a_i = 0$, $i = 1, \ldots, n$;*

$$w(\underline{t}) \;=\; r_1^{t_1} \cdots r_n^{t_n}(\alpha \cos(\theta_1 t_1 + \cdots + \theta_n t_n) + \beta \sin(\theta_1 t_1 + \cdots + \theta_n t_n))$$

for some $\alpha, \beta \in \mathbb{R}$.

3. *For $\mathcal{A} = \mathbb{R}^{\mathbb{Z}^n}$, write $a_i = r_i e^{\imath\theta_i}$, $i = 1, \ldots, n$;*

$$w(\underline{t}) \;=\; r_1^{t_1} \cdots r_n^{t_n}(\alpha \cos(\theta_1 t_1 + \cdots + \theta_n t_n) + \beta \sin(\theta_1 t_1 + \cdots + \theta_n t_n))$$

for some $\alpha, \beta \in \mathbb{R}$.

4. *For $\mathcal{A} = \mathcal{C}^\infty(\mathbb{R}^n, \mathbb{C})$ or $\mathcal{A} = \mathcal{D}'(\mathbb{R}^n, \mathbb{C})$,*

$$w(\underline{t}) \;=\; \alpha e^{a_1 t_1 + \cdots + a_n t_n}$$

for some $\alpha \in \mathbb{C}$.

5. *For $\mathcal{A} = \mathcal{C}^\infty(\mathbb{R}^n, \mathbb{R})$ or $\mathcal{A} = \mathcal{D}'(\mathbb{R}^n, \mathbb{R})$, with $a_i = g_i + \imath h_i$, $i = 1, \ldots, n$,*

$$w(\underline{t}) \;=\; e^{g_1 t_1 + \cdots + g_n t_n}(\alpha \cos(h_1 t_1 + \cdots + h_n t_n) + \beta \sin(h_1 t_1 + \cdots + h_n t_n))$$

for some $\alpha, \beta \in \mathbb{R}$.

A trajectory of $\mathcal{A}^q$ is annihilated by $I(\underline{a})$ if and only if every component is of the above form. Such a trajectory is called an exponential trajectory with frequency $\underline{a}$.

The full exponential-type properties of a behavior can however only be captured by considering trajectories of a more general form, as described in the next result. A trajectory $w \in \mathcal{A}$ is called *locally finite* if the submodule $\mathcal{R} \cdot w$ generated by w is finite-dimensional over k [25]. The $\mathcal{R}$-submodule of $\mathcal{A}$ consisting of all locally finite trajectories is denoted by $\mathcal{A}_{\mathrm{lf}}$.

THEOREM AND DEFINITION 4.2 (see [25]). *Define $a, I(\underline{a})$ as above. A trajectory $w(\underline{t}) \in \mathcal{A}$ is annihilated by some power of $I(\underline{a})$ if and only if it is of the following form (dependent on the signal space):*

1. *For $\mathcal{A} = \mathbb{C}^{\mathbb{N}^n}$, write $S = \{i \in \{1, \dots, n\} \,:\, a_i \neq 0\}$;*

$$w(\underline{t}) \;=\; p(\underline{t}) \prod_{i \in S} a_i^{t_i}$$

for some $p(\underline{t})$ which is the componentwise product of a polynomial function $p^{(1)}(\underline{t})$ of the parameters $t_i, i \in S$ only, and a finitely supported function $p^{(2)}(\underline{t})$ of the parameters $t_j, j \notin S$ only.

2. *For $\mathcal{A} = \mathbb{C}^{\mathbb{Z}^n}$,*

$$w(\underline{t}) \;=\; p(\underline{t}) a_1^{t_1} \cdots a_n^{t_n}$$

for some $p(\underline{t}) \in \mathbb{C}[\underline{t}]$.

3. *For $\mathcal{A} = \mathbb{R}^{\mathbb{N}^n}$, write $S = \{i \in \{1, \dots, n\} \,:\, a_i \neq 0\}$, and, for $i \in S$, define $r_i, \theta_i \in \mathbb{R}$ by $a_i = r_i e^{\imath \theta_i}$. For $i \notin S$, set $\theta_i = 0$.*

$$w(\underline{t}) \;=\; p(\underline{t}) \left(\prod_{i \in S} r_i^{t_i} \right) (\alpha \cos(\theta_1 t_1 + \cdots + \theta_n t_n) + \beta \sin(\theta_1 t_1 + \cdots + \theta_n t_n))$$

for some $\alpha, \beta \in \mathbb{R}$ and some $p(\underline{t})$ which is the componentwise product of a real polynomial function $p^{(1)}(\underline{t})$ of the parameters $t_i, i \in S$ only, and a finitely supported real function $p^{(2)}(\underline{t})$ of the parameters $t_j, j \notin S$ only.

4. *For $\mathcal{A} = \mathbb{R}^{\mathbb{Z}^n}$, for $i = 1, \dots, n$, define $r_i, \theta_i \in \mathbb{R}$ by $a_i = r_i e^{\imath \theta_i}$, $i = 1 \dots, n$;*

$$w(\underline{t}) \;=\; p(\underline{t}) r_1^{t_1} \cdots r_n^{t_n} (\alpha \cos(\theta_1 t_1 + \cdots + \theta_n t_n) + \beta \sin(\theta_1 t_1 + \cdots + \theta_n t_n))$$

for some $\alpha, \beta \in \mathbb{R}$ and some $p(\underline{t}) \in \mathbb{R}[\underline{t}]$.

5. *For $\mathcal{A} = \mathcal{C}^\infty(\mathbb{R}^n, \mathbb{C})$ or $\mathcal{A} = \mathcal{D}'(\mathbb{R}^n, \mathbb{C})$,*

$$w(\underline{t}) \;=\; p(\underline{t}) e^{a_1 t_1 + \cdots + a_n t_n}$$

for some $p(\underline{t}) \in \mathbb{C}[\underline{t}]$.

6. *For $\mathcal{A} = \mathcal{C}^\infty(\mathbb{R}^n, \mathbb{R})$ or $\mathcal{A} = \mathcal{D}'(\mathbb{R}^n, \mathbb{R})$, write $a_i = g_i + \imath h_i$, $g_i, h_i \in \mathbb{R}$, $i = 1, \dots, n$;*

$$w(\underline{t}) \;=\; p(\underline{t}) e^{g_1 t_1 + \cdots + g_n t_n} (\alpha \cos(h_1 t_1 + \cdots + h_n t_n) + \beta \sin(h_1 t_1 + \cdots + h_n t_n))$$

for some $p(\underline{t}) \in \mathbb{R}[\underline{t}]$.

Moreover,

$$\mathcal{A}_{\mathrm{lf}} \;=\; \bigoplus_{\underline{a}} \left\{ w \in \mathcal{A} \;\middle|\; I(\underline{a})^l \text{ annihilates } w \text{ for some } l \in \mathbb{Z}^+ \right\}, \tag{4.1}$$

where for $k = \mathbb{C}$ the direct sum runs over all $\underline{a}$ in either $\mathbb{C}^n$ or $(\mathbb{C}\backslash 0)^n$, according to $\mathcal{R}$; for $k = \mathbb{R}$ the direct sum runs over a set of representatives of the one- or two-element subsets $\{\underline{a}, \overline{\underline{a}}\}, \underline{a} \in \mathbb{C}^n$. Here $\overline{\underline{a}}$ denotes the componentwise complex conjugate of $\underline{a}$.

The locally finite trajectories in $\mathcal{A}_{\mathrm{lf}}^q$ for any q are called polynomial exponential functions/trajectories.

Note that a finite-dimensional behavior consists entirely of polynomial exponential trajectories, as it contains $\mathcal{R}w$ for each trajectory w, and so all trajectories are locally finite.

As discussed in [25], the space $\mathcal{A}_{\mathrm{lf}}$ is itself an injective cogenerator of the category of $\mathcal{R}$-modules, i.e., Theorem 2.4 applies with $D(M) = \mathrm{Hom}_{\mathcal{R}}(M, \mathcal{A})$ replaced by $D_{\mathrm{lf}}(M) = \mathrm{Hom}_{\mathcal{R}}(M, \mathcal{A}_{\mathrm{lf}})$. This enables the vast majority of the duality theory of [24] to be applied to differential/difference behaviors in $\mathcal{A}_{\mathrm{lf}}^q$. In particular, the equation

$$(\mathcal{B} \cap \mathcal{A}_{\mathrm{lf}}^q)^{\perp} \;=\; \mathrm{Im}_{\mathcal{R}}\, E \;=\; \mathcal{B}^{\perp}$$

for any $\mathcal{B} = \ker_{\mathcal{A}} E$ shows that *the subbehavior $\mathcal{B} \cap \mathcal{A}_{\mathrm{lf}}^q$ of polynomial exponential trajectories determines $\mathcal{B}$ uniquely.* More specifically, in the case of a $\mathcal{C}^\infty$ differential behavior, Hörmander gives the following result [18, Thm. 7.6.14].

THEOREM 4.3. *Let $\mathcal{B}$ be a differential behavior with signal space $\mathcal{C}^\infty(\mathbb{R}^n, \mathbb{C})$. Then $\mathcal{B}$ is equal to the closure of the set of its polynomial exponential trajectories.*

Even in the general case, we have seen that a differential or difference behavior is described fully by those trajectories annihilated by some power of an ideal $I(\underline{a})$, $\underline{a} \in \mathbb{C}^n$. As any differential/difference behavior containing such a trajectory must also contain a trajectory annihilated by $I(\underline{a})$ itself, from section 4.2 to section 6 we will discuss only exponential trajectories. In section 7 we will look briefly at the integral representation theorem, which for a behavior with signal space $\mathcal{C}^\infty(\mathbb{R}^n, \mathbb{C})$ describes how any system trajectory can be written as an integral of polynomial exponential trajectories.

4.2. The characteristic variety. We now look at the concept of the characteristic variety, which contains the essential information on the exponential trajectories of a behavior.

Let $J \subseteq k[\underline{\mathbf{z}}]$, $k = \mathbb{R}, \mathbb{C}$. Then we define the variety $V(J)$ as

$$V(J) \;:=\; \{\underline{a} \in \mathbb{C}^n \;\mid\; p(\underline{a}) = 0 \text{ for all } p \in J\}. \tag{4.2}$$

Note that $V(J)$ is defined as a subset of $\mathbb{C}^n$ even when $k = \mathbb{R}$. In the case where the ring of interest is $\mathcal{R} = k[\underline{\mathbf{z}}, \underline{\mathbf{z}}^{-1}]$, the definition is the same, except that only points $\underline{a} \in (\mathbb{C}\backslash 0)^n$ are considered. Henceforth we write $\underline{a} \in \mathbb{C}^n$ ($(\mathbb{C}\backslash 0)^n$), on the understanding that the former applies except when working with the signal space $\mathcal{A} = k^{\mathbb{Z}^n}$.

The next result and definition are fundamental to the paper and come from the theory of PDEs; see [2, p. 200, p. 340] and also [28, p. 138–139], where the term *variety associated with the finitely generated module M* is used. The characteristic variety is also investigated in [30, 31].

THEOREM AND DEFINITION 4.4. *Let $\mathcal{B} = D(M)$ be a behavior with kernel representation E. Let $\underline{a}$ be a point in $\mathbb{C}^n$ ($(\mathbb{C}\backslash 0)^n$). The following are equivalent:*

1. *$\underline{a} \in V(\mathrm{ann}\, \mathcal{B}) = V(\mathrm{ann}\, M)$.*
2. *$E(\underline{a})$ has less than full column rank.*
3. *$\mathcal{B}$ contains a nonzero exponential trajectory w with frequency $\underline{a}$.*

If $\underline{a}$ satisfies these conditions, then it is called a characteristic point *of $\mathcal{B}$ or of M. The variety $V(\operatorname{ann}\mathcal{B})$ of all such points is called the* characteristic variety *of $\mathcal{B}$ or of M.*

Proof. Let g be the number of rows of E. By Lemma 2.6, duality preserves the annihilator, so $V(\operatorname{ann}\mathcal{B}) = V(\operatorname{ann} M)$. Write $I(\underline{a})$ for the set of all polynomials of $\mathcal{R}$ which vanish at the complex point $\underline{a}$. $I(\underline{a})$ is a maximal ideal of $\mathcal{R}$. The equivalence of the first two conditions follow since ann M and $\operatorname{Fitt}_0 M$, the ideal of qth order minors of E, are known to have the same radical [10, Prop. 20.6] and therefore vanish at the same points. Now we have that $(\mathcal{R}/I(\underline{a}))^{1,q}/(\mathcal{R}/I(\underline{a}))^{1,g}E(\underline{a}) = M/I(\underline{a})M$, and by Lemma 2.6 we also have

$$D(M/I(\underline{a})M) \;=\; [\mathcal{B} : I(\underline{a})],$$

the set of trajectories of $\mathcal{B}$ annihilated by $I(\underline{a})$. Hence $E(\underline{a})$ has less than full column rank if and only if $D(M/I(\underline{a})M) \neq 0$, or if and only if $\mathcal{B}$ contains a nonzero trajectory annihilated by $I(a)$. Finally, apply Theorem 4.1. □

The first condition of Theorem 4.4 has the advantage of being succinct, algebraic, and representation-independent. The second condition, which considers the points where the kernel representation has less than full column rank, is familiar from the classical theory of poles. The third condition gives a dynamic (trajectory) interpretation of the point, the details of which depend upon the signal space. The equivalence of the first two conditions of Theorem 4.4 is well known.

The characteristic variety is interesting only for autonomous behaviors, for it is different from $\mathbb{C}^n$ ($(\mathbb{C}\backslash 0)^n$) precisely when $\mathcal{B}$ is autonomous.

In fact, for a given characteristic point $\underline{a}$ of a behavior $\mathcal{B} = \ker_{\mathcal{A}} E$, it is possible to construct explicitly a nonzero exponential trajectory with frequency $\underline{a}$ in $\mathcal{B}$. Such a trajectory is given, for example, in the continuous complex case by $v_0 e^{a_1 t_1 + \cdots + a_n t_n}$, where v_0 is any nonzero element of the kernel of $E(\underline{a})$, acting on $\mathbb{C}^q$ on the right. Conversely, over the field $\mathbb{C}$ if $v_0 e^{a_1 t_1 + \cdots + a_n t_n} \in \mathcal{B}$, then $E(\underline{a})v_0 = 0$, because when working over $\mathbb{C}$ we can easily show for any exponential trajectory w that $E(\underline{\mathbf{z}})w = E(\underline{a})w$.

Note that the characteristic points are defined over the algebraic closure $\mathbb{C}$ of k. In the case $k = \mathbb{R}$, the subset of characteristic points in $\mathbb{R}^n$ or $(\mathbb{R}\backslash 0)^n$ correspond to exponential trajectories of the form $v_0 e^{a_1 t_1 + \cdots + a_n t_n}$ (continuous case) or $v_0 a_1^{t_1} \cdots a_n^{t_n}$ (discrete case). However, the information contained in this subset of the characteristic variety is insufficient to describe the exponential phenomena which the system may exhibit.

Example 4.5. Consider the 1D behavior over $\mathcal{C}^\infty(\mathbb{R}, \mathbb{R})$ given by the kernel representation:

$$\mathcal{B} \;=\; \ker_{\mathcal{A}} E, \qquad E \;=\; \begin{pmatrix} z-5 & -2 \\ z & z-1 \end{pmatrix}.$$

Now E loses rank nowhere in $\mathbb{R}$, for its determinant is $z^2 - 4z + 5$, which has no roots in $\mathbb{R}$. Over $\mathbb{C}$, we have the roots $2 + \imath$, $2 - \imath$, and indeed, $\mathcal{B}$ admits the trajectory

$$w(t) \;=\; \begin{pmatrix} e^{2t}(3\sin t - \cos t) \\ e^{2t}(3\cos t - 4\sin t) \end{pmatrix},$$

which is exponential with frequency $2 + \imath$. However, $\mathcal{B}$ has no trajectories of the form $w(t) = \alpha e^{at}$, $a \in \mathbb{R}$, and the exponentially increasing/decreasing trajectories can only be found by considering the characteristic points in $\mathbb{C}\backslash\mathbb{R}$.

4.3. Associated and coassociated primes. We will find it useful to consider objects which describe the structure of a behavior $\mathcal{B}$ in a similar but slightly more detailed manner than $\operatorname{ann}\mathcal{B}$ and its corresponding variety.

Recall that an $\mathcal{R}$-module M and a prime ideal J give rise to the *localization* $M_J = \{\frac{m}{s} \mid m \in M, s \in \mathcal{R}\backslash J\}$, which is a module over the local ring $R_J = \{\frac{r}{s} \mid r \in \mathcal{R}, s \in \mathcal{R}\backslash J\}$.

DEFINITION 4.6. *Let M be an $\mathcal{R}$-module. The* support $\operatorname{supp} M$ *of M is the set of prime ideals J such that $M_J \neq 0$. The set* $\operatorname{Ass} M$ *of* associated primes *of M is the set of prime ideals of $\mathcal{R}$ which are annihilators of elements of M. These will also be called the* coassociated primes *of $\mathcal{B} = D(M)$.*

The following standard result can be found in, e.g., [3, Prop. II.4.17, Thm. IV.1.2] or [10, Cor. 2.7, Thm. 3.1].

THEOREM 4.7. *Let M be a finitely generated nonzero $\mathcal{R}$-module. Then the support of M is the set of all prime ideals containing* $\operatorname{ann} M$*. The set* $\operatorname{Ass} M$ *is finite, nonempty, contained in* $\operatorname{supp} M$*, and contains all the minimal prime divisors of* $\operatorname{ann} M$*.*

From the first claim of Theorem 4.7, a point $\underline{a}$ is a characteristic point of $\mathcal{B} = D(M)$ if and only if $\operatorname{ann} M \subseteq I(\underline{a})$, or equivalently, if and only if $M_{I(\underline{a})} \neq 0$. Indeed, much of the theory in this paper can be expressed formally in the language of localization; we will however use it sparingly.

From the last claim of Theorem 4.7, we can write the characteristic variety of $\mathcal{B}$ as the union of the varieties $V(J)$, J a coassociated prime of $\mathcal{B}$. There is redundancy in this decomposition, as it is also possible to write the characteristic variety as the union of the varieties corresponding to the minimal prime divisors of $\operatorname{ann} M$, called the *minimal primes* of M, only. The remaining associated primes of M, called *embedded primes*, correspond to proper subvarieties of those corresponding to the minimal primes. It is however necessary to consider all associated primes, since this makes available the primary decomposition theorem, which will be applied in section 7. In general, the coassociated primes of a behavior contain more information than the characteristic variety.

The next standard result will prove invaluable in breaking down the set of poles into subsets (e.g., controllable poles) with certain properties.

LEMMA 4.8. *Let M, M', M'' denote finitely generated $\mathcal{R}$-modules which form an exact sequence*

$$0 \longrightarrow M' \longrightarrow M \longrightarrow M'' \longrightarrow 0.$$

Then $V(\operatorname{ann} M) = V(\operatorname{ann} M') \cup V(\operatorname{ann} M'')$ *and* $\operatorname{Ass} M' \subseteq \operatorname{Ass} M \subseteq (\operatorname{Ass} M') \cup (\operatorname{Ass} M'')$.

Proof. The first claim is found in [3, Chap. II, sect. 4.4] and the second in [3, Chap. IV, sect. 1.2] or [10, p. 92]. □

5. Controllable and uncontrollable poles. In this section we present our new definition of a pole of an nD system and provide a decomposition of the set of poles into *controllable* and *uncontrollable* poles. The controllable poles are poles of the controllable part of the system and are associated with the transfer matrix. The uncontrollable poles play the role of input-decoupling zeros.

5.1. Controllable and uncontrollable poles. In our motivational discussion at the beginning of section 4, we discussed the classical interpretation of a pole, as a frequency $\underline{a}$ for which the zero-input behavior contains a nonzero exponential

trajectory. We then discussed the characterization of such points $\underline{a}$ using the notion of the characteristic variety. We now use this notion to define a pole (point) in the obvious way. However, we do not yet consider states; these are introduced under the guise of latent variables in section 6. Note also the distinction between "poles," which are prime ideals, and "pole points," which are points in real or complex space.

DEFINITION 5.1. *Let $\mathcal{B} = D(M)$ be a behavior with a given input/output structure and controllable part $\mathcal{B}^c$.*

1. *The* pole variety*,* pole points*, and* poles *of $\mathcal{B}$ are defined to be the characteristic variety, characteristic points, and coassociated primes, respectively, of the zero-input behavior $\mathcal{B}_{0,y}$.*

2. *The* controllable pole variety*,* controllable pole points*, and* controllable poles *of $\mathcal{B}$ are defined to be the characteristic variety, characteristic points, and coassociated primes, respectively, of the behavior $(\mathcal{B}^c)_{0,y}$.*

3. *The* uncontrollable pole variety*,* uncontrollable pole points*, and* uncontrollable poles *of $\mathcal{B}$ are defined to be the characteristic variety, characteristic points, and coassociated primes, respectively, of the behavior $\mathcal{B}/\mathcal{B}^c$.*

These definitions are all new and are not trivially equivalent to any previously given definitions of which we are aware. The behaviors concerned in Definition 5.1, and the corresponding finitely generated modules, can be captured in a dual pair of commutative exact diagrams (5.1)–(5.2); compare with the diagrams on page 14 of [31].

$$
\begin{array}{ccccccccccc}
 & & & & & & 0 & & 0 & & \\
 & & & & & & \downarrow & & \downarrow & & \\
 & & & & & & F & = & F & & \\
 & & & & & & \downarrow & & \downarrow & & \\
(5.1) & 0 & \longrightarrow & tM & \longrightarrow & M & \longrightarrow & M/tM & \longrightarrow & 0 & \\
 & & & \| & & \downarrow & & \downarrow & & & \\
 & 0 & \longrightarrow & tM & \longrightarrow & M/F & \longrightarrow & M/(F \oplus tM) & \longrightarrow & 0 & \\
 & & & & & \downarrow & & \downarrow & & & \\
 & & & & & 0 & & 0 & & &
\end{array}
\qquad .
$$

In the diagram (5.1), F denotes a maximal free submodule of M, for example, that generated by the elements $e_i + \mathcal{B}^{\perp}$, where e_i is the ith natural basis vector in $\mathcal{R}^{1,q}$, for each index i corresponding to an input component. The dual of M/F is $\mathcal{B}_{0,y}$, the dual of M/tM is $\mathcal{B}^c$, as already discussed, and the dual of $M/(tM \oplus F)$ is $\mathcal{B}^c \cap \mathcal{B}_{0,y} = (\mathcal{B}^c)_{0,y}$. Diagram (5.1) dualizes to (5.2).

$$
\begin{array}{ccccccccc}
 & & & 0 & & 0 & & & \\
 & & & \downarrow & & \downarrow & & & \\
 & 0 & \longrightarrow & (\mathcal{B}^c)_{0,y} & \longrightarrow & \mathcal{B}_{0,y} & \longrightarrow & \mathcal{B}/\mathcal{B}^c & \longrightarrow & 0 \\
 & & & \downarrow & & \downarrow & & \| & & \\
(5.2) & 0 & \longrightarrow & \mathcal{B}^c & \longrightarrow & \mathcal{B} & \longrightarrow & \mathcal{B}/\mathcal{B}^c & \longrightarrow & 0 \\
 & & & \downarrow & & \downarrow & & & & \\
 & & & \mathcal{A}^m & = & \mathcal{A}^m & & & & \\
 & & & \downarrow & & \downarrow & & & & \\
 & & & 0 & & 0 & & & &
\end{array}
\qquad .
$$

Note that it is important not to confuse the zero-input behavior $(\mathcal{B}^c)_{0,y}$ with the behavior $(\mathcal{B}_{0,y})^c$. The latter is uninteresting as it is always zero, whereas the former in general is not, and we will henceforth write $\mathcal{B}^c_{0,y}$ for $(\mathcal{B}^c)_{0,y}$.

It is possible to regard the behavior $\mathcal{B}_{0,y}$ as the "pole module" of the system, though we feel that it would be more appropriate to give this name to the finitely generated module M/F to which $\mathcal{B}_{0,y}$ is dual. This terminology ties in loosely with that of Kalman [21], Conte and Perdon [6], and Wyman and Sain [45], and we could extend it to an "input decoupling zero module," etc., though there does not seem to be a strong link between the two sets of definitions. The significance of a "pole module" is that the pole points themselves are the points where the annihilator of the pole module vanishes. This interpretation is also possible in the earlier work of Kalman and others. The term "pole module," etc., has also been used by Bourlès and Fliess [4], where it is equivalent (for 1D systems) to the definition we have suggested; see also Pommaret and Quadrat [31].

One important consequence of our definitions is that the number of poles (and also the number of controllable poles and the number of uncontrollable poles) is finite. This follows from Theorem 4.7. Also, for any pole J, $V(J)$ is contained in the pole variety, and indeed the pole variety is the union of such $V(J)$'s, similarly for controllable and uncontrollable poles.

The following characterization of pole points is immediate from Theorem 4.4.

COROLLARY 5.2. *Let $\mathcal{B}$ be a behavior with a given input/output structure, and let $\underline{a}$ be a point in $\mathbb{C}^n$ $((\mathbb{C}\backslash 0)^n)$. Let $(-Q\ P)$ be a kernel representation of $\mathcal{B}$, where the submatrix P corresponds to the output variables. Then the following are equivalent:*

1. *$\underline{a}$ is a pole point of $\mathcal{B}$.*
2. *$P(\underline{a})$ has less than full column rank.*
3. *$\mathcal{B}_{0,y}$ contains a nonzero exponential trajectory of frequency $\underline{a}$.*

5.2. Controllable poles. The controllable poles (pole points, etc.) of $\mathcal{B}$ are poles (pole points, etc.) of the controllable part $\mathcal{B}^c$ of $\mathcal{B}$. Hence for any controllable pole point $\underline{a}$, there is an exponential trajectory $w(\underline{t}) \in \mathcal{B}^c_{0,y} \subseteq \mathcal{B}_{0,y}$ which can be concatenated with the zero trajectory in the sense of behavioral controllability, i.e., such a $w(\underline{t})$ can be controlled.

Recall that $\mathcal{B}^c$ is the unique minimal element of the transfer class of $\mathcal{B}$. Since there is a strong relationship between $\mathcal{B}^c$ and the transfer matrix G of $\mathcal{B}$, it is not surprising that the controllable pole points of a behavior are captured in the structure of its transfer matrix.

Not only the variety $V(\operatorname{ann} \mathcal{B}^c_{0,y})$ but also the ideal $\operatorname{ann} \mathcal{B}^c_{0,y}$ can be described via the transfer matrix, and this gives $\operatorname{ann} \mathcal{B}^c_{0,y}$ and also the corresponding coassociated primes a special structure.

THEOREM 5.3. *Let $\mathcal{B}$ be a behavior with a given input/output structure and transfer matrix G. Then we have*

$$\operatorname{ann} \mathcal{B}^c_{0,y} \;=\; \{r \in \mathcal{R} \;\mid\; rG \text{ is a polynomial matrix}\}. \tag{5.3}$$

In particular, $\operatorname{ann} \mathcal{B}^c_{0,y}$ *is principal, and indeed generated by the least common denominator of the entries of G. The controllable poles of $\mathcal{B}$ are also principal.*

Proof. Recall (3.6)

$$(\mathcal{B}^c_{0,y})^\perp \;=\; \{v \in \mathcal{R}^{1,p} \;\mid\; vG \in \mathcal{R}^{1,m}\}.$$

For any element v of $\mathcal{R}^{1,p}$, we therefore have

$$\operatorname{ann}\,(v + (\mathcal{B}^c_{0,y})^\perp) \;=\; \{r \in \mathcal{R} \;\mid\; (rv)G \in \mathcal{R}^{1,m}\}. \tag{5.4}$$

Taking the intersection of these sets for all $v \in \mathcal{R}^{1,p}$ gives us $\operatorname{ann} \mathcal{R}^{1,p}/(\mathcal{B}^c_{0,y})^\perp = \operatorname{ann} \mathcal{B}^c_{0,y}$, and thus we obtain (5.3). Furthermore, (5.4) also tells us that the annihilator of any $v + (\mathcal{B}^c_{0,y})^\perp$ is equal to (d), where d is the least common denominator of the entries of vG. In particular, each such annihilator is principal, including the coassociated primes of $\mathcal{B}^c_{0,y}$, which are the controllable poles. □

COROLLARY 5.4. *Let $\mathcal{B}$ be a behavior with a given input/output structure and transfer matrix G. Let $(-Q^c\ P^c)$ be a GFLP matrix such that $P^cG = Q^c$, e.g., P^c is square and $(P^c)^{-1}Q^c = G$ is a minor left coprime matrix fraction description of G, if such exists. Let $\underline{a} \in \mathbb{C}^n$ $((\mathbb{C}\backslash 0)^n)$. Then the following are equivalent:*

1. *$\underline{a}$ is a controllable pole point of $\mathcal{B}$.*
2. *$P^c(\underline{a})$ has less than full column rank.*
3. *The denominator of some entry of G vanishes at $\underline{a}$.*
4. *$\mathcal{B}^c_{0,y}$ contains a nonzero exponential trajectory with frequency $\underline{a}$.*

Proof. If $(-Q^c\ P^c)$ is GFLP and $P^cG = Q^c$, then $(-Q^c\ P^c)$ is a kernel representation of a controllable behavior with transfer matrix G, which is therefore the controllable part of $\mathcal{B}$. From [43, Cor. 7.9], any minor left prime matrix is GFLP, and therefore this situation particularly applies when $(P^c)^{-1}Q^c$ is a minor left coprime matrix fraction description of G.

P^c is then a kernel representation of $\mathcal{B}^c_{0,y}$ (up to trivial isomorphism), and the equivalence of 1, 2, and 4 now follows from Theorem 4.4. Condition 3 is immediate from (5.3) in Theorem 5.3. □

Note that the condition "the denominator of some entry of G vanishes at $\underline{a}$" can be expressed formally by the localization condition: $G \notin \mathcal{R}^{p,m}_{I(\underline{a})}$, where $I(\underline{a})$ is the maximal ideal corresponding to the point $\underline{a}$.

Corollary 5.4 also characterizes controllable pole points in terms of left matrix fraction descriptions (MFDs) of G; these must be minor left coprime for the characterization to apply, and not every transfer matrix has a minor left coprime MFD. The same characterization can be applied to right MFDs, since it is known that if $P^{-1}Q = \overline{Q}\overline{P}^{-1}$ are minor left coprime and minor right coprime MFDs of a transfer matrix, then $|P| = |\overline{P}|$ to within a unit [19, Thm. 2.12].

This correspondence between the controllable pole points and the transfer matrix is similar to the classical 1D characterization of transmission poles. However, transmission poles are defined with respect to a state-space representation, and so it is not accurate to identify controllable pole points with transmission poles. The nD behavioral equivalent of the classical transmission poles will be discussed in section 6.2.

5.3. Uncontrollable poles. The uncontrollable poles *are defined independently of any input/output structure*, i.e., a system has the same uncontrollable poles regardless of the input/output structure imposed. This is interesting only if we have a proper interpretation of an uncontrollable pole, but we will see that it corresponds in appropriate cases to the idea of an input decoupling zero.

The uncontrollable poles can be characterized as follows.

LEMMA 5.5. *Let $\mathcal{B} = D(M)$ be a behavior with a given input/output structure. The uncontrollable poles are the coassociated primes of*

$$\mathcal{B}/\mathcal{B}^c \quad \cong \quad \mathcal{B}_{0,y}/\mathcal{B}^c_{0,y} \quad \cong \quad D(tM). \tag{5.5}$$

The uncontrollable poles of $\mathcal{B}$ are precisely the nonzero coassociated primes of $\mathcal{B}$. Furthermore, a behavior is controllable if and only if it has no uncontrollable poles.

Proof. The duality $\mathcal{B}/\mathcal{B}^c \cong D(tM)$ is known (see (3.5) and the discussion following), and the isomorphism $\mathcal{B}/\mathcal{B}^c \cong \mathcal{B}_{0,y}/\mathcal{B}^c_{0,y}$ comes from the standard module isomorphism theorems; see also diagram (5.2). The uncontrollable poles of $\mathcal{B}$ are the associated primes of tM, and it is easy to see that these are the nonzero associated primes of M, i.e., the nonzero coassociated primes of $\mathcal{B}$. Finally, $\mathcal{B}$ has no uncontrollable poles if and only if tM has no associated primes, i.e., if and only if $tM = 0$, i.e., $\mathcal{B} = \mathcal{B}^c$. □

The last result of Lemma 5.5 is reminiscent of the result that a controllable state-space representation of a 1D system is precisely one with no input decoupling zeros. However it is not appropriate to identify the uncontrollable poles with input decoupling zeros, since the concept of an input decoupling zero only has meaning when inputs, outputs, and states are involved; we will develop this connection further in section 6.2.

We now consider the characterization of uncontrollable pole points in terms of properties of $\mathcal{B}$ and its representations.

THEOREM 5.6. *Let $\mathcal{B} = D(M)$ be a behavior with kernel representation E. Then*

1. *$\mathcal{B}$ is controllable if and only if it has no uncontrollable pole points.*

2. *Let $\underline{a} \in \mathbb{C}^n$ $((\mathbb{C}\backslash 0)^n)$, and write $I(\underline{a}) \subseteq \mathcal{R}$ for the ideal of polynomials vanishing at $\underline{a}$. We have that $\mathcal{B}\backslash\mathcal{B}^c$ contains a nonzero exponential trajectory with frequency $\underline{a}$:*

   $\Rightarrow$ *$\underline{a}$ is an uncontrollable pole point of $\mathcal{B}$*

   $\Rightarrow$ *E loses rank at $\underline{a}$, i.e.,* rank $E(\underline{a}) \lneq$ rank E, *or equivalently, the localization $M_{I(\underline{a})}$ is not free.*

3. *The statements in claim 2 are equivalent when $\mathcal{B}^c$ is strongly controllable, and, in particular, for a 1D system. More generally, if E^c is a kernel representation of $\mathcal{B}^c$, then the statements are equivalent for points $\underline{a}$ at which E^c does not lose rank. In particular, these statements are equivalent for points $\underline{a}$ which are not controllable pole points.*

4. *Suppose that $\underline{a}$ is not an uncontrollable pole point of $\mathcal{B}$. Then E loses rank at $\underline{a}$ if and only if E^c loses rank at $\underline{a}$.*

Proof.

1. $\mathcal{B}$ has no uncontrollable pole points if and only if ann $\mathcal{B}/\mathcal{B}^c = \mathcal{R}$, i.e., $\mathcal{B} = \mathcal{B}^c$.

2. Write $J = I(\underline{a})$, and suppose that $\mathcal{B}\backslash\mathcal{B}^c$ contains a nonzero exponential trajectory w with frequency $\underline{a}$. Then $Jw = 0 \in \mathcal{B}^c$, so $w + \mathcal{B}^c$ is an element of $\mathcal{B}/\mathcal{B}^c$ which is annihilated by J. By Theorem 4.4, $\underline{a}$ is a characteristic point of $\mathcal{B}/\mathcal{B}^c$, i.e., an uncontrollable pole point of $\mathcal{B}$. From this it follows that ann $tM \subseteq J$, or equivalently, by Theorem 4.7 that $t(M_J) = (tM)_J \neq 0$, i.e., M_J is not torsionfree. Hence M_J is not free, which by [24, Thm. 7.69] means that E loses rank at $\underline{a}$.

3. Now suppose that E loses rank at $\underline{a}$ but that E^c does not. Equivalently, $(M/tM)_J$ is free but M_J is not. Due to the exact sequence

$$0 \longrightarrow (tM)_J \longrightarrow M_J \longrightarrow (M/tM)_J \longrightarrow 0, \tag{5.6}$$

we must have $(tM)_J \neq 0$, and so by Theorem 4.7 ann $tM \subseteq J$, from which $\underline{a}$ is an uncontrollable pole point of $\mathcal{B}$. Next, if $\underline{a}$ is such a point, and furthermore E^c does not lose rank at $\underline{a}$, we again have the exact sequence (5.6). Since again $(M/tM)_J$ is free, (5.6) splits, and so we can tensor with $R_J/J_J = R/J$ to obtain the exact sequence

$$0 \longrightarrow tM/JtM \longrightarrow M/JM \longrightarrow M^c/JM^c \longrightarrow 0,$$

where M^c denotes M/tM and its dual

$$0 \longrightarrow [\mathcal{B}^c : J] \longrightarrow [\mathcal{B} : J] \xrightarrow{\rho} [\mathcal{B}/\mathcal{B}^c : J] \longrightarrow 0. \tag{5.7}$$

Since $\underline{a}$ is an uncontrollable pole point of $\mathcal{B}$, by Theorem 4.4 there exists a nonzero element v in $[\mathcal{B}/\mathcal{B}^c : J]$. From the exactness of (5.7), there must be an element $w \in [\mathcal{B} : J]$ with $\rho(w) = w + \mathcal{B}^c = v$. Since $v \neq 0$, we must have $w \notin \mathcal{B}^c$, and as J annihilates w, it is an exponential trajectory with frequency $\underline{a}$, as required. This proves the equivalence of the three conditions in claim 2 for points $\underline{a}$ at which E^c does not lose rank. If $E^c = (-Q^c \; P^c)$ is a decomposition of E^c with respect to an input/output structure, so that P^c is a kernel representation of $\mathcal{B}^c_{0,y}$, then we know that rank E^c = rank P^c. Now if $\underline{a}$ is not a controllable pole point of $\mathcal{B}$, so that by Corollary 5.4 P^c does not lose rank at $\underline{a}$, then E^c does not lose rank at $\underline{a}$, and the equivalence of the given statements applies.

If $\mathcal{B}^c$ is strongly controllable, then E^c is zero left prime, and so E^c loses rank nowhere. For a 1D system, $\mathcal{B}^c$ is always strongly controllable.

4. Again we have an exact sequence

$$0 \longrightarrow (tM)_J \longrightarrow M_J \longrightarrow (M/tM)_J \longrightarrow 0.$$

As in the proof of the previous claim, the condition that $\underline{a}$ is not an uncontrollable pole point of $\mathcal{B}$ means that $(tM)_J = 0$, from which $M_J \cong (M/tM)_J$. Hence M_J is free if and only if $(M/tM)_J$ is free. By [24, Thm. 7.69], an equivalent statement is that E loses rank at $\underline{a}$ if and only if E^c loses rank at $\underline{a}$. ∎

It is not true in general that an uncontrollable pole point of $\mathcal{B} = \ker_{\mathcal{A}} E$ is precisely a point where E loses rank. Consider the example

$$E \;=\; \begin{pmatrix} z_1 & 0 & z_2 \\ 0 & z_2 & z_3 \end{pmatrix}$$

and the corresponding behavior $\mathcal{B}$ with signal space $\mathcal{A} = \mathcal{C}^\infty(\mathbb{R}^3, \mathbb{C})$. The matrix E has a rank-loss point $(0, 0, a_3)$ for any $a_3 \in \mathbb{C}$, but it is minor left prime (i.e., the second-order minors have no common factor), and therefore GFLP by [43, Cor. 7.9]. $\mathcal{B}$ is therefore controllable by Theorem 3.8, and so by Theorem 5.6 it has no uncontrollable pole points.

Uncontrollable pole points can also be characterized in terms of other behaviors and representations.

THEOREM 5.7. *Let $\mathcal{B} = \ker_{\mathcal{A}} E$ be any behavior with controllable part $\mathcal{B}^c = \ker_{\mathcal{A}} E^c$. Let K be any polynomial matrix with $E = KE^c$, and let C be a minimal left annihilator of E^c; write*

$$L \;=\; \begin{pmatrix} K \\ C \end{pmatrix}.$$

Then L is a kernel representation of $E^c\mathcal{B} \cong \mathcal{B}/\mathcal{B}^c$, and the following are equivalent for any point $\underline{a} \in \mathbb{C}^n$ $((\mathbb{C}\backslash 0)^n)$:

1. *$\underline{a}$ is an uncontrollable pole point of $\mathcal{B}$.*
2. *$L(\underline{a})$ has less than full column rank.*
3. *$E^c\mathcal{B}$ contains a nonzero exponential trajectory with frequency $\underline{a}$.*

Proof. Given E, E^c, K, C, and L as described, it is easy to see that LE^c is a kernel representation of $\mathcal{B}$, and we also have $\ker_{\mathcal{A}} L \subseteq \operatorname{im}_{\mathcal{A}} E^c$. Now $w \in E^c\mathcal{B}$ implies

$Lw = 0$, and conversely $w \in \ker_{\mathcal{A}} L$ implies that $w = E^c v$ for some v, so $(LE^c)v = 0$, i.e., $v \in \mathcal{B}$. Hence $\ker_{\mathcal{A}} L = E^c\mathcal{B}$. As $E^c\mathcal{B} \cong \mathcal{B}/\mathcal{B}^c$, the uncontrollable pole points are the characteristic points of $E^c\mathcal{B}$, and the remaining assertions follow on applying Theorem 4.4. ∎

5.4. The uncontrollable–controllable decomposition. As implied by the nomenclature, poles are either uncontrollable poles or controllable poles (or possibly both), and similarly for pole points.

THEOREM 5.8.

1. *The union of the uncontrollable pole variety and the controllable pole variety is the pole variety.*

2. *The union of the set of uncontrollable poles and the set of controllable poles is the set of poles.*

Proof. From the commutative exact diagram (5.2), we have an exact sequence

$$0 \longrightarrow (\mathcal{B}^c)_{0,y} \longrightarrow \mathcal{B}_{0,y} \longrightarrow \mathcal{B} \,/\, \mathcal{B}^c \longrightarrow 0,$$

which is dual to a corresponding sequence of finitely generated modules in reverse order in the diagram (5.1). Applying Lemma 4.8 gives us the first claim concerning pole points immediately, and also the following:

$$\{\text{uncontrollable poles}\} \subseteq \{\text{poles}\} \subseteq \{\text{uncontrollable poles, controllable poles}\}.$$

It remains to show that every controllable pole is a pole. Write $M_{0,y}$ and $M^c_{0,y}$, respectively, for the modules to which $\mathcal{B}_{0,y}$ and $\mathcal{B}^c_{0,y}$ are dual. We now have the following:

$$\operatorname{Ass} M^c_{0,y} \subseteq \operatorname{supp} M^c_{0,y} \subseteq \operatorname{supp} M_{0,y}.$$

As $\operatorname{ann} M_{0,y} \neq 0$ (its dual is autonomous), $0 \notin \operatorname{supp} M_{0,y}$. Now by Theorem 5.3, every element of $\operatorname{Ass} M^c_{0,y}$ is principal and therefore has height 1. Each such element must therefore be minimal in $\operatorname{supp} M_{0,y}$, and so by Theorem 4.7 is therefore in $\operatorname{Ass} M_{0,y}$. This proves that $\operatorname{Ass} M^c_{0,y} \subseteq \operatorname{Ass} M_{0,y}$, i.e., the controllable poles are also poles, as required. ∎

In the proof above, the fact that every controllable pole is a pole relies on the fact that every controllable pole is principal (from Theorem 5.3). In section 6 we will encounter decompositions of sets of poles where such an observation does not apply, and therefore only partial results of this nature can be obtained.

Note that, contrary to the implication of the nomenclature, it is possible for a pole to be both a controllable pole and an uncontrollable pole, and similarly for pole points! A complete partitioning of poles into controllable and uncontrollable ones would require a suitable notion of multiplicity, which is beyond the scope of this paper.

5.5. The dimension of the characteristic variety. Another important question for an nD system is how "large" the poles are, in a geometric sense. For a 1D system, the pole variety is just a set of isolated points. For a 2D system, the pole variety may be a set of isolated points, or it may include curves. For a 3D system, there is the additional possibility that the pole variety may contain a 2D surface. The dimensions of the pole variety, controllable pole variety, and uncontrollable pole variety, or more generally of the characteristic variety, are therefore of interest.

LEMMA 5.9. *Let $\mathcal{B} = \ker_{\mathcal{A}} E$. Then*

1. *For $\mathcal{B} \neq 0$, the dimension of the characteristic variety of $\mathcal{B}$ is n minus the "right primeness degree" [43] of E.*

2. *$\mathcal{B}$ is finite-dimensional if and only if it has a finite number of characteristic points.*

3. *The controllable pole variety has dimension either -1 (it is empty) or $n-1$. The former case occurs if and only if the transfer matrix is a polynomial matrix. In the latter case, each maximal irreducible subvariety of the controllable pole variety has dimension $n-1$.*

4. *If the transfer matrix is not a polynomial matrix, the pole variety has dimension $n-1$.*

Proof. The first claim is immediate from Theorem 4.4 and the definition of the right primeness degree. The second claim is given in [26, 29].

From Theorem 5.3, ann $\mathcal{B}^c_{0,y}$ is principal, and it must therefore be either $\mathcal{R}$ or an ideal of height 1. Hence $V(\text{ann}\,\mathcal{B}^c_{0,y})$ is either empty or of dimension $n-1$. Again by Theorem 5.3, the former occurs precisely when the transfer matrix is polynomial. In the other case, Theorem 5.3 says further that every coassociated prime of $\mathcal{B}^c_{0,y}$, including the minimal prime divisors of ann $\mathcal{B}^c_{0,y}$, has height 1, so the maximal irreducible subvarieties of $V(\text{ann}\,\mathcal{B}^c_{0,y})$ have dimension $n-1$.

Claim 4 follows from claim 3, as the pole variety contains the controllable pole variety. ∎

The dimension of the characteristic variety (pole variety, etc.) can be found algorithmically, as discussed in [43]. For a system with signal space $k^{\mathbb{N}^n}$, this quantity can also be interpreted as a measure of the order of magnitude of the system's initial condition set, the "autonomy degree" [43]. The interpretation for systems with other signal spaces is still open.

When the transfer matrix is not a polynomial matrix, Lemma 5.9 states that every maximal irreducible subvariety of the controllable pole variety has dimension $n-1$. In consequence, every maximal irreducible subvariety of the pole variety having dimension less than $n-1$ must also be a maximal irreducible subvariety of the uncontrollable pole variety. This is also trivially true when the transfer matrix is a polynomial matrix.

6. Observable and unobservable poles. In this section we will look at the additional structure which occurs when the behavior is a latent variable representation of some other behavior. The principal example of this is when the latent variables are "state variables," e.g., in the context of the classical 1D theory. When the behavior is defined through a Rosenbrock system matrix (or "polynomial matrix description"), this additional structure allows us to relate certain submatrices to various sets of poles.

6.1. Observable and unobservable poles. Consider now a behavior $\mathcal{B}_{l,w}$ which is a latent variable description of some other behavior $\mathcal{B}_w$. The behavior $\mathcal{B}_w$ is called the *manifest behavior* of the *full behavior* $\mathcal{B}_{l,w}$, the variables of w are called *manifest variables*, and the variables of l are called *latent variables*. These ideas are described in [40] for the 1D case and [22, 32] for the 2D case. Formally, we have the following for some polynomial matrices E, F:

$$\mathcal{B}_{l,w} = \left\{ \begin{pmatrix} l \\ w \end{pmatrix} \in \mathcal{A}^{d+q} \;\middle|\; Ew = Fl \right\}, \tag{6.1}$$

$$\mathcal{B}_w = \left\{ w \in \mathcal{A}^q \;\middle|\; \exists\, l \in \mathcal{A}^d \text{ such that } \begin{pmatrix} l \\ w \end{pmatrix} \in \mathcal{B}_{l,w} \right\}. \tag{6.2}$$

It is easy to derive a kernel representation of $\mathcal{B}_w$: we construct a minimal left annihilator L of F, and we have $\mathcal{B}_w = \ker_{\mathcal{A}}(LE)$ [22], [24, p. 27]. There is also the behavior $\mathcal{B}_{l,0}$, which we will call the *unobservable behavior*, consisting of all elements of $\mathcal{B}_{l,w}$ with $w = 0$. $\mathcal{B}_{l,0}$ is isomorphic to $\ker_{\mathcal{A}} F$, and vanishes precisely when the latent variables are observable (in the behavioral sense) from the manifest variables. We can show that $\mathcal{B}_w$ is the factor module of $\mathcal{B}_{l,w}$ by $\mathcal{B}_{l,0}$, so we have an exact sequence

$$0 \longrightarrow \mathcal{B}_{l,0} \longrightarrow \mathcal{B}_{l,w} \longrightarrow \mathcal{B}_w \longrightarrow 0. \tag{6.3}$$

Now consider the pole structure of these behaviors. Take an input/output structure on $\mathcal{B}_{l,w}$ for which the number of inputs which are in the variables w is maximized. (When the latent variables are state variables, all the input variables can of course be taken to be manifest variables.) We will refer to such an input/output structure on $\mathcal{B}_{l,w}$ as a *maximally manifest input/output structure*. Divide the components w and l into input and output components, so $w = (w_1^T, w_2^T)^T$ and $l = (l_1^T, l_2^T)^T$, where $(l_1^T, w_1^T)^T$ is a complete input vector for $\mathcal{B}_{l,w}$. It follows that (w_1, w_2) is an input/output structure on $\mathcal{B}_w$ and (l_1, l_2) is one on $\mathcal{B}_{l,0}$. (It is for this reason that we restrict our consideration to input/output structures which are maximally manifest.) These input/output structures are called the *induced input/output structures* on $\mathcal{B}_w$ and $\mathcal{B}_{l,0}$.

The poles of $\mathcal{B}_{l,w}$ fall naturally into two categories. The first is the set of poles which are observable from the behavior $\mathcal{B}_w$ and the second is the set of poles which cannot be determined from the manifest behavior. (They are poles of the unobservable behavior.)

DEFINITION 6.1. *Let $\mathcal{B}_{l,w}$ be a behavior with latent variables l and a maximally manifest input/output structure with input variables (w_1, l_1) and output variables (w_2, l_2). Take the induced input/output structures on $\mathcal{B}_w$ and $\mathcal{B}_{l,0}$.*

1. *The* unobservable pole variety*,* unobservable pole points*, and* unobservable poles *of $\mathcal{B}_{l,w}$ are defined to be the pole variety, pole points, and poles of the unobservable behavior $\mathcal{B}_{l,0}$.*

2. *The* observable pole variety*,* observable pole points*, and* observable poles *of $\mathcal{B}_{l,w}$ are defined to be the pole variety, pole points, and poles of the manifest behavior $\mathcal{B}_w$.*

In effect we have defined the unobservable poles as the coassociated primes of the behavior $\mathcal{B}_{l_1=0,l_2,w_1=0,w_2=0}$ (notation interpreted in the obvious way) and the observable poles as those of the behavior $\mathcal{B}_{w_1=0,w_2} \subseteq \mathcal{B}_w$. The poles of $\mathcal{B}_{l,w}$ itself are the coassociated primes of $\mathcal{B}_{l_1=0,l_2,w_1=0,w_2}$, and $\mathcal{B}_{w_1=0,w_2}$ is the factor module of $\mathcal{B}_{l_1=0,l_2,w_1=0,w_2}$ by $\mathcal{B}_{l_1=0,l_2,w_1=0,w_2=0}$, so we have an exact sequence

$$0 \longrightarrow \mathcal{B}_{l_1=0,l_2,w_1=0,w_2=0} \longrightarrow \mathcal{B}_{l_1=0,l_2,w_1=0,w_2} \longrightarrow \mathcal{B}_{w_1=0,w_2} \longrightarrow 0. \tag{6.4}$$

THEOREM 6.2. *Let $\mathcal{B}_{l,w}$ be a behavior with latent variables l and a maximally manifest input/output structure. Then the poles/pole points of the full behavior $\mathcal{B}_{l,w}$ have the following properties:*

1. *A point $\underline{a}$ is an observable pole point if and only if there exists $(l, w) \in \mathcal{B}_{l,w}$ such that $w_1 = 0$ and w_2 is nonzero and exponential with frequency $\underline{a}$.*

2. *A point $\underline{a}$ is an unobservable pole point if and only if there exists $(l, w) \in \mathcal{B}_{l,w}$ such that $w = 0, l_1 = 0$ and l_2 is nonzero and exponential with frequency $\underline{a}$.*

3. *Suppose that the latent variables l contain no free variables, i.e., $\mathcal{B}_{l,0}$ is autonomous. Then $\mathcal{B}_{l,w}$ is observable if and only if it has no unobservable poles or, equivalently, no unobservable pole points.*

4. *The union of the unobservable pole variety and the observable pole variety is the pole variety.*

5. *Every observable pole is a pole. Every pole is either an unobservable pole or an observable pole (or both).*

Proof. The first two claims follow on applying Theorem 4.4 and the last two are immediate from Lemma 4.8 together with the exact sequence (6.4). For claim 3, the condition that $\mathcal{B}_{l,w}$ is observable is equivalent to the condition $\mathcal{B}_{l,0} = 0$, or alternatively, to the vanishing of the module to which $\mathcal{B}_{l,0}$ is dual. The result follows since a finitely generated module M is 0 precisely when it has no associated primes (Theorem 4.7), or equivalently, when $V(\operatorname{ann} M) = \emptyset$. ∎

We do not claim that every unobservable pole is a pole. Claim 3 of Theorem 6.2 is similar to the classical 1D result that a system in state-space form is observable if and only if it has no output decoupling zeros. Note that the condition on this last result is satisfied if the latent variables l are interpreted as states.

6.2. Complete classification of poles. We can combine the partitioning of pole points into unobservable and observable pole points with the partitioning into uncontrollable and controllable pole points. This involves the controllable part of $\mathcal{B}_{l,w}$ and its associated unobservable and manifest behaviors. We denote by $\mathcal{B}^c_{l,w}$ the controllable part of $\mathcal{B}_{l,w}$, by $\mathcal{B}^c_{l,0}$ the unobservable behavior of $\mathcal{B}^c_{l,w}$, by $\mathcal{B}^c_w$ the manifest behavior of $\mathcal{B}^c_{l,w}$, and so on. Care must be taken with this notation, as it is not generally the case that $\mathcal{B}^c_{l,0} = (\mathcal{B}_{l,0})^c$.

Our final refinement of the pole structure is defined as follows.

DEFINITION 6.3. *Let $\mathcal{B}_{l,w}$ be a behavior with latent variables l and a maximally manifest input/output structure. Then*

1. *The* unobservable uncontrollable pole variety*,* unobservable uncontrollable pole points*, and* unobservable uncontrollable poles *are the characteristic variety, characteristic points, and coassociated primes of $\mathcal{B}_{l,0}$ / $\mathcal{B}^c_{l,0}$.*

2. *The* unobservable controllable pole variety*,* unobservable controllable poles points*, and* unobservable controllable poles *are the pole variety, pole points, and poles of $\mathcal{B}^c_{l,0}$.*

3. *The* observable uncontrollable pole variety*,* observable uncontrollable pole points*, and* observable uncontrollable poles *are the characteristic variety, characteristic points, and coassociated primes of $\mathcal{B}_w$ / $\mathcal{B}^c_w$.*

4. *The* observable controllable pole variety*,* observable controllable pole points*, and* observable controllable poles *are the pole variety, pole points, and poles of $\mathcal{B}^c_w$.*

To show that the unobservable uncontrollable poles are meaningfully defined, we have to show that $(\mathcal{B}^c)_{l,0}$ is a subbehavior of $\mathcal{B}_{l,0}$, which is evident from the fact that $(\mathcal{B}^c)_{l,0} = \mathcal{B}_{l,0} \cap \mathcal{B}^c$. To show that the observable uncontrollable poles are meaningfully defined, we have to show that $(\mathcal{B}^c)_w$ is a subbehavior of $\mathcal{B}_w$, which follows by definition. In fact, a stronger claim is possible.

THEOREM 6.4. *Let $\mathcal{B}_{l,w}$ be a behavior with latent variables l. Then the manifest behavior of the controllable part of $\mathcal{B}_{l,w}$ is equal to the controllable part of the manifest behavior of $\mathcal{B}_{l,w}$.*

Proof. We write $\mathcal{B} = \mathcal{B}_{l,w}$ as usual. Clearly $(\mathcal{B}^c)_w$ is a subbehavior of $\mathcal{B}_w$. It is also clear from the definition of controllability that as $\mathcal{B}^c$ is controllable, so is $(\mathcal{B}^c)_w$. Now write E for any kernel representation of $\mathcal{B}$ and E^c for any kernel representation of $\mathcal{B}^c$. Since $\mathcal{B}$ and $\mathcal{B}^c$ have the same input/output structures, any submatrices comprised of corresponding columns of E and of E^c must have the same rank. In particular,

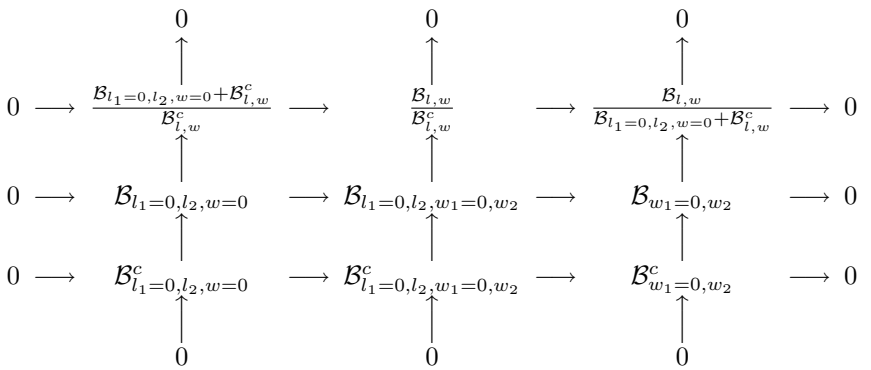


FIG. 6.1. *Exact commutative diagram for complete classification of poles. Note that the columns of this diagram are read upwards, contrary to convention; this is to achieve symmetry in Figure* 6.2.

$m(\mathcal{B}_{l,0}) = m((\mathcal{B}^c)_{l,0})$. Now we use the additivity of the number of free variables:

$$\begin{aligned} m((\mathcal{B}^c)_w) &= m((\mathcal{B}^c)_{l,w}) - m((\mathcal{B}^c)_{l,0}) \\ &= m(\mathcal{B}_{l,w}) - m(\mathcal{B}_{l,0}) \\ &= m(\mathcal{B}_w). \end{aligned}$$

Summarizing, $(\mathcal{B}^c)_w$ is a subbehavior of $\mathcal{B}_w$ which has the same number of free variables and is controllable. This suffices to prove that $(\mathcal{B}^c)_w$ is the controllable part of $\mathcal{B}_w$. ∎

Theorem 6.4 proves that the observable controllable poles are equal to the controllable poles of the manifest behavior (if you like, the "controllable observable poles"). This result also shows that we can use the notation $\mathcal{B}^c_w$ without fear of ambiguity.

Currently we have a pair of exact sequences of autonomous behaviors whose coassociated primes correspond to various sets of poles:

$$0 \longrightarrow \mathcal{B}_{l_1=0,l_2,w_1=0,w_2=0} \longrightarrow \mathcal{B}_{l_1=0,l_2,w_1=0,w_2} \longrightarrow \mathcal{B}_{w_1=0,w_2} \longrightarrow 0, \tag{6.5}$$

$$0 \longrightarrow \mathcal{B}^c_{l_1=0,l_2,w_1=0,w_2=0} \longrightarrow \mathcal{B}^c_{l_1=0,l_2,w_1=0,w_2} \longrightarrow \mathcal{B}^c_{w_1=0,w_2} \longrightarrow 0, \tag{6.6}$$

and each module in the sequence (6.6) is a submodule of the corresponding module in the upper sequence (6.5). Let us look at the corresponding factor modules. Since $\mathcal{B}^c_{l_1=0,l_2,w_1=0,w_2=0} = \mathcal{B}_{l_1=0,l_2,w_1=0,w_2=0} \cap \mathcal{B}^c_{l_1,l_2,w_1,w_2}$, the factor module is

$$\frac{\mathcal{B}_{l_1=0,l_2,w_1=0,w_2=0}}{\mathcal{B}_{l_1=0,l_2,w_1=0,w_2=0} \cap \mathcal{B}^c_{l_1,l_2,w_1,w_2}} \cong \frac{\mathcal{B}_{l_1=0,l_2,w_1=0,w_2=0} + \mathcal{B}^c_{l_1,l_2,w_1,w_2}}{\mathcal{B}^c_{l_1,l_2,w_1,w_2}}. \tag{6.7}$$

This is obviously a submodule of $\mathcal{B}_{l_1,l_2,w_1,w_2} \,/\, \mathcal{B}^c_{l_1,l_2,w_1,w_2}$, which by (5.5) is isomorphic to $\mathcal{B}_{l_1=0,l_2,w_1=0,w_2}/\mathcal{B}^c_{l_1=0,l_2,w_1=0,w_2}$. These behaviors are all shown in Figure 6.1. By the Snake lemma (e.g., [10, Ex. A3.10]), we also have an isomorphism

$$\frac{\mathcal{B}_{l_1,l_2,w_1,w_2}}{\mathcal{B}_{l_1=0,l_2,w_1=0,w_2=0} + \mathcal{B}^c_{l_1,l_2,w_1,w_2}} \cong \frac{\mathcal{B}_{w_1=0,w_2}}{\mathcal{B}^c_{w_1=0,w_2}}, \tag{6.8}$$

and the diagram in Figure 6.1 is commutative and exact.

Reading the nontrivial behaviors in Figure 6.1 from left to right and from top to bottom, we see that the coassociated primes are, respectively, the following: the

unobservable uncontrollable poles, the uncontrollable poles, the observable uncontrollable poles, the unobservable poles, the poles, the observable poles, the unobservable controllable poles, the controllable poles, and the observable controllable poles.

We now apply Lemma 4.8 again. This gives us the basic structure of the poles of a behavior given by a latent variable description.

1. The union of the unobservable uncontrollable pole variety and the observable uncontrollable pole variety is the uncontrollable pole variety. Every observable uncontrollable pole is an uncontrollable pole. Every uncontrollable pole is either an unobservable uncontrollable pole or an observable uncontrollable pole.

2. The union of the unobservable pole variety and the observable pole variety is the pole variety. Every observable pole is a pole. Every pole is either an unobservable pole or an observable pole.

3. The union of the unobservable controllable pole variety and the observable controllable pole variety is the controllable pole variety. Every observable controllable pole is a controllable pole. Every controllable pole is either an unobservable controllable pole or an observable controllable pole.

4. The union of the unobservable uncontrollable pole variety and the unobservable controllable pole variety is the unobservable pole variety. Every unobservable uncontrollable pole is an unobservable pole. Every unobservable pole is either an unobservable controllable pole or an unobservable uncontrollable pole.

5. The union of the uncontrollable pole variety and the controllable pole variety is the pole variety. The union of the set of uncontrollable poles and the set of controllable poles is the set of poles.

6. The union of the observable uncontrollable pole variety and the observable controllable pole variety is the observable pole variety. The union of the set of observable uncontrollable poles and the set of observable controllable poles is the set of observable poles.

In general, we do not expect the inclusions of sets of poles missing from the list above to hold (e.g., we do not expect every unobservable uncontrollable pole to be an uncontrollable pole). Note however that we are able to make a full decomposition in the last case of observable poles. This is possible because the observable controllable poles are a subset of the controllable poles, and are therefore principal by Theorem 5.3. We can therefore apply the same argument as in the proof of Theorem 5.8 to deduce that every observable controllable pole is an observable pole. This reasoning cannot be applied to cases 1–4.

We summarize the relationships between the various sets of poles in a conceptual exact diagram (Figure 6.2). This diagram should be interpreted as follows: if $0 \longrightarrow$ Set 1 $\longrightarrow$ Set 2 $\longrightarrow$ Set 3 $\longrightarrow 0$ is a row or column, then the union of the pole variety of Set 1 and the pole variety of Set 3 is the pole variety of Set 2. The corresponding statements for the poles themselves is slightly weaker, except for the middle and right-hand columns, for which an analogous law holds.

Example. Take the 2D behavior given by the following polynomial matrix:

$$\mathcal{B}_{l,u,y} = \ker_{\mathcal{A}} E_1,$$
$$E_1 = \left(\begin{array}{c|c|c} (z_1^2 - z_2)(z_1 + z_2)(z_1 - 1)z_2 & 0 & z_2^2(z_1^2 - z_2) \\ (z_1^2 - z_2)(z_1 + z_2)(z_1 - 1)z_1 & (z_1 + z_2)^2 & z_1 z_2 (z_1^2 - z_2) \end{array} \middle| \begin{array}{c} -(z_1^2 + z_2^2 - 1) \\ (z_1^2 + z_2^2 - 1) \end{array}\right).$$

The signal space $\mathcal{A}$ can be taken as $k^{\mathbb{N}^n}$, $\mathcal{C}^\infty(\mathbb{R}^n, k)$, or $\mathcal{D}'(\mathbb{R}^n, k)$ for $k = \mathbb{R}$ or $\mathbb{C}$; the ring $\mathcal{R} = k[\mathbf{z}]$. The behavior $\mathcal{B}_{l,u,y}$ has 2 inputs, 1 output, and 1 latent variable.

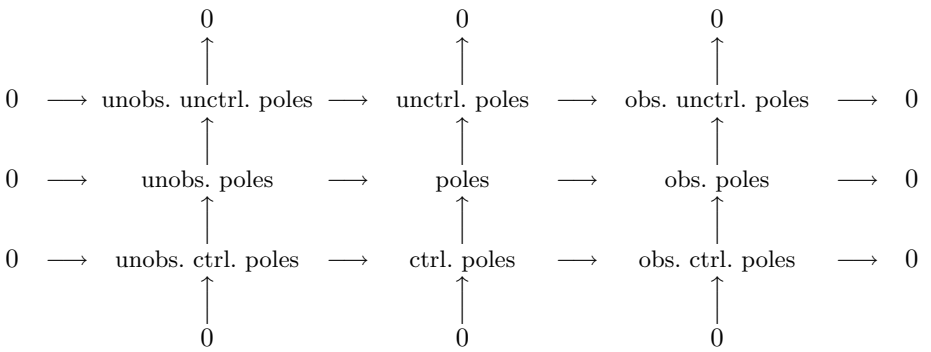


FIG. 6.2. *Relationships between sets of poles.*

We will now find the 9 pole varieties of this behavior. This requires constructing representations for 9 related behaviors—those appearing in Figure 6.1. To begin with, we have the zero-input and unobservable behaviors:

$$\mathcal{B}_{l,u=0,y} \cong \ker_{\mathcal{A}} E_2, \qquad E_2 = \left(\begin{array}{c|c} (z_1^2 - z_2)(z_1 + z_2)(z_1 - 1)z_2 & -(z_1^2 + z_2^2 - 1) \\ (z_1^2 - z_2)(z_1 + z_2)(z_1 - 1)z_1 & (z_1^2 + z_2^2 - 1) \end{array}\right),$$
$$\mathcal{B}_{l,u=0,y=0} \cong \ker_{\mathcal{A}} E_3, \qquad E_3 = \left(\begin{array}{c} (z_1^2 - z_2)(z_1 + z_2)(z_1 - 1)z_2 \\ (z_1^2 - z_2)(z_1 + z_2)(z_1 - 1)z_1 \end{array}\right).$$

Next, observe that we can write $E_1 = LE_4$, where L and E_4 are given by

$$L = \left(\begin{array}{cc} z_2 & -1 \\ z_1 & 1 \end{array}\right),$$
$$E_4 = \left(\begin{array}{c|cc|c} (z_1^2 - z_2)(z_1 + z_2)(z_1 - 1) & (z_1 + z_2) & z_2(z_1^2 - z_2) & 0 \\ 0 & (z_1 + z_2)z_2 & 0 & (z_1^2 + z_2^2 - 1) \end{array}\right).$$

Since E_4 is minor left prime, $\ker_{\mathcal{A}} E_4 = \mathcal{B}^c_{l,u,y}$. Hence we have the following representations:

$$\mathcal{B}^c_{l,u=0,y} \cong \ker_{\mathcal{A}} E_5, \qquad E_5 = \left(\begin{array}{c|c} (z_1^2 - z_2)(z_1 + z_2)(z_1 - 1) & 0 \\ 0 & (z_1^2 + z_2^2 - 1) \end{array}\right),$$
$$\mathcal{B}^c_{l,u=0,y=0} \cong \ker_{\mathcal{A}} E_6, \qquad E_6 = \left(\begin{array}{c} (z_1^2 - z_2)(z_1 + z_2)(z_1 - 1) \\ 0 \end{array}\right).$$

For the manifest behaviors of $\mathcal{B}_{l,u,y}$ and $\mathcal{B}^c_{l,u,y}$, we need minimal left annihilators of E_3 and E_6, and these are easily found to be the matrices $E_7 = (z_1 \ \ -z_2)$ and $E_8 = (0 \ \ 1)$, respectively. The zero-input manifest behaviors are now found by taking the appropriate submatrices of E_7E_1 and E_8E_4, respectively:

$$\mathcal{B}_{u=0,y} \cong \ker_{\mathcal{A}} E_9, \qquad E_9 = (-(z_1^2 + z_2^2 - 1)(z_1 + z_2)),$$
$$\mathcal{B}^c_{u=0,y} \cong \ker_{\mathcal{A}} E_{10}, \qquad E_{10} = (z_1^2 + z_2^2 - 1).$$

Since E_4 has full row rank, Theorem 5.7 tells us that a kernel representation of $E_4\mathcal{B}_{l,u,y} \cong \mathcal{B}_{l,u,y}/\mathcal{B}^c_{l,u,y}$ is the matrix L. Similarly, the representation found for $\mathcal{B}^c_{u,y}$ also has full row rank, and so a kernel representation for $\mathcal{B}_{u,y}/\mathcal{B}^c_{u,y}$ is the matrix $E_{11} =$

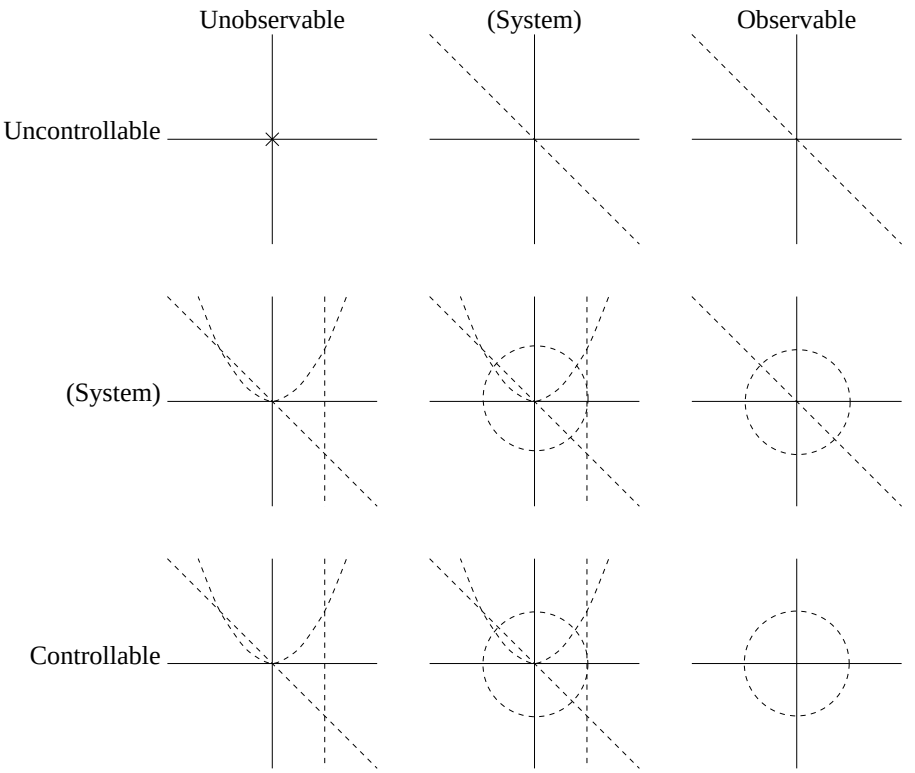


FIG. 6.3. *Example of pole varieties.*

$(-(z_1 + z_2))$. Using a similar trick, a kernel representation for $\mathcal{B}_{l,u=0,y=0}/\mathcal{B}^c_{l,u=0,y=0}$ is

$$E_{12} \;=\; \begin{pmatrix} z_2 \\ z_1 \end{pmatrix}.$$

Now the 9 sets of pole points, 9 behaviors of relevance, and 9 kernel representations are listed as follows:

Pole point set	Behavior	Rep. matrix
Unobservable uncontrollable pole points	$\mathcal{B}_{l,u=0,y=0}/\mathcal{B}^c_{l,u=0,y=0}$	E_{12}
Uncontrollable pole points	$\mathcal{B}_{l,u,y}/\mathcal{B}^c_{l,u,y}$	L
Observable uncontrollable pole points	$\mathcal{B}_{u,y}/\mathcal{B}^c_{u,y}$	E_{11}
Unobservable pole points	$\mathcal{B}_{l,u=0,y=0}$	E_3
(System) pole points	$\mathcal{B}_{l,u=0,y}$	E_2
Observable pole points	$\mathcal{B}_{u=0,y}$	E_9
Unobservable controllable pole points	$\mathcal{B}^c_{l,u=0,y=0}$	E_6
Controllable pole points	$\mathcal{B}^c_{l,u=0,y}$	E_5
Observable controllable pole points	$\mathcal{B}^c_{u=0,y}$	E_{10}

For this example, it is now easy to plot the various pole varieties as the rank-loss varieties of the corresponding matrices. (However in general we suspect that computation of the annihilator is a more efficient method.) These varieties are shown in Figure 6.3.

Note that the behavior $\mathcal{B}_{l,u=0,y=0}/\mathcal{B}^c_{l,u=0,y=0}$ giving the unobservable uncontrollable pole points has a single pole (z_1, z_2). We have conjectured that not every un-

observable uncontrollable pole is an uncontrollable pole. We can confirm this using the current example; we have to show that (z_1, z_2) is not an associated prime of $\mathrm{Coker}_{\mathcal{R}} L$, where L is the matrix given above. Equivalently, we show that no element of $\mathcal{R}^{1,2} \backslash \mathrm{Im}_{\mathcal{R}} L$ is taken into $\mathrm{Im}_{\mathcal{R}} L$ by the actions of z_1 and z_2. Let N denote the set of all elements x of $\mathcal{R}^{1,2}$ such that $z_1 x \in \mathrm{Im}_{\mathcal{R}} L$ and $z_2 x \in \mathrm{Im}_{\mathcal{R}} L$. Then N is the set of polynomial vectors corresponding to the first two components of the kernel (natural left action) of the matrix

$$K = \left(\begin{array}{cc|cccc} z_1 & 0 & z_2 & z_1 & 0 & 0 \\ 0 & z_1 & -1 & 1 & 0 & 0 \\ z_2 & 0 & 0 & 0 & z_2 & z_1 \\ 0 & z_2 & 0 & 0 & -1 & 1 \end{array} \right)^T.$$

However, it is possible to show that the kernel of this matrix is equal to the image of the matrix

$$\left(\begin{array}{cc|cccc} z_2 & -1 & -z_1 & 0 & -z_2 & 0 \\ z_1 & 1 & 0 & -z_1 & 0 & -z_2 \end{array} \right),$$

and hence $N = \mathrm{Im}_{\mathcal{R}} L$. This proves that (z_1, z_2) is not an associated prime of $\mathrm{Coker}_{\mathcal{R}} L$, and hence is not an uncontrollable pole of the behavior, although it is an unobservable uncontrollable pole.

6.3. Rosenbrock system matrices. One important class of representations of 1D and nD systems is the class of Rosenbrock system matrices, also called polynomial matrix descriptions; see, e.g., [33] for the 1D case and [19] for the nD case.

We write here x for the latent variables, u for the (manifest) free variables, and y for the remaining manifest variables. We can think of u as the input, x as the state, and y as the output. Since x is to be interpreted as a vector of state variables, it is assumed to contain no free variables, i.e., the behavior with zero input and zero output is assumed to be autonomous. We consider state-input-output behaviors given by the following classical system of equations:

$$T(\underline{\mathbf{z}}) x(\underline{t}) = U(\underline{\mathbf{z}}) u(\underline{t}), \tag{6.9}$$

$$y(\underline{t}) = V(\underline{\mathbf{z}}) x(\underline{t}) + W(\underline{\mathbf{z}}) u(\underline{t}). \tag{6.10}$$

We can write the resulting behavior in kernel form as follows:

$$\mathcal{B}_{x,u,y} = \ker_{\mathcal{A}} \left(\begin{array}{ccc} T & -U & 0 \\ V & W & -I \end{array} \right). \tag{6.11}$$

The submatrix $\left(\begin{smallmatrix} T & -U \\ V & W \end{smallmatrix} \right)$ is called a *Rosenbrock system matrix*. We define the *pole points of the Rosenbrock system matrix* to be the points where T fails to have full column rank. This is the definition previously given in the 1D case (e.g., in [20]).

Note that the standard input/output structure on $\mathcal{B}_{x,u,y}$ is trivially a maximally manifest input/output structure, with the states being treated as additional outputs in the behavioral sense.

LEMMA 6.5. *Let $\mathcal{B}_{x,u,y}$ be a differential or difference behavior given by a Rosenbrock system matrix* (6.9–6.10). *Then*

1. *The pole points of $\mathcal{B}_{x,u,y}$ are precisely the pole points of the Rosenbrock system matrix.*

2. *The unobservable pole points of* $\mathcal{B}_{x,u,y}$ *are precisely the points where the matrix* $(T^T\ V^T)^T$ *has less than full column rank.*

3. *Each uncontrollable pole point of* $\mathcal{B}_{x,u,y}$ *is a point where* $(T\ -U)$ *has less than its normal rank; the converse holds for* 1*D systems.*

4. *The observable controllable pole points of* $\mathcal{B}_{x,u,y}$ *are precisely the points* $\underline{a}$ *such that the denominator of some entry of the input-to-output transfer matrix of the Rosenbrock system matrix vanishes at* $\underline{a}$.

Proof.

1. A kernel representation of $\mathcal{B}_{x,0,y}$ is $\begin{pmatrix} T & O \\ V & -I \end{pmatrix}$ (to within isomorphism). By Corollary 5.2, the pole points are the points where this matrix, or equivalently T, fails to have full column rank.

2. The unobservable pole points are the characteristic points of $\mathcal{B}_{x,0,0} \cong \ker_{\mathcal{A}} \begin{pmatrix} T \\ V \end{pmatrix}$. By Theorem 4.4, the unobservable pole points are the points where this matrix fails to have full column rank.

3. By Theorem 5.6, in the 1D case the uncontrollable pole points are the points where the whole matrix $\begin{pmatrix} T & -U & 0 \\ V & W & -I \end{pmatrix}$ has less than its normal rank. These are equal to the rank-loss points of $(T\ -U)$.

4. By Theorem 6.4, the observable controllable pole points are the controllable pole points of $\mathcal{B}_{u,y}$, the transfer matrix of which is the input-to-output transfer matrix of the PMD. The claim now follows from Corollary 5.4. ∎

Lemma 6.5 is enough to prove that, in the 1D case, the pole points, unobservable pole points, uncontrollable pole points, and observable controllable pole points of $\mathcal{B}_{x,0,y}$ correspond, respectively, to the poles, output decoupling zeros, input decoupling zeros, and transmission poles of the system, as defined classically. Note also that, in the general nD case, we have proved in Theorems 5.6 and 6.2 that a behavior is controllable (resp., observable) if and only if it has no uncontrollable pole points (resp., no unobservable pole points).

We can readily apply our theory to characterize the pole points of special classes of nD systems, for example, systems governed by the Roesser or Fornasini–Marchesini state space models. In these cases we find that the pole points are the points where the "characteristic polynomial" vanishes, as has been anticipated by many other authors. Such an analysis is however beyond the scope of this paper.

7. Polar decomposition and integral representation. In this section we discuss a decomposition of a behavior that emerges from the theory of poles, and also the related integral representation theorem.

7.1. Polar and decoupling zero decompositions. In a 1D system, any trajectory with zero input can be written as a sum of polynomial exponential trajectories with frequencies corresponding to the system poles. This results from a decomposition of the behavior as a sum of certain subbehaviors, a decomposition which is dual to the classical algebraic "primary decomposition." We can apply the same principles in the nD case.

Consider the ring $\mathcal{R} = k[\underline{\mathbf{z}}]$ or $\mathcal{R} = k[\underline{\mathbf{z}}, \underline{\mathbf{z}}^{-1}]$ (or more generally any Noetherian ring) and a finitely generated $\mathcal{R}$-module L. Given a proper submodule N of L, N has a *primary decomposition* (e.g., [3, 5, 10]) into submodules of L containing N:

$$N = \bigcap_{i=1}^{l} N_i, \tag{7.1}$$

where each quotient L/N_i is J_i-coprimary for some prime J_i, i.e., Ass (L/N_i) consists

of the single prime J_i, and the J_i's include all the associated primes of L/N. The intersection can be taken to be such that the J_i's are precisely the associated primes of L/N, each occurring precisely once, in which case it is called a *minimal primary decomposition*. Such a decomposition is in general nonunique.

The dual of (7.1) allows us to express any behavior $\mathcal{B}$ as the sum of subbehaviors $\mathcal{B}_i$, where each $\mathcal{B}_i$ has a single coassociated prime, and these primes are precisely the coassociated primes of $\mathcal{B}$. This idea has two interesting applications.

THEOREM 7.1. *Let $\mathcal{B}$ be a differential or difference behavior. Then the zero-input behavior has a decomposition*

$$\mathcal{B}_{0,y} = \sum_{i=1}^{l} \mathcal{B}_i, \tag{7.2}$$

where each behavior $\mathcal{B}_i$ has only a single coassociated prime J_i, and the J_i's are precisely the poles of $\mathcal{B}$. We call this decomposition a polar decomposition *of $\mathcal{B}_{0,y}$.*

The behavior $\mathcal{B}$ itself also has a decomposition

$$\mathcal{B} = \mathcal{B}^c + \sum_{l=1}^{r} \mathcal{B}'_l, \tag{7.3}$$

where each behavior $\mathcal{B}'_l$ has only a single coassociated prime J'_l, and the J'_l's are precisely the uncontrollable poles of $\mathcal{B}$.

Proof. The first claim follows immediately by application of the dual version of (7.1) to the zero-input behavior $\mathcal{B}_{0,y}$. For the second part, assume first that $\mathcal{B}$ is not autonomous, since otherwise $\mathcal{B}^c = 0$, the uncontrollable poles are the coassociated primes of $\mathcal{B}$, and the result is immediate by the dual of (7.1). Hence by Theorem 3.3 M is not a torsion module and so 0 is an associated prime. Thus $\mathcal{B}^\perp$ has a minimal primary decomposition: $\mathcal{B}^\perp = \bigcap_{l=0}^{r} N_l$, where, in particular, N_0 is the module such that $\mathrm{Ass}\,(\mathcal{R}^{1,q}/N_0) = \{0\}$. Define $\mathcal{B}'_l = (N_l)^\perp$ for $l = 0, 1, \ldots, r$. This gives us

$$\mathcal{B} = \sum_{l=0}^{r} \mathcal{B}'_l. \tag{7.4}$$

Now by Lemma 5.5, the uncontrollable poles of $\mathcal{B} = D(M)$ are precisely the coassociated primes of the behaviors $\mathcal{B}'_l$ for $l = 1, \ldots, r$. Finally, $\mathcal{R}^{1,q}/N_0$ is 0-primary, or equivalently, torsionfree (e.g., on applying [10, Prop. 3.4]), and so by Theorem 3.8 $\mathcal{B}'_0$ is controllable. The sum $\sum_{l=1}^{r} \mathcal{B}'_l$ is autonomous as the sum of autonomous behaviors, and so by the controllable–autonomous decomposition $\mathcal{B}'_0$ is the controllable part of $\mathcal{B}$. This completes the proof. ∎

Note that the polar decomposition also leads to a decomposition of $\mathcal{B}$ itself, by combination with the controllable–autonomous decomposition. However such a decomposition of $\mathcal{B}$ must be redundant, since by (7.3) it is possible to do this using only the uncontrollable poles. Indeed, the uncontrollable pole decomposition of $\mathcal{B}$ is a refinement of the controllable–autonomous decomposition. The 2D polar decomposition is also similar to a decomposition discussed by Valcher [38], in which any 2D autonomous behavior can be decomposed into the sum of a finite-dimensional behavior and a behavior with a square kernel representation.

The annihilator of a finitely generated module M can be computed using Gröbner basis techniques; an algorithm has been presented in [43]. This is in general the closest we can get to computing the characteristic variety of a behavior. The decompositions

in Theorem 7.1 can be computed via the primary decomposition, algorithms for which have been given in the algebraic literature, for example, [1, 11, 17]. The construction of a primary decomposition necessarily includes an identification of the associated primes, i.e., the poles or uncontrollable poles.

7.2. The integral representation theorem. The integral representation theorem was originally formulated by Ehrenpreis in 1961 [9] and then proven in full generality by Palamodov [28, Thm. VI.4.1]. A simplified special case was presented by Björk in [2, Thm. 8.1.3]. Roughly, the integral representation theorem gives an explicit form for any $\mathcal{C}^\infty$ behavior, in which each trajectory is expressed as a sum of integrals of polynomial exponential functions which also lie in the behavior. In the following, we write $\langle \underline{a}, \underline{t} \rangle$ for the inner product of two vectors $\underline{a}, \underline{t}$. Using the language of the behavioral approach, Björk's statement of the theorem reads as follows.

THEOREM 7.2 (integral representation theorem). *Let $\mathcal{B}$ be a differential behavior with signal space $\mathcal{C}^\infty(\mathbb{R}^n, \mathbb{C})$. Then there exist polynomial vectors $p_1, \ldots, p_L$ in $2n$ variables with the following property. For any compact and convex subset $\mathcal{H}$ of $\mathbb{R}^n$ with nonempty interior $\mathcal{K}$, and for any trajectory $w \in \mathcal{B}$, there exists measures $\mu_1, \ldots, \mu_L$ on $\mathbb{C}^n$ for which*

$$w(\underline{t}) \;=\; \sum_{i=1}^{L} \int_{\mathbb{C}^n} p_i(\underline{a}, \underline{t}) e^{\langle \underline{a}, \underline{t} \rangle} \; d\mu_i(\underline{a}) \tag{7.5}$$

for all $\underline{t} \in \mathcal{H}$, and for which the following conditions also hold:

1. *Each function $p_i(\underline{a}, \underline{t}) e^{\langle \underline{a}, \underline{t} \rangle}$ is a polynomial exponential trajectory of $\mathcal{B}$ for any $\underline{a}$ in the support of the measure μ_i. In particular, the support of μ_i is contained in the characteristic variety of $\mathcal{B}$.*
2. *We have*

$$\int_{\mathbb{C}^n} e^{\mathcal{H}(\underline{a})} (1 + ||\underline{a}||^c) \, d|\mu_i|(\underline{a}) \;<\; \infty$$

for all c, where $\mathcal{H}(\underline{a}) = \sup_{\underline{t} \in \mathcal{H}} (\langle \Re e(\underline{a}), \underline{t} \rangle)$. In particular, the integrals on the right of (7.5) are absolutely convergent and can be differentiated with respect to $\underline{t}$ under the integral sign for $\underline{t} \in \mathcal{K}$.

The integral representation (7.5) can be broken down by first applying the decomposition (7.3) in Theorem 7.1 in order to write $\mathcal{B}$ as a sum of subbehaviors $\mathcal{B}_i$ each with a single coassociated prime. For each $\mathcal{B}_i$, we then obtain an integral representation, and when summed these give a representation of $\mathcal{B}$.

Assume therefore that $\mathcal{B}$ has a single coassociated prime J, or equivalently, that $\mathcal{R}^{1,q}/\mathcal{B}^\perp$ is a J-coprimary. Then according to [28, Cor. IV.4.2] or [2, Thm. 8.4.5], there is a $q \times L$ matrix $A(\underline{\mathbf{z}}, \partial/\partial \underline{\mathbf{z}})$ of differential operators with polynomial coefficients such that

$$\mathcal{B}^\perp \;=\; \{ v \in \mathcal{R}^{1,q} \mid A^T(\underline{a}, \partial/\partial \underline{a}) v^T(\underline{a}) = 0 \text{ for all } \underline{a} \in V \},$$

where V denotes the characteristic variety of $\mathcal{B}$. Such a matrix A is called a *Noetherian operator* for $\mathcal{B}^\perp \subseteq \mathcal{R}^{1,q}$, and we can show that, for all $\underline{a} \in V$,

$$E(\partial/\partial \underline{t})(A(\underline{a}, \underline{t}) e^{\langle \underline{a}, \underline{t} \rangle}) \;=\; (A^T(\underline{a}, \partial/\partial \underline{a}) E^T(\underline{a}) e^{\langle \underline{a}, \underline{t} \rangle})^T \;=\; 0,$$

i.e., that $A(\underline{a}, \underline{t}) e^{\langle \underline{a}, \underline{t} \rangle}$ is a trajectory of $\mathcal{B}$. The columns of $A(\underline{a}, \underline{t})$ then become the polynomial vectors $p_i(\underline{a}, \underline{t})$ in the right-hand side of (7.5); see [2, proof of Thm. 8.1.3].

It is easy to construct a Noetherian operator in the special case when we know in addition that $\operatorname{ann} \mathcal{R}^{1,q}/\mathcal{B}^\perp$ is a radical ideal, i.e., when the annihilator of $\mathcal{R}^{1,q}/\mathcal{B}^\perp$ is equal to its single associated prime. In this case, $M = \mathcal{R}^{1,q}/\mathcal{B}^\perp$ is a torsionfree $\mathcal{R}/J$-module. We write $\overline{\mathcal{R}} = \mathcal{R}/J$, and for any matrix Q over $\mathcal{R}$ we use the notation $\overline{Q}$ for the matrix over $\overline{\mathcal{R}}$ obtained by projection of each entry. Now consider a kernel representation $E \in \mathcal{R}^{g,q}$ of $\mathcal{B}$, and construct an $A \in \mathcal{R}^{q,L}$ such that

$$\operatorname{Im}_{\overline{\mathcal{R}}} \overline{E} \;=\; \operatorname{Ker}_{\overline{\mathcal{R}}} \overline{A},$$

which is possible as $\operatorname{Coker}_{\overline{\mathcal{R}}} \overline{E}$ is torsionfree over $\overline{\mathcal{R}}$. We now have

$$\begin{aligned}\mathcal{B}^\perp &= \operatorname{Im}_{\mathcal{R}} E \\ &= \{v \in \mathcal{R}^{1,q} \;\mid\; \overline{vA} = \overline{v}\overline{A} = 0 \in \overline{\mathcal{R}}^{1,L}\} \\ &= \{v \in \mathcal{R}^{1,q} \;\mid\; A^T(\underline{a})v^T(\underline{a}) = 0 \text{ for all } \underline{a} \in V(\operatorname{ann}\mathcal{B})\}.\end{aligned}$$

This proves that $A(\underline{z}, \partial/\partial\underline{z}) = A(\underline{z})$ is a Noetherian operator for M. The polynomial vectors in the integral representation are then extracted as the columns of A.

The preceding argument can be applied in the case where $\mathcal{B}$ is controllable (and with no other conditions). For it then holds that $\mathcal{B}$ has the single coassociated prime $\{0\}$, which is equal to $\operatorname{ann}\mathcal{B}$. The ring $\overline{\mathcal{R}}$ is simply equal to $\mathcal{R}$, and so the Noetherian operator A is just any matrix such that $\operatorname{Im}_{\mathcal{R}} E = \operatorname{Ker}_{\mathcal{R}} A$ or equivalently $\ker_{\mathcal{A}} E = \operatorname{im}_{\mathcal{A}} A$, i.e., A is any image representation of $\mathcal{B}$. Thus in the controllable case we obtain the following corollary.

COROLLARY 7.3. *Let $\mathcal{B}$ be a controllable differential behavior with signal space $\mathcal{C}^\infty(\mathbb{R}^n, \mathbb{C})$ and an image representation F. Let $F_1, \dots, F_L$ denote the columns of F. Then for any $\underline{a} \in \mathbb{C}^n$, $F_i(\underline{a})e^{\langle \underline{a}, \underline{t}\rangle}$ is an exponential trajectory of $\mathcal{B}$, and every trajectory w of $\mathcal{B}$ can be written*

$$w(\underline{t}) \;=\; \sum_{i=1}^{L} \int_{\mathbb{C}^n} F_i(\underline{a}) e^{\langle \underline{a}, \underline{t}\rangle} \; d\mu_i(\underline{a}) \tag{7.6}$$

for suitable measures $\mu_1, \dots, \mu_L$ as in Theorem 7.2.

COROLLARY 7.4. *Let $u \in \mathcal{A}^m$, $y \in \mathcal{A}^p$ satisfy*

$$P(\underline{\mathbf{z}})y \;=\; u,$$

where $\mathcal{A} = \mathcal{C}^\infty(\mathbb{R}^n, \mathbb{C})$ and P is a square nonsingular matrix. Then there are measures $\mu_1, \dots, \mu_p$ on $\mathbb{C}^n$ such that

$$y_j(\underline{t}) = \int_{\mathbb{C}^n} e^{\langle \underline{a}, \underline{t}\rangle} \, d\mu_j(\underline{a}), \qquad j = 1, \dots, p,$$

and

$$u(\underline{t}) = \sum_{i=1}^{p} \int_{\mathbb{C}^n} P_i(\underline{a}) e^{\langle \underline{a}, \underline{t}\rangle} \, d\mu_i(\underline{a})$$

as in Theorem 7.2, where P_i denotes the ith column of P.

Proof. This result is immediate from Corollary 7.3, since $(U^T \; I_p)^T$ is an image representation of $\ker_{\mathcal{A}} (-I_p \; U)$, the given controllable behavior. □

In the case of a finite-dimensional behavior $\mathcal{B}$, the characteristic variety is finite (see Lemma 5.9), and each integral in the summation of (7.5) can be taken as an evaluation of the integrand at a given characteristic point, multiplied by a suitable

constant. In this situation we can recover the result that every trajectory is a polynomial exponential trajectory [26], which generalizes the well-known decomposition of trajectories of 1D systems. Results for the case of a zero-dimensional characteristic variety are also given in [2, p. 365]. The general construction of Noetherian operators is addressed in [27].

REFERENCES

[1] D. Bayer and D. Mumford, *What can be computed in algebraic geometry*?, in Computational Algebraic Geometry and Commutative Algebra, Sympos. Math. 34, Cambridge University Press, Cambridge, UK, 1993, pp. 1–48.

[2] J.-E. Björk, *Rings of Differential Operators*, North-Holland, Amsterdam, 1979.

[3] N. Bourbaki, *Commutative Algebra*, Hermann, Paris, 1972.

[4] H. Bourlès and M. Fliess, *Finite poles and zeros of linear systems: An intrinsic approach*, Internat. J. Control, 68 (1997), pp. 897–922.

[5] P. Cohn, *Algebra*, Vol. 2, 2nd ed., John Wiley and Sons, Chichester, UK, 1989.

[6] G. Conte and A. M. Perdon, *Zeros, poles and modules in linear system theory*, in Three Decades of Mathematical System Theory, Lecture Notes in Control and Inform. Sci. 135, Springer-Verlag, Berlin, 1989, pp. 79–100.

[7] D. Cox, J. Little, and D. O'Shea, *Ideals, Varieties, and Algorithms*, 2nd ed., Springer-Verlag, New York, 1997.

[8] C. A. Desoer and J. D. Schulman, *Zeros and poles of matrix transfer functions and their dynamical interpretation*, IEEE Trans. Circuits and Systems, 21 (1974), pp. 3–8.

[9] L. Ehrenpreis, *Fourier Analysis in Several Complex Variables*, Wiley-Interscience Publishers, New York, 1970.

[10] D. Eisenbud, *Commutative Algebra: With a View Toward Algebraic Geometry*, Grad. Texts in Math. 150, Springer-Verlag, New York, 1995.

[11] D. Eisenbud, C. Huneke, and W. Vasconcelos, *Direct methods for primary decomposition*, Invent. Math., 110 (1992), pp. 207–235.

[12] M. Fliess, *Some basic structural properties of generalized linear systems*, Systems Control Lett., 15 (1990), pp. 391–396.

[13] M. Fliess, *Controllability revisited*, in Mathematical System Theory (The influence of R. E. Kalman), A.C. Antoulas, ed., Springer-Verlag, Berlin, 1991, pp. 463–474.

[14] M. Fliess, *Une interprétation algébrique de la transformation de Laplace et des matrices de transfert*, Linear Algebra Appl., 203–204 (1994), pp. 429–442.

[15] E. Fornasini, P. Rocha, and S. Zampieri, *State space realization of* 2*-D finite-dimensional behaviors*, SIAM J. Control Optim., 31 (1993), pp. 1502–1517.

[16] S. Fröhler and U. Oberst, *Continuous time-varying linear systems*, Systems Control Lett., 35 (1998), pp. 97–110.

[17] P. Gianni, B. Trager, and G. Zacharias, *Gröbner basis and primary decomposition of polynomial ideals*, J. Symbolic Comput., 6 (1988), pp. 149–167.

[18] L. Hörmander, *An Introduction to Complex Analysis in Several Variables*, North-Holland/ American Elsevier, New York, 1973.

[19] D. Johnson, *Coprimeness in Multidimensional Systems and Symbolic Computation*, Ph.D. thesis, Department of Mathematics, Loughborough University of Technology, Loughborough, UK, 1993.

[20] T. Kailath, *Linear Systems*, Prentice-Hall, Inc., Englewood Cliffs, NJ, 1980.

[21] R. Kalman, *Algebraic structure of linear dynamic systems.* I. *The module of* Σ, Proc. Nat. Acad. Sci. U.S.A., 54 (1965), pp. 1503–1508.

[22] J. Komorník, P. Rocha, and J. C. Willems, *Closed subspaces, polynomial operators in the shift, and ARMA representations*, Appl. Math. Lett., 4 (1991), pp. 15–19.

[23] A. G. J. Macfarlane and N. Karcanias, *Poles and zeros of linear multivariable systems: A survey of the algebraic, geometric and complex-variable theory*, Internat. J. Control, 24 (1976), pp. 33–74.

[24] U. Oberst, *Multidimensional constant linear systems*, Acta Appl. Math., 20 (1990), pp. 1–175.

[25] U. Oberst, *Variations on the fundamental principle for linear systems of partial differential and difference equations with constant coefficients*, Appl. Algebra Engrg., Comm. Comput., 6 (1995), pp. 211–243.

[26] U. Oberst, *Finite-dimensional systems of partial differential or difference equations*, Adv. Appl. Math., 17 (1996), pp. 337–356.

[27] U. Oberst, *The construction of Noetherian operators*, J. Algebra, to appear.
[28] V. P. Palamodov, *Linear Differential Operators with Constant Coefficients*, Springer-Verlag, New York, 1970.
[29] H. Pillai and S. Shankar, *A behavioral approach to control of distributed systems*, SIAM J. Control Optim., 37 (1999), pp. 388–408.
[30] J.-F. Pommaret, *Partial Differential Equations and Group Theory: New Perspectives for Applications*, Math. Appl., 293, Kluwer, Dordrecht, 1994.
[31] J.-F. Pommaret and A. Quadrat, *Algebraic Analysis of Linear Multidimensional Control Systems*, Tech. Report 98–131, CERMICS/INRIA, ENPC, France, 1998; IMA J. Math. Control Inform., to appear.
[32] P. Rocha, *Structure and Representation of* 2*-D Systems*, Ph.D. thesis, Department of Mathematics, University of Groningen, The Netherlands, 1990.
[33] H. Rosenbrock, *State-Space and Multivariable Theory*, John Wiley & Sons, New York, 1970.
[34] J. Rudolph, *Duality in time-varying linear systems: A module theoretic approach*, Linear Algebra Appl., 245 (1996), pp. 83–106.
[35] C. B. Schrader and M. K. Sain, *Research on system zeros: A survey*, Internat. J. Control, 50 (1989), pp. 1407–1433.
[36] S. Shankar and V. R. Sule, *Algebraic geometric aspects of feedback stabilization*, SIAM J. Control Optim., 30 (1992), pp. 11–30.
[37] V. Sule, *PhD thesis*, Dept. of Elec. Engineering, Indian Institute of Technology, Bombay, 1990.
[38] M. Valcher, *Characteristic cones and stability properties of two-dimensional autonomous behaviors*. IEEE Trans. Circuits Systems I Fund. Theory Appl., (1999), to appear.
[39] J. C. Willems, *From time series to linear system. Part* I*: Finite-dimensional linear time invariant systems*, Automatica, 22 (1986), pp. 561–580.
[40] J. C. Willems, *Paradigms and puzzles in the theory of dynamical systems*, IEEE Trans. Automat. Control, 36 (1991), pp. 259–294.
[41] J. C. Willems, *On interconnections, control, and feedback*, IEEE Trans. Automat. Control, 42 (1997), pp. 326–339.
[42] J. Wood, E. Rogers, and D. H. Owens, *Controllable and autonomous nD systems*, Multidimens. Systems Signal Process., 10 (1999), pp. 33–69.
[43] J. Wood, E. Rogers, and D. Owens, *A formal theory of matrix primeness*, Math. Control Signals Systems, 11 (1998), pp. 40–78.
[44] J. Wood and E. Zerz, *Notes on the definition of behavioral controllability*, Systems Control Lett., 37 (1999), pp. 31–37.
[45] B. F. Wyman and M. K. Sain, *Module theoretic zero structures for system matrices*, SIAM J. Control Optim., 25 (1987), pp. 87–99.
[46] E. Zerz, *Primeness of multivariate polynomial matrices*, Systems Control Lett., 29 (1996), pp. 139–145.
[47] Y. Zheng, J. C. Willems, and C. Zhang, *Common factors and controllability of nonlinear systems*, in Proceedings of the 36th IEEE Conference on Decision and Control, San Diego, CA, IEEE, Piscataway, NJ, 1997, pp. 2584–2589.

SIAM J. CONTROL OPTIM.
Vol. 38, No. 2, pp. 662–663

ERRATUM: ASYNCHRONOUS STOCHASTIC APPROXIMATIONS*

VIVEK S. BORKAR†

PII. S0363012998345913

The proofs of Lemma 2.1 and Theorem 3.2 of [1] have a common error, the correction of which needs additional conditions: the use of l'Hôpital's rule in both equates the limiting ratio of derivatives of functions with the limiting ratio of the functions themselves, but this presupposes that the latter exist. Of these, the conclusions of Lemma 2.1 can simply be adopted as an additional assumption on stepsizes (a harmless one, as it is satisfied by all the usual examples). The proof of Theorem 3.2 will continue to hold if we make the following additional assumption.

For all $x > 0$ and

$$N(n,x) = \min\left\{ m > n : \sum_{k=n+1}^{m} \tilde{a}(k) > x \right\},$$

the limit

$$\lim_{n\to\infty} \frac{\sum_{k=\nu(n,i)}^{\nu(N(n,x),i)} a(k,i)}{\sum_{k=\nu(n,j)}^{\nu(N(n,x),j)} a(k,j)} \tag{1}$$

exists almost surely (a.s.) for all i, j.

Intuitively, this implies that the updating of different components is "evenly spread." Under this condition,

$$\lim_{t\to\infty} \frac{\int_0^x \mu_{t+y}(j)dy}{\int_0^x \mu_{t+y}(i)dy}$$

is guaranteed to exist a.s. a priori, justifying the use of l'Hôpital's rule.

Since $\{\tilde{a}(n)\}$ were defined after the artificial "unfolding" of iterates, (1) may look a little contrived. One can alternatively use the following condition defined in terms of the original stepsizes: for all $x > 0$ and

$$N(n,x) = \min\left\{ m > n : \sum_{k=n+1}^{m} \sum_{i} \bar{a}(k,i) I\{i \in Y_n\} > x \right\},$$

the limit

$$\lim_{n\to\infty} \frac{\sum_{k=\nu(n,i)}^{\nu(N(n,x),i)} a(k,i)}{\sum_{k=\nu(n,j)}^{\nu(N(n,x),j)} a(k,j)}$$

exists a.s. for all i, j.

Under this, an appropriate variant of Theorem 3.2 can be proved without the unfolding. See also Lemma 4.8, [2].

*Received by the editors October 12, 1998; accepted for publication (in revised form) May 11, 1999; published electronically February 9, 2000.
http://www.siam.org/journals/sicon/38-2/34591.html
†School of Technology and Computer Science, Tata Institute of Fundamental Research, Homi Bhabha Road, Mumbai 400005, India (borkar@tifr.res.in).

REFERENCES

[1] V.S. Borkar, *Asynchronous stochastic approximations*, SIAM J. Control Optim., 36 (1998), pp. 840–851.
[2] V.R. Konda and V.S. Borkar, *Actor-critic–type learning algorithms for Markov decision processes*, SIAM J. Control Optim., 38 (1999), pp. 94–123.

SIAM J. CONTROL OPTIM.
Vol. 38, No. 3, pp. 665–682

DYNAMIC L^p-HEDGING IN DISCRETE TIME UNDER CONE CONSTRAINTS*

HUYÊN PHAM†

Abstract. We consider a general discrete time process of a financial market with cone constraints on trading strategies. In this framework, we study the problem of minimizing the expected l_p-loss function of the shortfall of a given contingent claim in L^p. This stochastic control problem is solved by using results on superhedging and a convex duality approach.

Key words. hedging, superhedging, constrained portfolios, stochastic control problem, duality theory

AMS subject classifications. 90A09, 90A46, 93E20, 60G48

PII. S0363012998341095

1. Introduction. In a frictionless financial market which is free of arbitrage opportunities and complete, the problem of pricing and hedging is well understood. Any contingent claim H is attainable: starting from an initial wealth v_0, an agent can find a trading strategy θ^H that will allow his (self-financed) wealth V^{v_0,θ^H} to achieve exact replication of the claim at the expiration date T, $V_T^{v_0,\theta^H} = H$, almost surely (a.s.). The cost of replication v_0 is given by the expected value of the contingent claim under the unique equivalent martingale measure.

In an incomplete market and/or with constraints on trading strategies, not every contingent claim is attainable. However, it is still possible to hedge without risk the contingent claim H at time T, whenever one starts with a large enough initial wealth x:

$$V_T^{x,\theta} \geq H, \quad \text{a.s. for some trading strategy } \theta. \tag{1.1}$$

The least initial wealth v_0 allowing (1.1) is called superreplication cost of the contingent claim, and the corresponding strategy θ^H such that $V_T^{v_0,\theta^H} \geq H$, a.s. is called superhedging strategy of H. The superreplication cost v_0 is given by the supremum of the expected values over a suitable set of equivalent martingale measures; see El Karoui and Quenez [13], Kramkov [26], Schäl [36], Föllmer and Kabanov [14] in an incomplete market context, and Cvitanić and Karatzas [5] and Föllmer and Kramkov [15] in a continuous time model with convex constraints on trading strategies. But in many situations, the superhedging strategy is the trivial buy-and-hold strategy and the superreplication cost is too high; see, e.g., Eberlein and Jacod [11] and Cvitanić, Pham, and Touzi [7].

We consider the position of an agent who is unwilling to commit the initial amount required for a superhedging strategy. Then the contingent claim carries an intrinsic risk and the question is how to quantify this risk. Various criteria have been proposed in the literature. The mean-variance hedging approach, initiated by Duffie and Richardson [10] and studied by Schäl [35], Schweizer [38, 39], Monat and Stricker [29],

*Received by the editors June 25, 1998; accepted for publication (in revised form) September 22, 1999; published electronically February 24, 2000.
http://www.siam.org/journals/sicon/38-3/34109.html
†Laboratoire de Probabilités et Modéles Aleatoires, UMR 7599, Université Paris 7, Case 7012, 2 Place Jussieu, 75251 Paris Cedex 05, France (pham@gauss.math.jussieu.fr).

Gouriéroux, Laurent, and Pham [18], and Rheinländer and Schweizer [31], among others, consists of approximating in $L^2(P)$-norm the contingent claim H by the terminal value of a self-financed wealth process:

$$\inf_\theta E\left[\left(H - V_T^{x,\theta}\right)^2\right]. \tag{1.2}$$

Expectation in (1.2) is under the objective probability P. The main criticism of this criterion is that it gives equal weight to upside and downside risks. In a general continuous time semimartingale model, Föllmer and Leukert [16] consider the problem of maximizing the probability of a hedge without risk,

$$\sup_\theta P\left[V_T^{x,\theta} \geq H\right], \tag{1.3}$$

by making use of the Neyman–Pearson lemma. This criterion does not take into account the size of the shortfall $(H - V_T^{x,\theta})^+$ but only the probability of its occurrence. In the context of a complete diffusion model and in the spirit of the paper of Artzner et al. [1], Cvitanić and Karatzas [6] study the following measure of the risk of the contingent claim H,

$$\inf_\theta E\left[\left(H - V_T^{x,\theta}\right)^+\right], \tag{1.4}$$

by adopting a convex duality approach. Notice that in both papers of Föllmer and Leukert [16, 17] and in Cvitanić and Karatzas [6], wealth processes have to satisfy a fixed lower bound requirement.

In this paper, we consider a general discrete time price process of a financial market. We also impose cone constraints on trading strategies. This includes the incomplete market case as well as the case of no short-selling constraints on portfolios. Given a contingent claim H in $L^p(P)$ and a loss function $l_p(x) = x^p/p$, $x \geq 0$, $p \in (1, \infty)$, we propose to measure the intrinsic risk of H by the quantity

$$J(x) = \inf_\theta E\left[l_p\left(\left(H - V_T^{x,\theta}\right)^+\right)\right]. \tag{1.5}$$

This is the smallest expected l_p-loss function of the shortfall of the contingent claim that can be achieved by trading in the financial market. Our purpose is to study this stochastic control problem (1.5) by deriving some properties of the value function J and by characterizing the trading strategy, if it exists, that attains the infimum. Recently, Föllmer and Leukert [17] have studied the problem of minimizing $E[l(H - V_T^{x,\theta})^+]$ for a general loss function l and in a general continuous time incomplete model, focusing on the Neyman–Pearson approach. In Cvitanić [3], convex duality approach is applied for a linear loss function, $l(x) = x$, in the context of an incomplete or constrained diffusion model.

We present in section 2 the model and the precise formulation of the problem. Within this $L^p(P)$-framework with cone constraints, we introduce in section 3 sets of martingale measures related to the no arbitrage condition and we state duality results on the superreplication cost. In section 4, we combine the technique of superhedging and a convex duality approach to solve the stochastic control problem (1.5). Our main result states that the optimal hedging strategy is given by the superhedging strategy of the modified contingent claim $H - l_q'(y^*(x)Z^*(y^*(x))$, where q is the conjugate of

p; i.e., $1/p + 1/q = 1$, $y^*(x) > 0$ is some Lagrange multiplier, and $Z^*(y^*(x))$ is a martingale measure, depending on H in general, solution of a dual problem whose existence is proved. Section 5 contains applications and examples of our results. We first show that the well-known equivalence between uniqueness of a martingale measure and attainability of every contingent claim in an unconstrained market can be extended to the case of symmetrical cone constraints on trading strategies. The determination of the optimal hedging strategy in this framework of complete market model is illustrated with the example of the Cox–Ross–Rubinstein model. In a second application, we note a case where H is attainable in a model with symmetrical cone constraints. Then, the dual problem leads to the problem of the so-called $L^q(P)$-optimal martingale measure, which does not depend on H. Finally, we consider the problem of hedging a riskless asset and we give an explicit example of computation of the optimal hedging strategy in a no short-selling constraints model.

2. Formulation of the problem. Let $(\Omega, \mathcal{F}, P)$ be a probability space with a filtration $\mathbb{F} = \{\mathcal{F}_k,\ k = 0, \dots, T\}$ for some $T \in \mathbb{N}$. For simplicity, we assume that $\mathcal{F}_0$ is trivial. The evolution of the discounted price process of d risky assets is modeled by an $\mathbb{R}^d$-valued $\mathbb{F}$-adapted stochastic process $S = \{S_k,\ k = 0, \dots, T\}$. A trading strategy is an $\mathbb{R}^d$-valued process $\theta = \{\theta_k,\ k = 1, \dots, T\}$, $\mathbb{F}$-predictable; i.e., θ_k is $\mathcal{F}_{k-1}$-measurable $\forall k = 1, \dots, T$. Here θ_k represents the number of units of risky assets held by the investor during $(k-1, k]$ $\forall k = 1, \dots, T$. We denote by Θ the set of all trading strategies. Given an initial wealth $x \in \mathbb{R}$ and a trading strategy $\theta \in \Theta$, we define the discounted (self-financed) wealth process $V^{x,\theta}$ by

$$V_k^{x,\theta} = x + \sum_{j=1}^{k} \theta_j . \Delta S_j, \quad k = 1, \dots, T,$$
$$V_0^{x,\theta} = x,$$

where $\Delta S_j = S_j - S_{j-1}$. Here given two elements $a, b \in \mathbb{R}^d$, $a.b$ denotes their Euclidean scalar product. When using discounted prices and a discounted self-financed wealth process, we assume there is an additional bank account whose interest rate is zero. This reduced form is explained in detail in Harrison and Kreps [19]. Notice that by considering discounted prices, we leave aside all problems related to the case of stochastic interest rate. These questions typically arise in the context of foreign exchange markets. This would certainly be an important extension of our model but is beyond the scope of the present paper.

Let us consider a nonempty closed convex cone C in $\mathbb{R}^d$. A trading strategy θ is called C-constrained if

$$\theta_k \in C, \qquad k = 1, \dots, T,\ P \text{ a.s.}$$

We denote by $\Theta(C)$ the set of all C-constrained trading strategies. The role of the convex cone C is to model some constraints on the trading strategies. We denote by C^o the polar cone of C. Since C is a cone, we have (see [32])

$$C^o = \{b \in \mathbb{R}^d : a.b \leq 0\ \forall a \in C\}.$$

The following are some examples of such constraints:

(a) *Unconstrained case*: $C = \mathbb{R}^d$. Then $C^o = \{0\}$.

(b) *Prohibition of short-selling of some risky assets*: $C = \{a \in \mathbb{R}^d : a^i \geq 0\ \forall i \in I\}$, where I is some subset of $\{1, \dots, d\}$. Then $C^o = \{b \in \mathbb{R}^d : b^i \leq 0\ \forall i \in I, b^i = 0 \text{ otherwise}\}$.

Consider now a contingent claim given by an $\mathcal{F}_T$-measurable random variable H. Given an initial wealth x and a trading strategy θ, the resulting shortfall of H, defined as the excess of the contingent claim over the final portfolio value, is $(H - V_T^{x,\theta})^+$. We are interested in the problem of minimizing the $L^p(P)$-norm $(1 < p < \infty)$ of the shortfall of a contingent claim $H \in L^p(P)$. We introduce the subset of trading strategies

$$\Theta_p(C) = \left\{ \theta \in \Theta(C) : \exists X \in L^p(P),\ V_T^{0,\theta} \geq X,\ P\ a.s. \right\},$$

and we consider the following optimization problem:

$$(\mathcal{P}_p(x))J(x) \ = \ \inf_{\theta \in \Theta_p(C)} E\left[l_p\left(\left(H - V_T^{x,\theta} \right)^+ \right) \right], \quad x \in \mathbb{R},$$

where l_p is a loss function defined on $\mathbb{R}_+$ by $l_p(y) = y^p/p$. In this L^p-framework, we shall assume that $S_k \in L^p(P)$, $k = 0, \ldots, T$. We mention that the main result of the paper remains valid under the more general case when S is locally in $L^p(P)$ (see Remark 3.1). We analyze problem $(\mathcal{P}_p(x))$ by combining the technique of superhedging in a L^p-framework under cone constraints with a convex duality approach.

Remark 2.1. Föllmer and Leukert [17] study the problem of minimizing $E[l(H - V_T^{x,\theta})^+]$ for a general loss function l and in a general continuous time incomplete model. They focus mainly on the Neyman–Pearson lemma approach as in their previous paper [16]. Cvitanić and Karatzas [6] and Cvitanić [3] analyze a similar problem for a linear loss function, $l(x) = x$, in a diffusion model including also constraints on strategies and margin requirements. Notice that in these papers a nonnegativity constraint and, more generally, a given lower bound constraint on the wealth process is imposed. Here, since we allow for contingent claims with a nonconstant sign, we consider a more general admissibility condition by requiring only a lower bound on the terminal wealth value by some arbitrary $L^p(P)$ random variable.

3. Martingale measures, arbitrage, and superhedging strategy. In an incomplete market model, there is a known duality relation between the superreplication cost of a contingent claim H, defined as the least initial wealth for hedging H without risk, and the largest arbitrage-free prices defined as the supremum of expectations of H under a set of martingale measures; see, e.g., Delbaen and Schachermayer [9], El Karoui and Quenez [13], Kramkov [26], Schäl [36], and Föllmer and Kabanov [14]. In markets with convex constraints on trading strategies, analogous results are generalized by Cvitanić and Karatzas [5] in a diffusion model and by Föllmer and Kramkov [15] in a general continuous time semimartingale model; see also Brannath [2] in a discrete time framework. Most of the cited papers consider the case of nonnegative contingent claims. For the sake of completeness and the purpose of our paper, we state similar results in a general discrete time model with cone constraints and for a contingent claim in $L^p(P)$.

Let us introduce some notations. By $L^{0,d}(\mathcal{F}_k, P)$, we denote the space of all $\mathbb{R}^d$-valued $\mathcal{F}_k$-measurable random variables. When $d = 1$, we shall omit exponent d. We consider the convex cone K of $L^0(P) := L^0(\mathcal{F}_T, P)$,

$$K = \left\{ V_T^{0,\theta} : \theta \in \Theta(C) \right\},$$

and we say that there is no arbitrage opportunity if

$$\text{(NA)} \quad K \cap L^0_+(P) = \{0\}.$$

In the unconstrained case on trading strategies, there is a basic result relating the no arbitrage condition (NA) to the existence of some equivalent martingale measure. This duality result, known as the fundamental theorem of asset pricing, was proved by Dalang, Morton and Willinger [8], Schachermayer [34], Kabanov and Kramkov [24], and Rogers [33] in a discrete time framework. Extensions of this result to the constrained case are proved by Jouini and Kallal [23] and Schürger [37] in a no short-selling constraints model, by Pham and Touzi [30] in the case of cone constraints on portfolios, and by Brannath [2] in the case of convex constraints. For the sake of completeness, we briefly recall these results.

In the rest of this paper, for any $p \in [1, \infty]$, we denote by $q \in [1, \infty]$ its conjugate, i.e., $1/p + 1/q = 1$. Given $p \in [1, \infty]$, we introduce the following set of martingale measures:

$$\mathcal{M}_q(P) = \left\{ Q \ll P : \frac{dQ}{dP} \in L^q(P), \text{ and } E^Q[\Delta S_k | \mathcal{F}_{k-1}] \in C^o,\ k = 1, \dots, T,\ Q \text{ a.s.} \right\},$$

$$\mathcal{M}_q^e(P) = \{ Q \sim P : Q \in \mathcal{M}_q(P) \}.$$

In the unconstrained case $C = \mathbb{R}^d$, $\mathcal{M}_q^e(P)$ is the set of equivalent probability measures with density in $L^q(P)$ under which S is a martingale. In the no short-selling constraints case $C = (0, \infty)^d$, $\mathcal{M}_q^e(P)$ is the set of equivalent probability measures with density in $L^q(P)$ under which S is a supermartingale. Writing that $V_k^{0,\theta} = V_{k-1}^{0,\theta} + \theta_k.\Delta S_k$, it is easily seen that for any $\theta \in \Theta_p(C)$, $V^{0,\theta}$ is a supermartingale under any $Q \in \mathcal{M}_q(P)$. Since $\Delta S_k \in L^p(P)$, $k = 1, \dots, T$, it is clear that $\mathcal{M}_q(P)$ is closed in $L^q(P)$ (when identifying a probability measure $Q \ll P$ with its Radon–Nikodym dQ/dP).

Remark 3.1. In the case where S is assumed only to be local in $L^p(P)$, one introduces the set

$$\mathcal{M}_q^{loc}(P) = \left\{ Q \ll P : \frac{dQ}{dP} \in L^q(P),\ E^Q[|\Delta S_k||\mathcal{F}_{k-1}] < \infty, \text{ and } E^Q[\Delta S_k | \mathcal{F}_{k-1}] \in C^o,\ k = 1, \dots, T,\ Q \text{ a.s.} \right\},$$

which is equal to $\mathcal{M}_q(P)$ when $S \in L^p(P)$. It can be proved that $\mathcal{M}_q^{loc}(P)$ is closed in $L^q(P)$.

As is now standard in the literature on arbitrage and superreplication, a key result is to state the closedness property of the set of dominated final payoffs, $K - L_+^0(P)$ for the topology of convergence in probability. While this property is satisfied in the unconstrained case (see, e.g., Schachermayer [34]), this does not hold true in the presence of constraints on strategies (see Brannath [2]). We shall then make a nondegeneracy assumption on the price process in order to ensure this closedness property. Let us introduce, $\forall k = 1, \dots, T$, the subsets of $L^{0,d}(\mathcal{F}_{k-1}, P)$:

$$N(k-1) = \left\{ \eta \in L^{0,d}(\mathcal{F}_{k-1}, P) : \eta.\Delta S_k = 0,\ P \text{ a.s.} \right\}.$$

We assume that

(B) $N(k-1) = \{0\}\ \forall k = 1, \dots, T$.

Remark 3.2. Notice that if $S \in L^2(P)$, then condition (B) is obviously implied by the nondegeneracy condition on the price process:

(B′) $\mathrm{Var}(S_k|\mathcal{F}_{k-1})$ is invertible $\forall k = 1, \ldots, T$.

Condition (B′) is satisfied by most models considered in the financial literature. For instance, suppose that $\{S_k, k = 1, \ldots, T\}$ is obtained from the geometric Brownian motion $dS_t = S_t(\mu dt + \sigma dW_t)$ as in the Black–Scholes model. Then, a straightforward calculation shows that

$$\mathrm{Var}\left(S_k|\mathcal{F}_{k-1}\right) = S_{k-1}^2 e^{2\mu}\left(e^{\sigma^2} - 1\right),$$

and the validity of the hypothesis of invertibility of $\mathrm{Var}\left(S_t|\mathcal{F}_{t-1}\right)$ is guaranteed whenever $\sigma > 0$. It is also easy to check that condition (B′) is satisfied in the Cox–Ross–Rubinstein model (see section 5).

Under condition (B), we have the dual characterization of the no arbitrage condition (NA): there is no arbitrage opportunity if and only if $\mathcal{M}_\infty^e(P) \neq \emptyset$ (see Pham and Touzi [30], Brannath [2]).

We define the superreplication cost of a contingent claim $H \in L^p(P)$, $p \in [1, \infty]$, by

$$v_0 = \inf\left\{x \in \mathbb{R} : \ \exists \theta \in \Theta_p(C),\ V_T^{x,\theta} \geq H,\ P \text{ a.s.}\right\}.$$

The next result provides a dual characterization of the superreplication cost in terms of martingale measures within a $L^p(P)$-framework in a general discrete time model with cone constraints portfolios. In what follows, to alleviate notations, we write $\mathcal{M}_q = \mathcal{M}_q(P)$.

THEOREM 3.1. *Assume that condition* (NA) *is satisfied and assumption* (B) *holds. Then the superreplication cost of a contingent claim $H \in L^p(P)$, $p \in [1, \infty]$, is given by*

$$v_0 = \sup_{Q \in \mathcal{M}_q^e} E^Q[H]. \tag{3.1}$$

In (3.1)*, the supremum can also be taken with respect to $\mathcal{M}_q$. Moreover, we have the following assertions:*

(1) *Suppose that $\sup_{Q \in \mathcal{M}_q^e} E^Q[H] < \infty$; then there exists $\theta^H \in \Theta_p(C)$ such that $V_T^{v_0,\theta^H} \geq H$, P a.s. θ^H is called superhedging strategy for H.*

(2) *Suppose that the supremum in* (3.1) *is attained for some $\hat{Q} \in \mathcal{M}_q^e$. Then the superhedging strategy θ^H is actually a replicating strategy, i.e., $V_T^{v_0,\theta^H} = H$, P a.s.*

Proof. See Appendix A. □

4. $L^p(P)$-hedging and dual formulation. In a first step, we reformulate the dynamic problem $(\mathcal{P}_p(x))$ into an equivalent static problem by using results of section 3 on superhedging strategy. For $p \in (1, \infty)$, $x \in \mathbb{R}$, and a given contingent claim H in $L^p(P)$, we denote

$$\mathcal{C}_p(x) = \left\{X \in L^p(P) : X \leq H,\ P \text{ a.s. and } E[ZX] \leq x\ \forall Z \in \mathcal{M}_q\right\}.$$

Here we identify a probability measure Q in $\mathcal{M}_q$ with its Radon–Nikodým density $Z = dQ/dP$. We consider then the static problem

$$(\mathcal{S}_p(x)) \quad \inf_{X \in \mathcal{C}_p(x)} E\left[l_p\left(H - X\right)\right].$$

PROPOSITION 4.1. *Assume that condition* (NA) *is satisfied and assumption* (B) *holds.*

(1) *The value functions of problems* $(\mathcal{P}_p(x))$ *and* $(\mathcal{S}_p(x))$, $x \in \mathbb{R}$, *coincide:*

$$J(x) = \inf_{\theta \in \Theta_p(C)} E\left[l_p\left(\left(H - V_T^{x,\theta}\right)^+\right)\right] = \inf_{X \in \mathcal{C}_p(x)} E\left[l_p\left(H - X\right)\right] \quad \forall\, x \in \mathbb{R}.$$

(2) *Assume that* $X^*(x)$ *is a solution to problem* $(\mathcal{S}_p(x))$. *Then there exists a superhedging strategy* $\theta^*(x)$ *for* $X^*(x)$ *and* $\theta^*(x)$ *solves the dynamic problem* $(\mathcal{P}_p(x))$.

(3) *Conversely, assume that* $\theta^*(x)$ *is a solution to problem* $(\mathcal{P}_p(x))$. *Then* $X^*(x) = H - (H - V_T^{x,\theta^*(x)})^+$ *solves problem* $(\mathcal{S}_p(x))$.

Proof. (1) Let $\theta \in \Theta_p(C)$. Then $X = H - (H - V_T^{x,\theta})^+ = \min(H, V_T^{x,\theta}) \in L^p(P)$. We have $X \le H$ and $X \le V_T^{x,\theta}$, so that by the supermartingale property of $V^{x,\theta}$ under $Q = Z.P \in \mathcal{M}_q$, $E[ZX] \le x$. It follows that $X \in \mathcal{C}_p(x)$. Denoting by $\bar{J}$ the value function of problem $(\mathcal{S}_p(x))$, $x \in \mathbb{R}$, we have

$$E\left[l_p\left(\left(H - V_T^{x,\theta}\right)^+\right)\right] = E\left[l_p\left(H - X\right)\right] \ge \bar{J}(x),$$

and therefore $J(x) \ge \bar{J}(x)$. Conversely, let $X \in \mathcal{C}_p(x)$. We then have

$$x_0 := \sup_{Z \in \mathcal{M}_q} E[ZX] \le x < \infty.$$

We deduce by Theorem 3.1 that there exists $\theta^X \in \Theta_p(C)$, superhedging strategy for X, such that

$$V_T^{x,\theta^X} \ge V_T^{x_0,\theta^X} \ge X, \tag{4.1}$$

and therefore, recalling that $X \le H$,

$$\left(H - V_T^{x,\theta^X}\right)^+ \le H - X. \tag{4.2}$$

Now, since the function l_p is nondecreasing, we obtain

$$J(x) \le E\left[l_p\left(\left(H - V_T^{x,\theta^X}\right)^+\right)\right] \le E\left[l_p\left(H - X\right)\right],$$

which proves that $J(x) \le \bar{J}(x)$ and finally the equality $J = \bar{J}$.

(2) As in (4.1)–(4.2), there exists a superhedging strategy $\theta^*(x)$ for $X^*(x)$, and we have

$$\left(H - V_T^{x,\theta^*(x)}\right)^+ \le H - X^*(x);$$

hence,

$$E\left[l_p\left(\left(H - V_T^{x,\theta^*(x)}\right)^+\right)\right] \le E\left[l_p\left(H - X^*(x)\right)\right] = J(x),$$

which proves that $\theta^*(x)$ solves problem $(\mathcal{P}_p(x))$.

(3) As in the proof of (1), the random variable $X^*(x) = H - (H - V_T^{x,\theta^*(x)})^+$ lies in $\mathcal{C}_p(x)$ and we have

$$E\left[l_p\left(H - X^*(x)\right)\right] = E\left[l_p\left(\left(H - V_T^{x,\theta^*(x)}\right)^+\right)\right] = J(x),$$

which proves that $X^*(x)$ solves problem $(\mathcal{S}_p(x))$. ∎

Remark 4.1. A trivial case where problem $(\mathcal{P}_p(x))$ is directly solved arises when the investor has an initial wealth larger than the superreplication cost of the contingent claim; i.e., $x \geq v_0 = \sup_{Z\in\mathcal{M}_q} E[ZH]$. Then the solution to problem $(\mathcal{S}_p(x))$ is obviously $X^*(x) = H$ and a solution to problem $(\mathcal{P}_p(x))$ is the superhedging strategy θ^H for H. In this case, the resulting shortfall of H is zero and $J(x) = 0$.

In Föllmer and Leukert [17], wealth processes are required to be nonnegative. This admissibility condition is crucial for solving the static problem associated to (1.3) thanks to the Neyman–Pearson lemma. Here in our L^p-framework, we do not impose a fixed lower bound on wealth process so that the static problem $(\mathcal{S}_p(x))$ can not be solved by using the Neyman–Pearson lemma. In what follows, we fix $p \in (1,\infty)$, $H \in L^p(P)$, and we study the static problem $(\mathcal{S}_p(x))$ for an initial capital $x < v_0$, by adopting a convex duality approach. Starting with the strictly convex function $x \mapsto l_p(H-x)$, defined on the random interval $(-\infty, H]$, we consider its stochastic Fenchel–Legendre transform:

$$\tilde{L}(y,\omega) = \max_{x\leq H}[-l_p(H-x) - xy] \;=\; l_q(y) - yH, \qquad y > 0. \tag{4.3}$$

The maximum in expression (4.3) is attained by the random $\mathcal{F}_T$-measurable function:

$$\chi(y,\omega) = H - l_q'(y), \quad y > 0. \tag{4.4}$$

We consider then the dual problem of $(\mathcal{S}_p(x))$:

$$(\mathcal{D}_q(y))\tilde{J}(y) \;=\; \inf_{Z\in\mathcal{M}_q} E\left[\tilde{L}(yZ,\omega)\right].$$

Notice that compared to the dual problem arising in the problem of maximizing the expected utility of terminal wealth in an incomplete market (see Karatzas et al. [25], He and Pearson [21], Cvitanić and Karatzas [4], and Kramkov and Schachermayer [27]), there is an additional term depending on the contingent claim in the formulation of the dual problem $(\mathcal{D}_q(y))$. We mention that Föllmer and Leukert [17] also use a convex duality approach in their context of nonnegative wealth processes.

The next proposition states the existence of a unique solution to the dual problem.

PROPOSITION 4.2. *Assume that $\mathcal{M}_q \neq \emptyset$. Then $\forall y > 0$, there exists a unique solution $Z^*(y)$ to the dual problem $(\mathcal{D}_q(y))$. Moreover, the function $\tilde{J}$ is differentiable and strictly convex on $(0,\infty)$, and we have*

$$\tilde{J}'(y) = -E\left[Z^*(y)\chi(yZ^*(y),\omega))\right], \quad y > 0. \tag{4.5}$$

Proof. Let $(Z_n)_n$ be a sequence in $\mathcal{M}_q$ such that

$$\lim_{n\to\infty} E\left[\tilde{L}(yZ_n,\omega)\right] = \tilde{J}(y). \tag{4.6}$$

Then, for n sufficiently large, we have

$$E\left[l_q(yZ_n) - yZ_nH\right] \leq \tilde{J}(y) + 1. \tag{4.7}$$

Since $\mathcal{M}_q \neq \emptyset$, we have $\tilde{J}(y) < \infty \;\forall\; y > 0$. By Hölder inequality and since $H \in L^p(P)$, we deduce from (4.7) that

$$\| Z_n \|^q_{L^q(P)} \leq \text{const.}\left(1 + \| Z_n \|_{L^q(P)}\right), \tag{4.8}$$

where const. is a positive constant independent of n. Since $q > 1$, inequality (4.8) proves that $(Z_n)_n$ is bounded in $L^q(P)$. This implies that Z_n converges weakly in $L^q(P)$ to some $Z^*(y)$, possibly along a subsequence. By Mazur's theorem (see Ekeland and Temam [12, Chap. I], there exists a sequence $\tilde{Z}_n \in \text{conv}(Z_n, Z_{n+1}, \ldots)$ such that $\tilde{Z}_n$ converges to $Z^*(y)$ in $L^q(P)$-norm. It is clear that $\tilde{Z}_n \in \mathcal{M}_q$. Moreover, $\mathcal{M}_q$ is closed in $L^q(P)$ and so $Z^*(y) \in \mathcal{M}_q$. From the convexity of the function $z \in \mathbb{R}_+ \mapsto \tilde{L}(yz, \omega)$, P a.s., we have

$$E\left[\tilde{L}(y\tilde{Z}_n, \omega)\right] \leq \sup_{k \geq n} E\left[\tilde{L}(yZ_k, \omega)\right],$$

so that by (4.6), and since $\tilde{L}(y\tilde{Z}_n, \omega)$ converges to $\tilde{L}(yZ^*(y), \omega)$ in $L^1(P)$,

$$E\left[\tilde{L}(yZ^*(y), \omega)\right] = \lim_{n \to \infty} E\left[\tilde{L}(y\tilde{Z}_n, \omega)\right] = \tilde{J}(y),$$

which proves that $Z^*(y)$ is solution to $(\mathcal{D}_q(y))$. The uniqueness of the solution to $(\mathcal{D}_q(y))$ follows from the strict convexity of the function $z \in \mathbb{R}_+ \mapsto \tilde{L}(yz, \omega)$, P a.s. $\forall\ y > 0$. The strict convexity of $\tilde{J}$ is proved by same arguments as in Kramkov and Schachermayer [27, Lemma 3.5] by using the strict convexity of the function $x \in \mathbb{R}_+ \mapsto \tilde{L}(x, \omega)$, P a.s.

Let $y > 0$. Then $\forall\ h > 0$, we have

$$\begin{aligned}\frac{\tilde{J}(y+h) - \tilde{J}(y)}{h} &\leq \frac{1}{h} E\left[\tilde{L}((y+h)Z^*(y), \omega) - \tilde{L}(yZ^*(y), \omega)\right] \\ &= \frac{l_q(y+h) - l_q(y)}{h} E\left[Z^*(y)^q\right] - E\left[Z^*(y)H\right].\end{aligned}$$

This shows that

$$\limsup_{h \searrow 0^+} \frac{\tilde{J}(y+h) - \tilde{J}(y)}{h} \leq -E\left[Z^*(y)\chi(yZ^*(y), \omega)\right]. \tag{4.9}$$

Similarly, $\forall h < 0$, $y + h > 0$, we have

$$\frac{\tilde{J}(y+h) - \tilde{J}(y)}{h} \geq \frac{l_q(y+h) - l_q(y)}{h} E\left[Z^*(y)^q\right] - E\left[Z^*(y)H\right],$$

which shows that

$$\liminf_{h \nearrow 0^-} \frac{\tilde{J}(y+h) - \tilde{J}(y)}{h} \geq -E\left[Z^*(y)\chi(yZ^*(y), \omega)\right]. \tag{4.10}$$

Finally, (4.9)–(4.10) and convexity of the function $\tilde{J}$ imply the differentiability of $\tilde{J}$ and provides the expression (4.5) of $\tilde{J}'$. □

Remark 4.2. Cvitanić [3] also uses a convex duality approach in the case of a linear loss function. In this context, the value function of the dual problem fails to be everywhere differentiable.

LEMMA 4.1. *Under the assumptions of Proposition* 4.2, $\forall x < v_0$, *there exists a unique* $y^*(x) > 0$ *satisfying* $\tilde{J}'(y^*(x)) = -x$, *i.e.,*

$$E\left[Z^*(y^*(x))\chi\left(y^*(x)Z^*(y^*(x)), \omega\right)\right] = x. \tag{4.11}$$

Proof. The function $y \mapsto f_x(y) = \tilde{J}(y) + xy$ is (strictly) convex on $(0,\infty)$. We have $f_x(0^+) = 0$. Moreover, by Jensen's inequality and since $E[Z^*(y)] = 1$, we have

$$\tilde{J}(y) \geq l_q(y) - yE\left[Z^*(y)H\right];$$

hence,

$$f_x(y) \geq l_q(y) - y\left(E\left[Z^*(y)H\right] - x\right).$$

Since $q > 1$, this shows that $f_x(y) \to \infty$ as y goes to infinity. Let us now check that there exists $y_0 > 0$ such that $f_x(y_0) < 0$. If not, we should have

$$E\left[l_q(yZ) - yZH\right] + xy \geq 0 \quad \forall y > 0,\ \forall Z \in \mathcal{M}_q;$$

hence,

$$\frac{l_q(y)}{y} E\left[Z^q\right] + x \geq E\left[ZH\right] \quad \forall y > 0,\ \forall Z \in \mathcal{M}_q.$$

By sending y to zero, and from Theorem 3.1, we obtain

$$x \geq \sup_{Z \in \mathcal{M}_q} E\left[ZH\right] = v_0,$$

which is in contradiction to the fact that $x < v_0$. We have thus proved the existence of $y^*(x) > 0$ which attains the infimum of $f_x(y)$ over $y > 0$. The uniqueness of $y^*(x)$ follows from the strict convexity of f_x. Finally, since $\tilde{J}$ (and so f_x) is differentiable, we have $f_x'(y^*(x)) = 0$ or $\tilde{J}'(y^*(x)) = -x$, which provides the relation (4.11) by (4.5). □

We now state the main result of this paper by proving the existence of a unique solution to the dynamic problem $(\mathcal{P}_p(x))$ and relating it to the solution of a dual problem. Recall that when $x \geq v_0$, a solution to the problem $(\mathcal{P}_p(x))$ is the superhedging strategy for H; see Remark 4.1. In the case $x < v_0$, we have the following result.

THEOREM 4.1. *Assume that condition* (NA) *is satisfied and assumption* (B) *holds.*

(1) *For all $x < v_0$, there exists a unique solution $X^*(x)$ to problem $(\mathcal{S}_p(x))$, and we have the dual relation*

$$X^*(x) = \chi\left(y^*(x)Z^*(y^*(x)), \omega\right) \ = \ H - \left(y^*(x)Z^*(y^*(x))\right)^{q-1}, \tag{4.12}$$

where $y^*(x)$ is given by (4.11) *and $Z^*(y^*(x))$ is the solution to the dual problem $(\mathcal{D}_q(y^*(x)))$. Let $\theta^*(x)$ denote the superhedging strategy for the contingent claim $X^*(x) \in L^p(P)$. Then $\theta^*(x)$ is a solution to problem $(\mathcal{P}_p(x))$.*

(2) *The function J is differentiable, strictly decreasing, and strictly convex on $(-\infty, v_0)$, and we have the dual relations*

$$J(x) = \max_{y>0}\left[-\tilde{J}(y) - xy\right], \quad x < v_0, \tag{4.13}$$

$$\tilde{J}(y) = \max_{x<v_0}\left[-J(x) - xy\right], \quad y > 0, \tag{4.14}$$

$$J'(x) = -y^*(x), \quad x < v_0. \tag{4.15}$$

Proof. (1) Fix some $y > 0$ and let Z be an arbitrary element of $\mathcal{M}_q$. Denote

$$Z_\varepsilon = (1-\varepsilon)Z^*(y) + \varepsilon Z, \quad \varepsilon \in (0,1).$$

We have $Z_\varepsilon \in \mathcal{M}_q$ so that

$$\begin{aligned} 0 &\leq \frac{1}{\varepsilon} E\left[\tilde{L}(yZ_\varepsilon, \omega) - \tilde{L}(yZ^*(y), \omega)\right] \\ &\leq yE\left[\left(l_q'(yZ_\varepsilon) - H\right)\left(Z - Z^*(y)\right)\right], \end{aligned}$$

where the second inequality follows from the convexity of the function $y \mapsto \tilde{L}(y, \omega)$ and the fact that its derivative is equal to $l_q'(y) - yH$. Therefore we obtain

$$E\left[l_q'(yZ_\varepsilon)\left(Z^*(y) - Z\right)\right] \leq E\left[H\left(Z^*(y) - Z\right)\right]. \tag{4.16}$$

Notice that $l_q'(yZ_\varepsilon)\left(Z^*(y) - Z\right) \geq l_q'(yZ)\left(Z^*(y) - Z\right) \geq -l_q'(y)Z^q$ which is integrable. Hence, by Fatou's lemma, sending ε to zero in (4.16), we get

$$E\left[l_q'(yZ^*(y))\left(Z^*(y) - Z\right)\right] \leq E\left[H\left(Z^*(y) - Z\right)\right].$$

This last relation shows that

$$E\left[Z\chi(yZ^*(y), \omega)\right] \leq x_y := E\left[Z^*(y)\chi(yZ^*(y), \omega)\right] \quad \forall Z \in \mathcal{M}_q. \tag{4.17}$$

Therefore, $\chi(yZ^*(y), \omega) \in \mathcal{C}_p(x_y)$. Moreover, by (4.3)–(4.4), we have $\forall X \in \mathcal{C}_p(x_y)$

$$\tilde{L}(yZ^*(y), \omega) = -l_p(H - \chi(yZ^*(y), \omega)) - yZ^*(y)\chi(yZ^*(y), \omega), \tag{4.18}$$

$$\geq -l_p(H - X) - yZ^*(y)X; \tag{4.19}$$

hence,

$$E\left[l_p(H - \chi(yZ^*(y), \omega))\right] \leq E\left[l_p(H - X)\right], \tag{4.20}$$

since $E[Z^*(y)X] \leq x_y$. This shows that $\chi(yZ^*(y), \omega)$ is solution to problem $(\mathcal{S}_p(x_y))$. Now, by noting that for $y = y^*(x)$, we have $x_{y^*(x)} = x$ (see relation (4.11)) and $\chi(y^*(x)Z^*(y^*(x)), \omega) = X^*(x)$, we have proved that $X^*(x)$ is a solution to the static problem $(\mathcal{S}_p(x))$. The uniqueness of the solution to problem $(\mathcal{S}_p(x))$ follows from the strict convexity of the function $x \leq H \mapsto l_p(H - x)$. Now, from Proposition 4.1(2), there exists a superhedging strategy $\theta^*(x)$ for $X^*(x)$, and $\theta^*(x)$ is a solution to the dynamic problem $(\mathcal{P}_p(x))$.

(2) Proof of the duality results (4.13)–(4.15) is standard in utility optimization problem and is proved by same arguments as in Föllmer and Leukert [17, Thm. 7.3] or Kramkov and Schachermayer [27]. □

Part (1) of Theorem 4.1 states that the problem of $L^p(P)$-hedging a contingent claim H, when one has in hand an initial wealth strictly smaller than the superreplication cost of H, is solved by superhedging a modified option $X^*(x)$ given in (4.12) and expressed as the difference of H and a nonnegative contingent claim related to the density of an appropriate martingale measure. In Föllmer and Leukert [17] where wealth processes are required to be nonnegative, the optimal hedging strategy is the superhedging strategy for a modified option of the form $(X^*(x))^+$. Part (2) of Theorem 4.1 gives some properties of the measure of risk J, related to those of Artzner et al. [1].

5. Applications and examples.

5.1. The complete market case. We say that a contingent claim $H \in L^p(P)$ is attainable if there exist $x \in \mathbb{R}$ and $\theta \in \Theta_p(C)$ such that $H = V_T^{x,\theta}$. In this case, θ^H is called perfect replicating strategy for H. In the unconstrained case, $C = \mathbb{R}^d$, it is well known that if the set of equivalent martingale measures is reduced to a singleton, then every contingent claim is attainable, and vice versa; see, e.g., Harrison and Pliska [20] and Jacka [22]. We extend such results in the case of cone constraints. We say that C is a symmetrical cone if $\forall y \in C$, $-y \in C$. An example of a symmetrical cone in the case $d = 2$ is $C = \{(x, y) \in \mathbb{R}^2 : xy \geq 0\}$. This means that the investor is constrained to have either short or long positions in both assets.

PROPOSITION 5.1. *Let $p \in [1, \infty]$ and assume that* (B) *holds.*

(1) *Suppose that $\mathcal{M}_q^e = \{\hat{P}\}$. Then every contingent claim $H \in L^p(P)$ is attainable.*

(2) *Assume that C is a symmetrical cone and condition* (NA) *is satisfied. Suppose that every contingent claim $H \in L^p(P)$, with $p \leq 2$, is attainable. Then $\mathcal{M}_q^e$ is reduced to a singleton and $\mathcal{M}_q^e = \mathcal{M}_q = \{\hat{P}\}$.*

Proof. (1) By Theorem 3.1, since $x_0 := \sup_{Q \in \mathcal{M}_q^e} E^Q[H] = E^{\hat{P}}[H] < \infty$, there exists $\theta^H \in \Theta_p(C)$ such that

$$V_T^{x_0,\theta^H} \geq H. \tag{5.1}$$

By the supermartingale property of $V^{x,\theta}$ under $\hat{P}$, we have

$$E^{\hat{P}}\left[V_T^{x_0,\theta^H} - H\right] \leq x_0 - E^{\hat{P}}[H] = 0. \tag{5.2}$$

Relations (5.1) and (5.2) imply that $H = V_T^{x_0,\theta}$, P a.s.

(2) Let $\hat{P} \in \mathcal{M}_q^e$, which is nonempty by condition (NA), and suppose that there exists $Q \neq \hat{P}$ in $\mathcal{M}_q$. Setting $\hat{Z} = d\hat{P}/dP$ and $Z^Q = dQ/dP$, it follows that the contingent claim $H = \hat{Z} - Z^Q \in L^q(P) \subset L^p(P)$ (since $p \leq 2$). Hence H is attainable and there exist $x \in \mathbb{R}$ and $\theta \in \Theta_p(C)$ such that

$$\hat{Z} - Z^Q = V_T^{x,\theta}, \quad P \text{ a.s.} \tag{5.3}$$

Under the condition that C is a symmetrical cone and by noting that $V_T^{-x,-\theta} = -H \in L^p(P)$, we deduce that $-\theta \in \Theta_p(C)$. Therefore by the supermartingale property of $V^{x,\theta}$ and $V^{-x,-\theta}$ under $\hat{P}$ and Q, we obtain that $V^{x,\theta}$ is a martingale under $\hat{P}$ and Q. Using (5.3), this implies that

$$E^{\hat{P}}\left[\hat{Z} - Z^Q\right] = E^Q\left[\hat{Z} - Z^Q\right] = x,$$

and therefore that $E[(\hat{Z} - Z^Q)^2] = 0$, which proves that $Q = \hat{P}$, which is a contradiction. □

Remark 5.1. This last proposition means that when C is a symmetrical cone and under condition (NA) and assumption (B), the following assertions, for $p \in [1, 2]$, are equivalent:

(i) $\mathcal{M}_q^e = \{\hat{P}\}$,

(ii) every contingent claim $H \in L^p(P)$ is attainable.

Notice also that in this case, $\mathcal{M}_q = \mathcal{M}_q^e = \{\hat{P}\}$. We shall say that the market is complete. This extends the characterization of market completeness of Harrison and Pliska [20] and Jacka [22] to the symmetrical cone constraints.

In the complete market case, when $x \geq v_0 = E^{\hat{P}}[H]$, a solution to problem $(\mathcal{P}_p(x))$ is the perfect replicating strategy for H. When $x < v_0$, we have the following result for the $L^p(P)$-hedging of a contingent claim H in $L^p(P)$.

THEOREM 5.1. *Assume that C is a symmetrical cone and assumption* (B) *holds, and suppose that $\mathcal{M}_q = \mathcal{M}_q^e = \{\hat{P}\}$, $p \in (1, \infty)$. Let $x < v_0 = E^{\hat{P}}[H]$. Then the solution to problem $(\mathcal{P}_p(x))$ is the perfect replicating strategy for the contingent claim:*

$$X^*(x) = H - \left(y^*(x)\hat{Z}\right)^{q-1}, \tag{5.4}$$

where $\hat{Z} = d\hat{P}/dP$ and

$$y^*(x) = \left(\frac{v_0 - x}{E[\hat{Z}^q]}\right)^{\frac{1}{q-1}}. \tag{5.5}$$

Proof. Since $\mathcal{M}_q = \{\hat{P}\}$, the solution to the dual problem $(\mathcal{D}_q(y))$ is $Z^*(y) = \hat{Z} = d\hat{P}/dP \ \forall \ y > 0$. Hence $y^*(x)$ in (4.11) is explicitly given by (5.5) and the solution $X^*(x)$ to problem $(\mathcal{S}_p(x))$ is given by (5.4). By Proposition 5.1(1), $X^*(x) \in L^p(P)$ is attainable and so the superhedging strategy $\theta^*(x)$ for $X^*(x)$ is in fact the perfect replicating strategy, and it is the solution to the dynamic problem $(\mathcal{P}_p(x))$ by Theorem 4.1. □

Remark 5.2. By Remark 5.1, if $\mathcal{M}_q^e = \{\hat{P}\}$, with $p \leq 2$, then $\mathcal{M}_q = \mathcal{M}_q^e = \{\hat{P}\}$.

In the unconstrained case, $C = \mathbb{R}^d$, the perfect replicating strategy for $X^*(x)$ is $\theta^*(x) = \theta^H - \tilde{\theta}(x)$, where θ^H is the perfect replicating strategy for H and $\tilde{\theta}(x)$ is the perfect replicating strategy for $(y^*(x)\hat{Z})^{q-1}$.

Example: The Cox–Ross–Rubinstein model. Given the price S_k at date k, the price S_{k+1} at date $k+1$ can jump either upward to the value $S_k u$ with a probability $\pi \in (0,1)$ or downward to the value $S_k d$, where $d < 1 < u$. The probability space is then $\Omega = \{u, d\}^T$ and for any $\omega = (\omega_1, \ldots, \omega_T) \in \Omega$, the price process is defined by $S_k(\omega) = S_0 \prod_{j=1}^k \omega_j$. $\mathcal{F}_k$ is the filtration $\sigma(S_1, \ldots, S_k)$ generated by the random variables $S_1, \ldots, S_k$. It is clear that condition (B′) is satisfied. It is well known that in the unconstrained case, $C = \mathbb{R}$, the market is complete and there is a unique equivalent martingale measure $\hat{P}$ whose probability transition is given by

$$\hat{P}\left[S_{k+1} = S_k u | \mathcal{F}_k\right] = \frac{1-d}{u-d}, \hat{P}\left[S_{k+1} = S_k d | \mathcal{F}_k\right] = \frac{u-1}{u-d}.$$

The density of $\hat{P}$ with respect to the objective probability P is then written as

$$\hat{Z} = \frac{d\hat{P}}{dP} = \left(\frac{1-d}{\pi(u-d)}\right)^N \left(\frac{u-1}{(1-\pi)(u-d)}\right)^{T-N},$$

where $N(\omega) = \sum_{k=1}^T 1_{\{\omega_k = u\}}$ follows a binomial law $\mathcal{B}(T, \pi)$. A straightforward calculation shows that

$$E\left[\hat{Z}^q\right] = \frac{1}{(u-d)^{qT}} \left(\pi \left(\frac{1-d}{\pi}\right)^q + (1-\pi)\left(\frac{u-1}{1-\pi}\right)^q\right)^T.$$

By Theorem 5.1 and Remark 5.2, the solution to problem $(\mathcal{P}_p(x))$, for $x < E^{\hat{P}}[H]$, is $\theta^H - \tilde{\theta}(x)$, where θ^H is the perfect replicating strategy for H and $\tilde{\theta}(x)$ is the perfect replicating strategy for $(y^*(x)\hat{Z})^{q-1}$.

5.2. The case of attainable contingent claims. In this paragraph, we assume that C is a symmetrical cone and we consider the $L^p(P)$-hedging of an attainable contingent claim $H \in L^p(P)$. There exist $c \in \mathbb{R}$ and $\theta \in \Theta_p(C)$ such that $H = V_T^{c,\theta}$. Since C is a symmetrical cone and $V_T^{-c,-\theta} = -H$, we have $-\theta \in \Theta_p(C)$. By the supermartingale property of $V^{c,\theta}$ and $V^{-c,-\theta}$ under any martingale measure in $\mathcal{M}_q$, we deduce that $E[ZH] = E[ZV_T^{c,\theta}] = c \ \forall Z \in \mathcal{M}_q$. Therefore the dual problem $(\mathcal{D}_q(y))$ can be written equivalently as

$$\inf_{Z \in \mathcal{M}_q} E[Z^q].$$

Hence, the solution Z_q^* of the dual problem does not depend on y and H. Notice that in the unconstrained case and for $q = 2$, Z_2^* is the variance-optimal martingale measure introduced by Schweizer [39]. We refer to this paper and to Laurent and Pham [28] for explicit computations of the variance-optimal martingale measure in different models of an incomplete market. By analogy, we call Z_q^* the $L^q(P)$-optimal martingale measure.

For $x < \sup_{Z \in \mathcal{M}_q} E[ZH]$, the Lagrange multiplier in (4.11) is explicitly given by

$$y^*(x) = \left(\frac{E[Z_q^* H] - x}{E[(Z_q^*)^q]} \right)^{\frac{1}{q-1}},$$

and the solution to problem $(\mathcal{P}_p(x))$ is the superhedging strategy for the contingent claim

$$X^*(x) = H - \left(y^*(x) Z_q^*\right)^{q-1}.$$

Remark 5.3. In the case of attainable claims, the solution of the primal optimization problem $(\mathcal{S}_p(x))$ can be directly solved without using the convex duality approach (see Appendix B).

5.3. $L^p(P)$-hedging of a riskless asset. In this paragraph, we consider the $L^p(P)$-hedging of the riskless asset $H = 1$. As in the previous paragraph, since $E[ZH] = E[Z] = 1 \ \ \forall \ \ Z \in \mathcal{M}_q$, the dual problem $(\mathcal{D}_q(y))$ can be written equivalently as

$$\inf_{Z \in \mathcal{M}_q} E[Z^q],$$

and the solution Z_q^* of this problem is still called $L^q(P)$-optimal martingale measure. The solution to problem $(\mathcal{P}_p(x))$, for $x < \sup_{Z \in \mathcal{M}_q} E[ZH] = 1$, is given by the superhedging strategy for the contingent claim:

$$X^*(x) = 1 - \frac{1 - x}{E[(Z_q^*)^q]} \left(Z_q^*\right)^{q-1}.$$

Example. We consider the case of no short-selling constraints, $C = [0, \infty)$, in a one-period binomial model. Given the initial value $S_0 = 1$, S_1 takes the values u and d, $d < u$, with objective probability π and $1 - \pi$, $\pi \in (0, 1)$. The probability space is $\Omega = \{u, d\}$ and $\mathcal{F}_1 = \sigma(S_1)$. In this context, problem $(\mathcal{P}_p(x))$ is written as

$$(\mathcal{P}_p(x)) \inf_{\theta \geq 0} E\left[l_p\left((1 - x - \theta \Delta S_1)^+ \right) \right].$$

Obviously, for $x \geq 1$, the solution of this problem is $\theta^*(x) = 0$. Fix now $x < 1$. A probability measure Q on $(\Omega, \mathcal{F}_1)$ is characterized by $\rho = Q(u) \in [0,1]$ and $Q \in \mathcal{M}_q$ if and only if $E^Q[S_1] \leq S_0$, i.e., $\rho \leq \hat{\pi} := (1-d)/(u-d)$. Assuming that $d < 1$, it follows that $\mathcal{M}_q^e \neq \emptyset$. The dual problem is written as

$$\min_{0 \leq \rho \leq \hat{\pi}} \left[\pi \left(\frac{\rho}{\pi} \right)^q + (1-\pi) \left(\frac{1-\rho}{1-\pi} \right)^q \right],$$

whose solution is given by

$$\rho^* = \begin{cases} \pi & \text{if } \pi \leq \hat{\pi}, \\ \hat{\pi} & \text{if } \pi > \hat{\pi}. \end{cases}$$

We shall therefore distinguish two cases:

Case 1. $\pi \leq \hat{\pi}$. The $L^q(P)$-optimal martingale measure is $Q^* = P$ and $Z^* = dQ^*/dP = 1$. Therefore $X^*(x) = x$ and so the solution to problem $(\mathcal{P}_p(x))$ is $\theta^*(x) = 0$.

Case 2. $\pi > \hat{\pi}$. This case implies that, in particular, we must have $u > 1$. The $L^q(P)$-optimal martingale measure is $Q^* = \hat{P}$, where $\hat{P}$ is the unique martingale measure in the unconstrained case and is characterized by $\hat{P}(u) = \hat{\pi}$. We have $Z^*(u) = \hat{\pi}/\pi$, $Z^*(d) = (1-\hat{\pi})/(1-\pi)$. The solution $\theta^*(x)$ to problem $(\mathcal{P}_p(x))$ is the superhedging strategy of $X^*(x) = 1 - (1-x)(Z^*)^{q-1}/E[(Z^*)^q]$, i.e., $x + \theta^*(x)\Delta S_1 \geq X^*(x)$. A straightforward calculation shows that $\theta^*(x)$ is in fact the perfect replicating strategy for $X^*(x)$ and is explicitly given by

$$\theta^*(x) = \frac{(\pi(u-1))^{q-1} - ((1-\pi)(1-d))^{q-1}}{\pi^{q-1}(u-1)^q - (1-\pi)^{q-1}(1-d)^q}(1-x).$$

Appendix A. Proof of Theorem 3.1. Let $x \in \mathbb{R}$ and $\theta \in \Theta_p(C)$ such that $V_T^{x,\theta} \geq H$, P a.s. Then, by the supermartingale property of $V^{x,\theta}$ under any $Q \in \mathcal{M}_q$, we have $\sup_{Q \in \mathcal{M}_q} E^Q[H] \leq x$ and therefore

$$\sup_{Q \in \mathcal{M}_q} E^Q[H] \leq v_0. \tag{A.1}$$

Consider now an arbitrary $x \in \mathbb{R}$ such that $x < v_0$. Then by definition of v_0, the element $H - x$ does not belong to K_p, where $K_p := (K - L^0_+(P)) \cap L^p(P)$. Now, by condition (B), the set $K - L^0_+(P)$ is closed for the topology of convergence in probability (see Brannath [2]) and so K_p is a closed convex cone in $L^p(P)$ containing 0. Therefore, by the Hahn–Banach separation theorem, there exists $Z \in L^q(P) \setminus \{0\}$ such that $\forall \theta \in \Theta(C)$, $Y \in L^0_+(P)$ with $V_T^{0,\theta} - Y \in L^p(P)$,

$$E[Z(V_T^{0,\theta} - Y)] \leq 0 < E[Z(H-x)]. \tag{A.2}$$

As usual, we can assume, without loss of generality, that $\Delta S_k \in L^1(P)$ $\forall k = 1, \ldots, T$. Indeed, if this is not the case, we can change P to an equivalent probability measure $\bar{P}$ with bounded density such that $\Delta S_k \in L^1(\bar{P})$ $\forall$ k (take, for example, $d\bar{P}/dP = \exp(-\max_k |\Delta S_k|)/E[\exp(-\max_k |\Delta S_k|)]$). Relation (A.2) for $\theta = 0$ and $Y = 1_A$, $A \in \mathcal{F}$, implies that $E[Z1_A] \geq 0$; hence $Z \geq 0$, P a.s. Let a be an arbitrary element of C, $A \in \mathcal{F}_{k-1}$ and $\theta \in \Theta(C)$ defined by $\theta_j = 0$ for $j \neq k$ and $\theta_k = a1_A$. Then $V_T^{0,\theta} = a.\Delta S_k 1_A \in L^p(P)$. Then, (A.2) and arbitrariness of $A \in \mathcal{F}_{k-1}$ imply that

$a.E[Z\Delta S_k|\mathcal{F}_{k-1}] \leq 0\ \forall a \in C$ and $k = 1, \ldots, T$. Defining the probability measure Q by $dQ/dP = Z/E[Z]$, it follows that $Q \in \mathcal{M}_q$ and by the right-hand side of (A.2), we have

$$x < E^Q[H]. \tag{A.3}$$

Now, from condition (NA), $\mathcal{M}^e_\infty \neq \emptyset$. Choose some $\tilde{Q} \in \mathcal{M}^e_\infty$ and set $Q_\varepsilon = (1-\varepsilon)Q + \varepsilon\tilde{Q}$, $\varepsilon \in (0,1)$. It is clear that $Q_\varepsilon \in \mathcal{M}^e_q$. Moreover, we have $E^{Q_\varepsilon}[H] \to E^Q[H]$ as ε goes to zero. Then, from (A.3), for a sufficiently small ε, we have $x < E^{Q_\varepsilon}[H]$. From the arbitrariness of $x < v_0$, this proves that

$$v_0 \leq \sup_{Q \in \mathcal{M}^e_q} E^Q[H]. \tag{A.4}$$

This last inequality combined with (A.1) proves (3.1), and also that the supremum can be taken over absolutely continuous martingale measures.

By noting that v_0 can also be written as

$$v_0 = \inf\left\{x \in \mathbb{R} :\ \exists U \in K - L^0_+(P),\ x + U = H,\ P \text{ a.s.}\right\}, \tag{A.5}$$

and since $(K - L^0_+(P)) \cap L^p(P)$ is closed in $L^p(P)$, we conclude that the infimum in (A.5) is attained whenever v_0 is finite, which proves assertion (1). Finally suppose that the supremum in (3.1) is attained for some $\hat{Q} \in \mathcal{M}^e_q$. Then v_0 is finite and there exists a superhedging strategy θ^H for H: $V_T^{v_0,\theta^H} - H \geq 0$ P a.s. By the supermartingale property of V^{v_0,θ^H} under $\hat{Q}$, we have $E^{\hat{Q}}[V_T^{v_0,\theta^H} - H] \leq v_0 - E^{\hat{Q}}[H] = 0$. Since $\hat{Q}$ is equivalent to P, we then conclude that $V_T^{v_0,\theta^H} = H$ P a.s.

Appendix B. Direct resolution in the case of attainable claims. In this paragraph, we directly derive the solution to the $L^p(P)$-hedging of an attainable contingent claim $H \in L^p(P)$, which then satisfies $E[ZH] = c\ \forall Z \in \mathcal{M}_q$. For $x \geq c = \sup_{Z \in \mathcal{M}_q} E[ZH]$, we already know that $X^*(x) = H$ is a solution to the static problem $(\mathcal{S}_p(x))$ associated to the $L^p(P)$-hedging problem $(\mathcal{P}_p(x))$. Fix now $x < c$, and consider the L^q-optimal martingale measure Z^*_q solution of

$$\inf_{Z \in \mathcal{M}_q} E[Z^q].$$

One easily obtains

$$E[(Z^*_q)^q] \leq E[Z(Z^*_q)^{q-1}] \quad \forall Z \in \mathcal{M}_q. \tag{B.1}$$

Now, let $X \in \mathcal{C}_p(x)$. By the Hölder inequality, we have $E[Z^*_q(H - X)] \leq (E[Z^*_q]^q)^{\frac{1}{q}}$ $(E[H - X]^p)^{\frac{1}{p}}$ and so

$$\frac{(c-x)^p}{\left(E[Z^*_q]^q\right)^{\frac{p}{q}}} \leq E[H - X]^p. \tag{B.2}$$

Now, set $y^*(x) = (\frac{c-x}{E[Z^*_q]^q})^{\frac{1}{q-1}}$, and define $X^*(x) = H - \left(y^*(x)Z^*_q\right)^{q-1}$. Then $X^*(x)$ $\leq H$ and $E[ZX^*(x)] = c - (y^*(x))^{q-1}\ E[Z(Z^*_q)^{q-1}] \leq c - (y^*(x))^{q-1}\ E[Z^*_q]^q = x$, by using (B.1) and definition of $y^*(x)$. Hence $X^*(x) \in \mathcal{C}_p(x)$. Moreover, we have

$$\begin{aligned} E[H - X^*(x)]^p \ &= \ (y^*(x))^q E[Z^*_q]^q = \frac{(c-x)^p}{\left(E[Z^*_q]^q\right)^{\frac{p}{q}}} \\ &\leq E[H - X]^p, \end{aligned}$$

by (B.2). This proves that $X^*(x)$ solves $(\mathcal{S}_p(x))$.

Acknowledgments. I would like to thank the anonymous referees for their helpful comments, and for pointing out the argument shown in Remark 5.3.

REFERENCES

[1] P. Artzner, F. Delbaen, J.M. Eber, and D. Heath, *A Characterization of Measures of Risk*, Math. Finance, to appear.
[2] W. Brannath, *No-Arbitrage and Martingale Measures in Option Pricing*, Doctoral Dissertation, Universität Wien, Vienna, Austria, 1997.
[3] J. Cvitanić, *Minimizing expected loss of hedging in incomplete and constrained markets*, SIAM J. Control Optim., to appear.
[4] J. Cvitanić and I. Karatzas, *Convex duality in convex portfolio optimization*, Ann. Appl. Probab., 2 (1992), pp. 767–818.
[5] J. Cvitanić and I. Karatzas, *Hedging contingent claims with constrained portfolios*, Ann. Appl. Prob., 3 (1993), pp. 652–681.
[6] J. Cvitanić and I. Karatzas, *On dynamic measures of risk*, Fin. and Stoch., to appear.
[7] J. Cvitanić, H. Pham, and N. Touzi, *Super-replication in stochastic volatility models under portfolios constraints*, J. Appl. Probab., 36 (1999), pp. 523–545.
[8] R. Dalang, A. Morton, and W. Willinger, *Equivalent martingale measures and no-arbitrage in stochastic securities market models*, Stochastics Stochastic Rep., 29 (1990), pp. 185–201.
[9] F. Delbaen and W. Schachermayer, *A general version of the fundamental theorem of asset pricing*, Math. Ann., 300 (1994), pp. 463–520.
[10] D. Duffie and H.R. Richardson, *Mean-variance hedging in continuous time*, Ann. Appl. Probab., 1 (1991), pp. 1–15.
[11] E. Eberlein and J. Jacod, *On the range of option prices*, Fin. and Stoch., 1 (1997), pp. 131–140.
[12] I. Ekeland and R. Temam, *Analyse Convexe et Problèmes Variationnels*, Dunod, Paris, 1974.
[13] N. El Karoui and M.-C. Quenez, *Dynamic programming and pricing of contingent claims in an incomplete market*, SIAM J. Control Optim., 33 (1995), pp. 29–66.
[14] H. Föllmer and Y. Kabanov, *Optional decomposition and Lagrange multipliers*, Fin. and Stoch., 2 (1998), pp. 69–81.
[15] H. Föllmer and D. Kramkov, *Optional decomposition under constraints*, Probab. Theory Related Fields, 109 (1997), pp. 1–25.
[16] H. Föllmer and P. Leukert, *Quantile hedging*, Fin. and Stoch., to appear.
[17] H. Föllmer and P. Leukert, *Efficient hedging: Cost versus and shortfall risk*, Fin. and Stoch., to appear.
[18] C. Gouriéroux, J.P. Laurent, and H. Pham, *Mean-variance hedging and numéraire*, Math. Finance, 8 (1998), pp. 179-200.
[19] J.M. Harrison and D. Kreps, *Martingale and arbitrage in multiperiods securities markets*, J. Econom. Theory, 20 (1979), pp. 381–408.
[20] J.M. Harrison and S.R. Pliska, *Martingales and stochastic integrals in the theory of continuous trading*, Stochastic Process. Appl., 11 (1981), pp. 215–260.
[21] H. He and N. Pearson, *Consumption and portfolio policies with incomplete markets and short-selling constraints: The infinite-dimensional case*, J. Econom. Theory, 54 (1991), pp. 259–304.
[22] S. Jacka, *A martingale representation result and an application to incomplete financial markets*, Math. Finance, 2 (1992), pp. 239–250.
[23] E. Jouini and H. Kallal, *Arbitrage in securities markets with short-sales constraints*, Math. Finance, 5 (1995), pp. 197–232.
[24] Y. Kabanov and D. Kramkov, *Nonarbitrage and equivalent martingale measures: A new proof of the Harrison–Pliska theorem*, Theory Probab. Appl., 39 (1994), pp. 523–527.
[25] I. Karatzas, J.P. Lehoczky, S.E. Shreve, and G.-L. Xu, *Martingale and duality methods for utility maximization in an incomplete market*, SIAM J. Control Optim., 29 (1991), pp. 702–730.
[26] D. Kramkov, *Optional decomposition of supermartingales and hedging contingent claims in incomplete security markets*, Probab. Theory Related Fields, 105 (1996), pp. 459–479.
[27] D. Kramkov and W. Schachermayer, *The asymptotic elasticity of utility functions and optimal investment in incomplete markets*, Ann. Appl. Probab., 9 (2000), pp. 904–950.
[28] J.P. Laurent and H. Pham, *Dynamic programming and mean-variance hedging*, Fin. and

Stoch., 23 (1999), pp. 83–110.
[29] P. Monat and C. Stricker, *Föllmer-Schweizer decomposition and mean-variance hedging of general claims*, Ann. Probab., 23 (1995), pp. 605–628.
[30] H. Pham and N. Touzi, *The fundamental theorem of asset pricing under cone constraints*, J. Math. Econom., 31 (1999), pp. 265–280.
[31] T. Rheinländer and M. Schweizer, *On L^2-projections on a space of stochastic integrals*, Ann. Probab., 25 (1997), pp. 1810–1831.
[32] R.T. Rockafellar, *Convex Analysis*, Princeton University Press, Princeton, NJ, 1970.
[33] L.C.G. Rogers, *Equivalent martingale measures and no-arbitrage*, Stochastics Stochastic Rep., 51 (1994), pp. 41–49.
[34] W. Schachermayer, *A Hilbert space of the fundamental theorem of asset pricing in finite discrete time*, Insurance Math. Econom., 11 (1992), pp. 249–257.
[35] M. Schäl, *On quadratic cost criteria for option hedging*, Math. Oper. Res., 19 (1994), pp. 121–131.
[36] M. Schäl, *Martingale measures and hedging for discrete time financial markets*, Math. Oper. Res., 24 (1999), pp. 509–528.
[37] K. Schürger, *On the existence of equivalent τ-measures in finite discrete time*, Stochastic Process. Appl., 61 (1996), pp. 109–128.
[38] M. Schweizer, *Variance-optimal hedging in discrete time*, Math. Oper. Res., 20 (1995), pp. 1–32.
[39] M. Schweizer, *Approximation pricing and the variance-optimal martingale measure*, Ann. Probab., 64 (1996), pp. 206–236.

SIAM J. CONTROL OPTIM.
Vol. 38, No. 3, pp. 683–710

A MAX-PLUS-BASED ALGORITHM FOR A HAMILTON–JACOBI–BELLMAN EQUATION OF NONLINEAR FILTERING*

WENDELL H. FLEMING† AND WILLIAM M. McENEANEY‡

Abstract. The Hamilton–Jacobi–Bellman (HJB) equation associated with the robust/H_∞ filter (as well as the Mortensen filter) is considered. These filters employ a model where the disturbances have finite power. The HJB equation for the filter information state is a first-order equation with a term that is quadratic in the gradient. Yet the solution operator is linear in the max-plus algebra. This property is exploited by the development of a numerical algorithm where the effect of the solution operator on a set of basis functions is computed off-line. The precomputed solutions are stored as vectors of coefficients of the basis functions. These coefficients are then used directly in the real-time computations.

Key words. nonlinear filtering, robust, H_∞, max-plus algebra, nonlinear HJB equations

AMS subject classifications. 35F25, 49L25, 93B36, 93B99, 93C10, 93C90, 93E11

PII. S0363012998332433

1. Introduction.

This paper is concerned with an algorithm for construction of viscosity solutions of a certain Hamilton–Jacobi–Bellman (HJB) equation. The solution of this partial differential equation (PDE) is a necessary step in the application of certain filtering techniques (such as robust filtering) to nonlinear systems. This will be discussed further below. First, we note that the PDE is first-order but nonlinear—possessing a term which is quadratic in the gradient. However, in the max-plus algebra, the solution operator is linear. The algorithm takes advantage of this linearity by using a max-plus basis representation of the solution. Propagation of the solution is reduced to a max-plus matrix multiplication on the coefficients of the basis representation.

We consider the problem of estimating the state of some system evolving continuously in time. The state space is $\Re^n$. The problem of estimating the current state based on knowledge (or lack thereof) of the initial state and the measurements up to the current moment is the filtering problem. One approach to the problem is to model the system via stochastic differential equations. This leads to the stochastic nonlinear filtering problem (we assume that the extended linear Kalman filter is not appropriate). The key step becomes the solution of a second-order stochastic partial differential–integral equation—the Kushner equation [19]—or a linear, second-order stochastic partial differential equation (SPDE)—the Zakai equation [6]. Alternatively, this can be reformulated via pathwise filtering, leading to a deterministic PDE parameterized by the measurement process [6], [11], [28].

One particular approach to the solution of the Zakai SPDE is the splitting method (see, for instance, [21]). If the measurements occur at discrete times, this approach

*Received by the editors January 12, 1998; accepted for publication (in revised form) February 15, 1999; published electronically February 24, 2000.
http://www.siam.org/journals/sicon/38-3/33243.html

†Division of Applied Mathematics, Brown University, Providence, RI 02912 (whf@cfm.brown.edu). The research of this author was partially supported by ONR grant N0014-96-1-0267, NSF grant DMS-9531276, and ARO grant DAAL03-92-G0115

‡Department of Mathematics, North Carolina State University, Raleigh, NC 27695–8205 (wmm@math.ncsu.edu). The research of this author was partially supported by ONR grant N0014-96-1-0267 and AFOSR grant F49620-95-1-0296.

takes advantage of the linearity of the solution operator via a basis function representation of the solution of the SPDE. In the continuous-time measurements case, one takes a limit [21], [16].

An alternate approach is taken by robust/H_∞ filters [24], [9], [5] and the Mortensen filter [29]. In this case, the noise in the system dynamics is modeled by an L^2 process that is unknown a priori, but to which we do not associate a probability measure. The measurement process is also affected by an a priori unknown disturbance. In contrast to the Kalman filter approach, for nonlinear systems this problem necessitates solution of a *first-order* PDE. (Note that in the formulations considered here this is a finite-dimensional PDE.) This first-order nature has certain implications. The foremost is that one has a "range-of-dependence" property: a change in initial conditions at one point does not instantaneously affect the solution everywhere in $\Re^n$; the solution propagates with finite speed. This has certain numerical advantages, the most obvious of which is the applicability of generalized characteristic methods [24], [27]. Second, the solution, even in nondegenerate cases, is likely to be nonsmooth, leading to the use of viscosity solution definitions of solution. However, this first-order PDE is not linear; it possesses a quadratic term in the gradient.

In the max-plus algebra (which may more correctly be termed a commutative semifield), the "addition operation" is defined by $a \oplus b = \max\{a, b\}$, and the "multiplication operation" is defined by $a \otimes b = a + b$. Interestingly, the solution operator for the PDE associated with the robust/H_∞ filter is linear in the max-plus algebra. As noted above, the algorithm described herein takes advantage of this linearity and a max-plus basis representation of the solution. This approach is analogous to the splitting method described above for the Kalman filter. In fact, that method provided, by analogy, an early influence on this work.

Both the case of discrete-time measurements and the case of continuous-time measurements are considered. In the discrete-time measurements case, the solution of the PDE between measurement times (2.14b) is of the form discussed just above, and the above approach is used, leading to the algorithm discussed in section 6. In the case of continuous-time measurements, a proof of convergence of this algorithm as the measurement time-step goes to zero is presented in section 7.

Section 2 reviews the robust/H_∞ filter (and Mortensen filter) and provides motivation for the solution of the relevant HJB equations. In section 3, the max-plus algebra is described, and the max-plus linearity of the solution operator is proved. The basis representation will be obtained in the class of continuous, semiconvex functions. Proofs of the semiconvexity of the solution of the PDE (under two sets of assumptions) are given in section 4. A semiconvex duality-based representation of the solution is also included there. In section 5, the semiconvex duality representation is used to obtain the basis representation, and certain properties of this representation are also presented. Sections 6 and 7 are as discussed above.

2. Review of the robust filter. In this section, the robust filter (as well as the Mortensen filter) are briefly reviewed. For a fuller discussion, see [24], [9], [10], [29].

Let $x(t)$ denote the state to be estimated at time $t \geq 0$, with $x(t) \in \Re^n$. The state dynamics are

$$\frac{dx}{dt} = f(x) + \sigma(x)w, \tag{2.1}$$

where $w \in L_2^{loc}([0, \infty); \Re^m)$ is the disturbance in the dynamics and σ is an $n \times m$

matrix-valued function. First we will describe the case of continuous-time measurements. Let the observations, taking values in $\Re^k$, be given by

$$y(t) = h(x(t)) + \rho(x(t))v(t), \tag{2.2}$$

where $v \in L_2^{loc}([0,\infty);\Re^l)$ is the observation disturbance, ρ is a $k \times l$ matrix-valued function, and $k \le l$. Let $\phi(x_0)$ be a measure of our uncertainty about the initial state $x(0) = x_0$.

We will assume that

$$\begin{aligned} &f, \sigma, h, \text{ and } \rho \text{ are } C^2; \\ &f, \sigma, \text{ and } h \text{ are globally Lipschitz in } x. \end{aligned} \tag{A1}$$

The differentiability assumptions are required in certain calculus of variations arguments. We assume that there exists $M < \infty$ such that

$$|\sigma(x)| \le M \qquad \forall\, x \in \Re^n \tag{A2}$$

and define $a \doteq \sigma\sigma^T$. We will also assume that

$$\text{Range}(\rho(x)) = \Re^k \qquad \forall\, x \in \Re^n, \tag{A3a}$$

which guarantees that for any y, x there exists some v satisfying (2.2).

Finally, we define ρ^{-1} by

$$\rho^{-1}(x)b = \text{argmin}\{|v| :\ \rho(x)v = b\}. \tag{2.3}$$

Assume that ρ^{-1} is uniformly bounded, that is, that there exists $C_\rho < \infty$ such that

$$\begin{aligned} &|\rho^{-1}(x)b| \le C_\rho |b| \qquad \forall\, x \in \Re^n \ \ \forall\, b \in \Re^l, \\ &\rho^{-1} \in C^2. \end{aligned} \tag{A3b}$$

Note also that these assumptions imply that if we view the integral version of (2.1),

$$x_T \doteq x(T) = x_0 + \int_0^T f(x(t)) + \sigma(x(t))w(t)\, dt, \tag{2.4}$$

as a mapping from x_0 to x_T, then this mapping is one-to-one and onto for any $w \in \mathrm{L}_2$.

Let $\phi(x_0)$ be a measure of our uncertainty about the initial state x_0, and suppose there exists $D_1 < \infty$ such that

$$\begin{aligned} &\phi(x) \le D_1 \qquad \forall\, x \in \Re^n, \\ &\phi \text{ is locally Lipschitz.} \end{aligned} \tag{A4}$$

It should be noted here that the algorithm described in this paper will produce the solution of the HJB equation corresponding to the robust/H_∞ and Mortensen filters. The algorithm could also be used to solve similar HJB equations arising from different applications. If it is not obvious that the desired solution is the same as the solution for the robust filter, then one would want a result indicating uniqueness of the (viscosity) solution of the HJB equation in some class. This uniqueness is known under the additional assumption that ϕ satisfies a quadratic growth assumption (references are given below). Since this uniqueness is not required for the specific result here, we did not include a quadratic growth assumption on ϕ in the above.

Suppose we wish to estimate the state at time T. Consider a cost criterion of the form

$$J(T, x_T, w(\cdot)) = \phi(x_0) - \frac{1}{2}\int_0^T |w(t)|^2 + |v(t)|^2\, dt \tag{2.5a}$$

$$= \phi(x_0) - \frac{1}{2}\int_0^T |w(t)|^2 + |\rho^{-1}(x(t))[y(t) - h(x(t))]|^2\, dt, \tag{2.5b}$$

where x_0 is given by (2.4) for any particular w. The information state is given by

$$P(T, x_T) = \sup_{w \in L_2} J(T, x_T, w). \tag{2.6}$$

We now briefly discuss the extraction of filter estimates from the information state. After this, we will return to a detailed discussion of aspects of (and representations for) the above information state. Further details may be found in the references included below. Readers who are interested mainly in the computational algorithm may choose to skip this discussion.

Mortensen [29] developed an estimator based on this information state in the 1960s. Mortensen's estimate is $\widehat{x}_T \in \operatorname{argmax}_x P(T, x)$ (the existence of which follows easily under reasonable assumptions [24], [9]).

For a robust filter, one is interested in a state estimate for which one has a bound on the effect of the disturbances on our estimate error by the product of some constant and the disturbance energy (where one hopes that this constant is small). In particular, one would like an estimate, $\hat{e}_T$, such that for some $\gamma^2 < \infty$ [24],

$$|x(T) - \hat{e}_T|^2 \le \gamma^2 \left[-\phi(x_0) + \frac{1}{2}\int_0^T \left[|w_t|^2 + |v_t|^2 \right] dt \right] \tag{2.7}$$

for all x_0, w, v in the case of continuous-time measurements with an analogous inequality for discrete-time measurements. (Of course this would require an additional assumption that $\phi(x) \le -k|x - \overline{x}|^2$ for some $k > 0$ and $\overline{x} \in \Re^n$; see [24].) Under reasonable assumptions (see [24]), there exists $\gamma^* < \infty$ such that for all $\gamma \ge \gamma^*$, such a robust estimator is given by

$$\hat{e}_T \doteq \operatorname{argmin}_e \max_x \left[|x - e|^2 + \gamma^2 P(T, x) \right]. \tag{2.8}$$

It should perhaps be noted that there will generally be multiple estimates that yield disturbance attenuation (2.7), and in fact, under certain conditions, the Mortensen estimator is a robust estimator in this sense [9]. We also note that $\hat{e}_T$ (given by (2.8)) is the risk-averse limit of a risk-sensitive stochastic filter [10]. Some readers may note that the above information state differs from some which include an integral running cost term. Such an information state is particularly suited to the problem of H_∞ control under partial information as opposed to the problem of robust state estimation considered here; see [1], [17].

Now we turn to the promised further discussion of the information state, P, itself. It will be convenient to note the following.

LEMMA 2.1. *Let $T, R < \infty$, $|x_T| \le R$. There exists M_1 (depending on T, R) such that for $\epsilon \le 1$, ϵ-optimal w for problem (2.6) satisfies $\|w\|^2_{L_2[0,T]} \le M_1$. Further, there exists an optimal w^* and M_2 (depending on T, R) such that $|w^*(t)| \le M_2$ for all $t \in [0, T]$.*

Proof. The first assertion follows by the standard technique of comparison with $w^0 \equiv 0$. Let $x^0(\cdot)$ be the solution of (2.1) corresponding to w^0 and $x^0(T) = x_T$. By (A1), there exist $K, K_h < \infty$ such that $|f(x)| \le K(1+|x|)$ and $|h(x)| \le K_h(1+|x|)$. Then $\frac{d}{dt}|x^0|^2 \le K(3|x^0|^2+1)$, and by Gronwall's inequality, one obtains $|x^0(t)|^2 \le R_1^2 \doteq (R^2+KT)(1+3KTe^{3KT})$ for all $t \in [0,T]$.

Let $\mu(\rho) = \min\{\phi(x) : |x| \le \rho\}$ and note that $\mu(\rho) > -\infty$ for all $\rho < \infty$ by (A4). Then

$$J(T, x_T, w^0(\cdot)) \ge \mu(R_1) - \frac{C_\rho^2}{2}\left[2K_h^2(1+R_1^2) + \|y\|^2_{L_2[0,T]}\right] \doteq \overline{M}_1.$$

Consequently, for ϵ-optimal w^ϵ, $-\frac{1}{2}\int_0^T |w^\epsilon|^2\,dt \ge \overline{M}_1 - \epsilon$, which yields the first assertion. The second assertion then follows by the proof of [9, Lemma 2.1(a)]. □

The following theorem and remark are easily obtained in the case where $y(\cdot)$ is continuous, and we outline the proof in that case. They are not directly needed in the sequel, and so we do not prove them in the case where $y \in L_2 \setminus C$. (Of course, they could be needed in an adaptation of this algorithm to the solution of PDEs not necessarily arising from this application.) Extensions of viscosity solutions to cases where the time-dependency of the Hamiltonian is only measurable are well known but more technical (see, for instance, [4], [14], [22], [32]).

THEOREM 2.2. *P is a continuous viscosity solution of the HJB equation*

$$\begin{aligned} 0 &= P_T + f^T(x)\nabla_x P - \frac{1}{2}\nabla_x P^T a(x)\nabla_x P + \frac{1}{2}|\rho^{-1}(x)(y(T)-h(x))|^2, \\ P(0,x) &= \phi(x). \end{aligned} \tag{2.9}$$

Proof. Fix $R < \infty$ and let $|x_T| \le R$. By Lemma 2.1, the optimal trajectory $x^*(\cdot)$ terminating at x_T satisfies $|x^*(t)| \le R_1$ for some $R_1 < \infty$ (depending on T, R) for any $x_t \in B_R \doteq \{x \in \Re^n : |x| \le R\}$. Let

$$\widetilde{\phi}(x) = \begin{cases} \phi(x) & \text{if } |x| \le R_1, \\ \phi(R_1 x/|x|) & \text{if } |x| > R_1. \end{cases}$$

Let $\widetilde{P}$ be the value given by (2.6) with $\widetilde{\phi}$ replacing ϕ. Then, by standard results (for instance, [26]), $\widetilde{P}$ is a viscosity solution of (2.9) with $\widetilde{\phi}$ replacing ϕ. Since $\widetilde{P} = P$ on $[0,T] \times B_R$ and $\widetilde{\phi} = \phi$ on B_{R_1}, P is a viscosity solution of (2.9) on $[0,T] \times B_R$. Letting $R \to \infty$, one obtains the result. □

Remark 2.3. In the case where $\phi(x) \ge -D_2(1+|x|^2)$ for some $D_2 < \infty$ and σ is constant, uniqueness among the class of continuous, quadratically growing viscosity solutions was obtained in [25]. Recent results of Da Lio and McEneaney [7], Bardi and Da Lio [2], and Ishii [15] allow the removal of the constant σ assumption.

Let us also develop the information state in the discrete-time measurement case.

In the discrete-time measurement case, the dynamics (2.1) are unchanged, but the measurement model is as follows. Suppose that at each measurement time, $t_j > 0$, we receive an observation that is modeled as

$$y_j = h(x(t_j)) + \rho(x(t_j))v_j. \tag{2.10}$$

Let the number of measurements in time $[0,T]$ be N. One may adapt the cost criterion

(2.5) and information state (2.6) in the following manner. Let the cost criterion be

$$J(T, x_T, w(\cdot)) = \phi(x_0) - \frac{1}{2}\int_0^T |w(t)|^2\,dt - \frac{\delta}{2}\sum_{j=1}^N |v_j|^2 \tag{2.11a}$$

$$= \phi(x_0) - \frac{1}{2}\int_0^T |w(t)|^2\,dt - \frac{\delta}{2}\sum_{j=1}^N |\rho^{-1}(x(t_j))[y_j - h(x(t_j))]|^2, \tag{2.11b}$$

where δ is the time interval between measurements, and let the value again be given by

$$P(T, x_T) = \sup_{w\in L_2} J(T, x_T, w). \tag{2.12}$$

Due to the discrete nature of the measurements, it is helpful to recall the form of the dynamic programming principle for such a system. Let $P(t_j^+, x)$ be the value at time t_j just after measurement j, and $P(t_j^-, x)$ be the value at time t_j just before the measurement. If T is not a measurement time, and $t \in (t_{k-1}, t_k)$, we have

$$P(T,x) = \sup_{w\in L_2([t,T];\Re^m)} \left\{ P(t,x(t)) - \frac{1}{2}\int_t^T |w(r)|^2\,dr - \frac{\delta}{2}\sum_{j=k}^N |\rho^{-1}(x(t_j))[y_j - h(x(t_j))]|^2 \right\}, \tag{2.13a}$$

and when T occurs at the time of measurement j, we have

$$P(T^+, x) = P(T^-, x) - \frac{\delta}{2}|\rho^{-1}(x)[y_j - h(x)]|^2. \tag{2.13b}$$

The corresponding dynamic programming equations are

$$P(0,x) = \phi(x), \qquad t = 0, \tag{2.14a}$$

$$0 = P_t - \sup_{w\in\Re^m}\left\{ -[f(x) + \sigma(x)w]^T \nabla_x P - \frac{1}{2}|w|^2 \right\}, \qquad t \in (t_j, t_{j+1}),$$

$$= P_t + f^T(x)\nabla_x P - \frac{1}{2}(\nabla_x P)^T a(x) \nabla_x P, \qquad t \in (t_j, t_{j+1}), \tag{2.14b}$$

$$P(t_j^+, x) = P(t_j^-, x) - \frac{\delta}{2}|\rho^{-1}(x)[y_j - h(x)]|^2, \qquad t = t_j \text{ for some } j. \tag{2.14c}$$

As in the continuous-time measurements case, we have the following [24]. (Since y does not appear on the right-hand side of (2.14b), there is no difficulty with discontinuous y in this case.)

THEOREM 2.4. *The information state, P, given by* (2.13)*, is continuous and is a viscosity solution of* (2.14) *between measurement times.*

In the case where ϕ satisfies a quadratic growth condition, uniqueness follows in a similar manner as that discussed in Remark 2.3.

As in the continuous-time measurement case, one obtains the robust estimator from (2.8), and an attenuation inequality similar to (2.7) follows (with summation of v_j replacing the integral of v_t).

Finally, see [1], [5], and [18] for related work.

3. Max-plus linearity. Recall the definition of max-plus addition, $\oplus$, and multiplication, $\otimes$, for elements of $\Re$ as

$$\begin{aligned} a \oplus b &\doteq \max\{a, b\}, \\ a \otimes b &\doteq a + b. \end{aligned} \tag{3.1}$$

Define these operations for $-\infty$ in the obvious way. It is well known that $(\Re \cup \{-\infty\}, \oplus, \otimes)$ is a commutative semifield which is referred to as the max-plus algebra. See [3], [20], [13] for a fuller discussion. We remark that, although the commutative semifield lacks additive inverses, there is an extension which allows one to solve linear systems in a certain sense. This extension is accomplished via the notion of a "balance"; see [3]. However, for the robust/H_∞ and Mortensen filter applications, we are interested only in real-valued quantities, and so nothing is gained by this extension. Consequently, we will not pursue this here.) The authors have recently learned that Theorem 3.1 was previously reported in 1987 [23].

As noted earlier, we are interested only in the discrete-time measurements case in this section; the continuous-time measurements case will be discussed in section 7. Let $\mathcal{S}_T$ be the solution operator for HJB equation (2.14b). Specifically, given initial condition $P(t_0, \cdot)$, the solution to (2.14b) at time $t_0 + T$ is given by

$$P(t_0 + T, x) = \mathcal{S}_T[P(t_0, \cdot)](x). \tag{3.2}$$

From the dynamic programming principle (2.13a) (noting that there are no measurements in $(t_0, t_0 + T)$ in order for (2.14b) to be applicable), one has

$$\begin{aligned} \mathcal{S}_T[P(t_0, \cdot)](x) &= P(t_0 + T, x) \\ &= \sup_{w \in L_2([t_0, t_0+T]; \Re^m)} \left\{ P(t_0, x(t_0)) - \frac{1}{2} \int_{t_0}^{t_0+T} |w(r)|^2 \, dr \right\}, \end{aligned} \tag{3.3}$$

where $x(t_0)$ is given by (2.1) with $x(t_0 + T) = x$.

We now prove that $\mathcal{S}_T$ is linear in the max-plus algebra. Let $c \in \Re$ and let $\phi, \psi : \Re^n \to \Re$ be locally Lipschitz and bounded from above. By (3.3), one easily sees that for any $x \in \Re^n$

$$\mathcal{S}_T[c \otimes \phi](x) = \mathcal{S}_T[c + \phi](x) = c + \mathcal{S}_T[\phi] = c \otimes \mathcal{S}_T[\phi](x). \tag{3.4}$$

Also, for any $x \in \Re^n$,

$$\begin{aligned} \mathcal{S}_T[\phi \oplus \psi](x) &= \sup_{w \in L_2([0,T]; \Re^m)} \left\{ \max[\phi(x(0)), \psi(x(0))] - \frac{1}{2} \int_0^T |w(r)|^2 \, dr \right\} \\ &\geq \sup_{w \in L_2([0,T]; \Re^m)} \left\{ \phi(x(0)) - \frac{1}{2} \int_0^T |w(r)|^2 \, dr \right\} \\ &= \mathcal{S}_T[\phi](x). \end{aligned} \tag{3.5}$$

By symmetry, one obtains an analogous inequality for ψ. Consequently,

$$\mathcal{S}_T[\phi \oplus \psi](x) \geq \max\{\mathcal{S}_T[\phi](x), \mathcal{S}_T[\psi](x)\} = \mathcal{S}_T[\phi](x) \oplus \mathcal{S}_T[\psi](x). \tag{3.6}$$

Now let w^* be optimal in (3.5). (The existence of the optimizer is discussed in section 2.) Then

$$\mathcal{S}_T[\phi \oplus \psi](x) = \max[\phi(x(0)), \psi(x(0))] - \frac{1}{2}\int_0^T |w^*(r)|^2\, dr,$$

and, in the case where $\phi(x(0)) \geq \psi(x(0))$,

$$\begin{aligned} &= \phi(x(0)) - \frac{1}{2}\int_0^T |w^*(r)|^2\, dr \\ &\leq \mathcal{S}_T[\phi](x). \end{aligned}$$

Considering also the case where $\phi(x(0)) < \psi(x(0))$, one obtains

$$\mathcal{S}_T[\phi \oplus \psi](x) \leq \mathcal{S}_T[\phi](x) \oplus \mathcal{S}_T[\psi](x).$$

Combining this with (3.6) yields

$$\mathcal{S}_T[\phi \oplus \psi](x) = \mathcal{S}_T[\phi](x) \oplus \mathcal{S}_T[\psi](x). \tag{3.7}$$

By (3.4) and (3.7), one has the following theorem.

THEOREM 3.1. *The solution operator for* (2.14b) *is linear in the max-plus algebra.*

4. Semiconvexity. A function ψ is called *semiconvex* if for every $R < \infty$, there exists C_R such that

$$\widehat{\psi}(x) \doteq \psi(x) + \frac{C_R}{2}|x|^2 \tag{4.1}$$

is convex on the ball $B_R \doteq \{x \in \Re^n : |x| \leq R\}$. The infimum over such C_R will be known as the semiconvexity constant for ψ over B_R. Note that any real-valued semiconvex function is locally Lipschitz. In fact, if $\widehat{\psi}$ is convex on the ball B_R, then $\widehat{\psi}$ is Lipschitz on $B_{R/2}$ with a Lipschitz constant depending only on $\max_{|x|\leq R} \widehat{\psi}(x) - \min_{|x|\leq R} \widehat{\psi}(x)$ (see [8, p. 111]).

In this section, it will be shown that if either the initial information state is semiconvex, or if σ is nondegenerate, then the information state will be semiconvex for all $T > 0$. Next, the convex duality representation for convex functions is adapted to the semiconvex case. Then, since the information state is semiconvex, one sees that this provides a semiconvex duality representation for the information state. This representation will be used in section 5 to obtain the basis representation.

Recall from section 3 that the solution operator (in the absence of observations) is defined as

$$\begin{aligned} \mathcal{S}_T[\phi](x) &= \sup_{w \in L_2([0,T];\Re^m)} J^0(T, x; w(\cdot)), \\ J^0(T, x; w(\cdot)) &= \phi(x_0) - \frac{1}{2}\int_0^T |w(t)|^2\, dt, \end{aligned} \tag{4.2}$$

where $x_0 = x(0)$ given $x(T) = x$ and dynamics (2.1).

THEOREM 4.1. *If ϕ is semiconvex, then $\mathcal{S}_T[\phi]$ is semiconvex for each $T > 0$. In fact, there exists $\Gamma_R(T)$ such that for $0 \leq \tau \leq T$, $R > 0$, $\mathcal{S}_\tau[\phi](x) + \Gamma_R(T)|x|^2$ is convex on the ball B_R.*

Proof. Let

$$\widetilde{J}(\tau, x; w(\cdot)) = J^0(\tau, x; w(\cdot)) + \Gamma|x|^2,$$
$$J^0(\tau, x, w(\cdot)) = \phi(x_0) - \frac{1}{2}\int_0^\tau |w(t)|^2\, dt,$$

where $\Gamma = \Gamma_R(T)$ is to be chosen suitably. Since the supremum of any family of convex functions is a convex function, it suffices to show that $\widetilde{J}(\tau, \cdot; w(\cdot))$ is convex on B_R for each $w(\cdot)$.

By smoothing via convolution with approximations to the identity, we may assume that ϕ is smooth. Then semiconvexity of ϕ is equivalent to $v^T\phi_{xx}(x)v \geq -C_R|v|^2$ for all $v \in \Re^n$, when $|x| \leq R$. To show convexity of $\widetilde{J}$ it suffices to show that

$$v^T J^0_{xx}(\tau, x; w(\cdot))v \geq -2\Gamma \tag{4.3}$$

whenever $|x| \leq R$, $|v| = 1$, and $\|w\|^2 \leq M_1(T, R)$ (see Lemma 2.1). Let $\zeta^1(t), \zeta^2(t)$ denote the first and second derivatives of $x(t)$ in the direction v (considered as a function of $x = x(\tau)$). Then

$$v^T J^0_{xx}(\tau, x; w(\cdot))v = \phi_x(x_0) \cdot \zeta^2(0) + (\zeta^1(0))^T\phi_{xx}(x_0)\zeta^1(0).$$

By the "range-of-dependence" property (which follows directly from Lemma 2.1), $|x_0| \leq R_1$ for suitable R_1. If we can also obtain bounds on $|\zeta^1(0)|, |\zeta^2(0)|$ depending only on T, R, then we will have (4.3) for suitable $\Gamma = \Gamma_R(T)$, which will complete the proof. (In fact, one could explicitly compute $\Gamma_R(T)$ from these bounds if so desired.) Bounds on $\zeta^1(0), \zeta^2(0)$ are obtained as follows. Let

$$\dot{x}^h = f(x^h) + \sigma(x^h)w,$$
$$x^h(\tau) = x + hv.$$

Then with $\zeta^1(t) \doteq \frac{d}{dh}x^h(t)\big|_{h=0}$, one has

$$\dot{\zeta}^1 = f_x(x^0)\zeta^1 + [\sigma_x(x^0)\zeta^1]w,$$
$$\zeta^1(\tau) = v.$$

By Lemma 2.1, one has bounds on x^0, w, and then by (A1), this last equation yields a bound on $|\zeta^1(0)|$. □

Also, letting $\zeta^{1,h}(t) = \frac{d}{dh}x^h(t)$ for h not necessarily zero, one has $\zeta^2(t) = \frac{d}{dh}\zeta^{1,h}(t)\big|_{h=0}$. Then

$$\dot{\zeta}^2 = f_{xx}(x^0)\zeta^1 \cdot \zeta^1 + f_x(x^0)\zeta^2 + [\sigma_{xx}(x^0)\zeta^1 \cdot \zeta^1 + \sigma_x(x^0)\zeta^2]w,$$
$$\zeta^2(\tau) = 0.$$

Again, employing bounds from Lemma 2.1 and (A1) with this equation leads to a bound for $|\zeta^2(0)|$. □

Fundamental solutions. If we assume nondegeneracy of the disturbance term in (2.1), then $\mathcal{S}_T[\phi]$ is semiconvex for $T > 0$ without assuming semiconvexity of ϕ. (The authors wish to thank P. Dupuis for pointing this out.) By nondegeneracy we mean that $m \geq n$ and

(A5) $\sigma^{-1}(x)$ is bounded uniformly in x, and also assume $\sigma^{-1} \in C^2$,

where we define the inverse by

$$\sigma^{-1}(x)b \doteq \operatorname{argmin}\{|a| : \sigma(x)a = b\}$$

for any $b \in \Re^n$. Note that assumption (A5) is only required if the initial information function, ϕ, is not semiconvex.

Let us introduce the following function, $V(T, x_0, x_T)$, which we call a *fundamental solution* of the PDE (2.14b). For $x_0, x_T \in \Re^n$, $T > 0$, let

$$V(T, x_0, x_T) = \sup_{w \in L_2([0,T];\Re^m)} \left\{ -\frac{1}{2} \int_0^T |w(t)|^2 \, dt : x(0) = x_0,\ x(T) = x_T \right\}, \tag{4.4}$$

where $x(\cdot)$ satisfies (2.1). From (4.2),

$$\mathcal{S}_T[\phi](x) = \sup_{x_0 \in \Re^n} \{\phi(x_0) + V(T, x_0, x)\}. \tag{4.5}$$

We can rewrite V in terms of the following calculus of variations problem with fixed end conditions. Let

$$L(x, \dot{x}) = \tfrac{1}{2}|\sigma^{-1}(x)(\dot{x} - f(x))|^2,$$
$$I(T, x_0, x_T; x(\cdot)) = -\textstyle\int_0^T L(x(t), \dot{x}(t)) \, dt.$$

Then

$$V(T, x_0, x_T) = \sup\{I(T, x_0, x_T; x(\cdot)) : x(0) = x_0,\ x(T) = x_T\}$$

with $x(\cdot)$ given by (2.1). Equivalence of (4.4) and this calculus of variations version follows easily by noting that for each path, $x(\cdot)$, there is a corresponding unique minimal-norm $w(\cdot)$ given in feedback form with this definition of σ^{-1}.

LEMMA 4.2. *Assume* (A5). *Then* $V(T, x_0, \cdot)$ *is semiconvex. In fact, given* $0 < T_1 < T_2 < \infty$, R, R_0, *there exists* Δ *(depending on* T_1, T_2, R, R_0*) such that* $V(T, x_0, x) + \Delta|x|^2$ *is convex on the ball* $B_R \doteq \{x \in \Re^n : |x| \le R\}$ *for any* $T \in [T_1, T_2]$ *and any* $|x_0| \le R_0$.

Proof. Given x_0, x, there exists $x^*(\cdot)$ such that $x^*(0) = x_0$, $x^*(T) = x$, and

$$V(T, x_0, x) = I(T, x_0, x; x^*(\cdot)).$$

Now suppose $0 < T_1 \le T \le T_2$, $|x_0| \le R_0$, and $|x| \le R$. Some "range-of-dependence" estimates will be obtained. For comparison with the optimal trajectory, let $\overline{x}(t) \doteq x_0 + (t/T)(x - x_0)$. Then there exists $C < \infty$ (depending on R_0, R, T_1) such that

$$\left|\dot{\overline{x}} - f(\overline{x}(t))\right| \le C \qquad \forall\, t \in [0, T],$$

and then by (A5), there exists $M_i < \infty$ such that

$$-L(\overline{x}, \dot{\overline{x}}) \ge -\frac{1}{2} M_i^2 C^2.$$

Then by the optimality of x^*,

$$I(T, x_0, x; x^*(\cdot)) \ge -\frac{1}{2} M_i^2 C^2 T,$$

and using (A2), this yields

$$\int_0^T |\dot{x}^* - f(x^*(t))|^2\,dt \le M^2 M_i^2 C^2 T. \tag{4.6}$$

Then note that by (A1), there exists $K < \infty$ such that

$$\begin{aligned}\frac{d}{dt}|x^*|^2 &= 2(x^*)^T(f(x^*) + (\dot{x}^* - f(x^*)))\\ &\le 3K|x^*|^2 + K + 2|x^*||\dot{x}^* - f(x^*)|\\ &\le (3K+1)|x^*|^2 + K + |\dot{x}^* - f(x^*)|^2.\end{aligned}$$

Employing Gronwall's inequality and (4.6), one finds that there exist $R_1, \overline{M} < \infty$ (depending on R_0, R, T_1, T_2) such that $|x^*(t)| \le R_1$ for all $t \in [0,T]$ and $\|\dot{x}^*(\cdot)\|_{L_2[0,T]} \le \overline{M}$.

Next, consider any direction $v \in \Re^n$ ($|v| = 1$) and scalar $h \in [-1,1]$. Let

$$x^h(t) \doteq x^*(t) + \frac{t}{T}hv.$$

Then $x^h(0) = x_0$ and $x^h(T) = x + hv$. Moreover,

$$\begin{aligned}&\frac{d^2}{dh^2} I(T, x_0, x+hv; x^h(\cdot)) = v^T \Lambda^h v,\\ &\Lambda^h \doteq \frac{1}{T^2}\int_0^T \left[t^2 L^h_{xx} + 2tL^h_{x\dot{x}} + L^h_{\dot{x},\dot{x}}\right] dt,\end{aligned}$$

$L^h_{xx} = L_{xx}(x^h(t), \dot{x}^h(t))$, and the other derivatives defined similarly. Since $|x^h(t)|$ and $\|\dot{x}^h(\cdot)\|_{L_2[0,T]}$ are uniformly bounded,

$$\frac{d^2}{dh^2} I(T, x_0, x+hv; x^h(\cdot)) \ge -2\Delta$$

for some constant Δ depending on T_1, T_2, R, R_0. Then $I(T, x_0, x+hv; x^h(\cdot)) + \Delta h^2$ is a convex function of h for $|h| \le 1$. Then

$$\begin{aligned}&V(T, x_0, x+hv) + V(T, x_0, x-hv) - 2V(T, x_0, x)\\ &\ge I(T, x_0, x+hv; x^h(\cdot)) + I(T, x_0, x-hv; x^{-h}(\cdot)) - 2I(T, x_0, x; x^*(\cdot))\\ &\ge -2\Delta h^2.\end{aligned}$$

This implies that $V(T, x_0, x) + \Delta|x|^2$ is convex on the ball $\{x \in \Re^n : |x| \le R\}$. ∎

THEOREM 4.3. *Assume* (A5). *For any* $\phi \le 0$, $T > 0$, $\mathcal{S}_T[\phi]$ *is semiconvex.*

Proof. If $|x_T| \le R$, by the "range-of-dependence" property, it suffices to consider those $x(\cdot)$ such that $|x(t)| \le R_1$, where $x_0 = x(0)$. If $x = x_T$, then by (4.5)

$$\begin{aligned}&\mathcal{S}_T[\phi](x) = \sup_{x_0 \in \Re^n} \{\phi(x_0) + V(T, x_0, x)\},\\ &\mathcal{S}_T[\phi](x) + \Delta|x|^2 = \sup_{x_0 \in \Re^n} \left\{\phi(x_0) + V(T, x_0, x) + \Delta|x|^2\right\},\end{aligned}$$

with Δ as in Lemma 4.2 Since the supremum of any family of convex functions is convex, $\mathcal{S}_T[\phi](x) + \Delta|x|^2$ is convex on the ball $\{x \in \Re^n : |x| \le R\}$. ∎

Theorems 4.1 and 4.3 imply (under different assumptions) that the information state (in the absence of measurement updates) will be semiconvex. To complete this discussion, one also needs to show that the measurement updates (2.14c) maintain semiconvexity.

LEMMA 4.4. *Suppose ψ is semiconvex. Then $\psi(x) - \frac{1}{2}|\rho^{-1}(x)[y - h(x)]|^2$ is semiconvex for any $y \in \Re^k$.*

Proof. Let $G(x) \doteq |\rho^{-1}(x)[y - h(x)]|^2$. Fix $R < \infty$. Then, since ψ is semiconvex, there exists $C_R < \infty$ such that $\psi(x) + \frac{C_R}{2}|x|^2$ is convex. But by (A1) and (A3b), there exists $C_{R,y} < \infty$ such that $|G_{xx}(x)| \leq C_{R,y}$ for all $x \in B_R$. Consequently $\psi(x) - \frac{1}{2}|\rho^{-1}(x)[y - h(x)]|^2 + \frac{C_R + C_{R,y}}{2}|x|^2$ is convex. □

By Theorem 2.4, Lemma 4.4, and either Theorem 4.1 or Theorem 4.3, we know that, under the given assumptions, the information state, P, will be continuous and semiconvex. Consequently, we define

$$\mathcal{P} \doteq \{\psi : \Re^n \to \Re : \psi \text{ is continuous and semiconvex}\}.$$

The algorithm to follow will depend on a max-plus basis representation of any $\psi \in \mathcal{P}$. The starting point for this representation will be the convex dual.

Let $R > 0$ and (as above) $B_R \doteq \{x \in \Re^n : |x| \leq R\}$. Let $\widehat{\psi} : B_R \to \Re$ be convex and Lipschitz over B_R with Lipschitz constant $L(R)$, and let $\widehat{\psi}(x) = +\infty$ for all $x \notin B_R$. The following result is a minor variation of a standard result in convex duality (see, for instance, [30], [31]). The slight variation here is that, in this case, one has existence of the maximizers in the statement of the lemma, which is a result of the restrictive class of functions, $\widehat{\psi}$, being considered here.

LEMMA 4.5 (convex duality representation).

$$\widehat{\psi}(x) = \max_{p \in B_{L(R)}} [p^T x + \overline{a}_p] \qquad \forall\, x \in B_R,$$

where

$$\overline{a}_p = -\max_{x \in B_R} [p^T x - \widehat{\psi}(x)],$$

and note that $\overline{a}_p$ may be $-\infty$.

Consider any $\psi \in \mathcal{P}$ over some ball B_R. Let $c > 0$ be a constant such that

$$\widehat{\psi}(x) \doteq \begin{cases} \psi(x) + \frac{c}{2}|x|^2 & \text{if } x \in B_R \\ +\infty & \text{otherwise} \end{cases} \quad \text{is convex.} \tag{4.7}$$

(The existence of c is guaranteed by $\psi \in \mathcal{P}$.) As noted above, by [8, p. 111], $\widehat{\psi}$ is Lipschitz over B_R with some constant $L(R)$. Further, if $|\psi(x)| \leq D_2(1 + |x|^2)$ for some D_2, then it can be shown using the explicit local Lipschitz bound in [8] that $L(R) \leq D_3(1 + R)$ for some D_3. The convex duality representation of $\widehat{\psi}$ in Lemma 4.5 leads easily to the following representation for ψ.

THEOREM 4.6 (semiconvex duality representation).

$$\psi(x) = \max_{\widetilde{x} \in B_{L(R)/c}} \left[-\frac{c}{2}|x - \widetilde{x}|^2 + a_{\widetilde{x}}\right] = \max_{\widetilde{x} \in \Re^n} \left[-\frac{c}{2}|x - \widetilde{x}|^2 + a_{\widetilde{x}}\right] \qquad \forall\, x \in B_R,$$

where

$$a_{\widetilde{x}} = -\max_{x \in B_R} \left[-\frac{c}{2}|x - \widetilde{x}|^2 - \psi(x)\right].$$

Proof. For any $x \in B_R$,

$$\psi(x) = \widehat{\psi}(x) - \frac{c}{2}|x|^2,$$

which by Lemma 4.5

$$\begin{aligned} &= \max_{p \in B_{L(R)}} \left[p^T x - \frac{c}{2}|x|^2 + \overline{a}_p \right] \\ &= \max_{p \in B_{L(R)}} \left[-\frac{c}{2}|x - p/c|^2 + \frac{1}{2c}|p|^2 + \overline{a}_p \right], \end{aligned}$$

which by letting $\widetilde{x} = p/c$ becomes

$$\begin{aligned} &= \max_{\widetilde{x} \in B_{L(R)/c}} \left[-\frac{c}{2}|x - \widetilde{x}|^2 + \frac{c}{2}|\widetilde{x}|^2 + \overline{a}_{c\widetilde{x}} \right] \\ &= \max_{\widetilde{x} \in B_{L(R)/c}} \left[-\frac{c}{2}|x - \widetilde{x}|^2 + a_{\widetilde{x}} \right], \end{aligned}$$

where

$$\begin{aligned} a_{\widetilde{x}} &= \frac{c}{2}|\widetilde{x}|^2 + \overline{a}_{c\widetilde{x}} \\ &= -\max_{x \in B_R} \left[(c\widetilde{x})^T x - \widehat{\psi}(x) - \frac{c}{2}|\widetilde{x}|^2 \right], \end{aligned}$$

which by (4.7)

$$\begin{aligned} &= -\max_{x \in B_R} \left[(c\widetilde{x})^T x - \psi(x) - \frac{c}{2}|x|^2 - \frac{c}{2}|\widetilde{x}|^2 \right] \\ &= -\max_{x \in B_R} \left[-\frac{c}{2}|x - \widetilde{x}|^2 - \psi(x) \right]. \end{aligned}$$

□

5. Basis representation. In this section we modify the representation of $\psi \in \mathcal{P}$ given by Theorem 4.6 to obtain a countable basis representation. Again, let $\psi \in \mathcal{P}$ and $R > 0$ with corresponding $\overline{c}$ such that (4.7) holds.

Suppose we have some countable dense subset $\{x_i\}_{i=1}^{\infty} \subseteq \Re^n$. Let $\widehat{c} \in (\overline{c}, \infty)$. Let

$$g_i(x) = -\frac{\widehat{c}}{2}|x - x_i|^2 \tag{5.1a}$$

and

$$a_i = -\max_{x \in B_R} [g_i(x) - \psi(x)]. \tag{5.1b}$$

Let $c \in (\widehat{c}, \infty)$ and

$$\overline{g}_{\widetilde{x}}(x) = -\frac{c}{2}|x - \widetilde{x}|^2.$$

By Theorem 4.6,

$$\psi(x) = \max_{\widetilde{x} \in B_{L(R)/c}} [\overline{g}_{\widetilde{x}}(x) + a_{\widetilde{x}}] \qquad \forall\, x \in B_R,$$

where $a_{\widetilde{x}}$ is given in Theorem 4.6. It will next be shown that, in fact,

$$\psi(x) = \sup_i [a_i + g_i(x)] = \bigoplus_{i=1}^{\infty} [a_i \otimes g_i(x)] \qquad \forall\, x \in B_R,$$

where $\bigoplus_{i=1}^{\infty}$ indicates max-plus summation. Some additional properties of this basis function representation will also be obtained. Lemma 5.1 will be useful in the two theorems to follow.

LEMMA 5.1. *Given $\widetilde{x} \in B_{L(R)/c}$ and $\epsilon > 0$, there exists i such that*

$$\psi(x) \geq g_i(x) + a_i \geq \overline{g}_{\widetilde{x}}(x) + a_{\widetilde{x}} - \epsilon \qquad \forall\, x \in B_R.$$

Further, for any $\delta > 0$ one may specify $|x_i - \widetilde{x}| \leq \frac{c-\widehat{c}}{c\widehat{c}}(cR + L(R)) + \delta$.

Proof. Let $\widetilde{x}$ and $\epsilon > 0$ be given. By Theorem 4.6, there exists $\overline{x} \in B_R$ such that

$$a_{\widetilde{x}} - \frac{c}{2}|\overline{x} - \widetilde{x}|^2 = \psi(\overline{x}). \tag{5.2}$$

Choose $\widehat{x} \in \Re^n$ such that

$$-\widehat{c}(\overline{x} - \widehat{x}) = -c(\overline{x} - \widetilde{x}) \tag{5.3}$$

(so that the gradient of $-\frac{\widehat{c}}{2}|x - \widehat{x}|^2$ matches that of $\overline{g}_{\widetilde{x}}$ at $\overline{x}$). Note that this implies $|\widehat{x} - \widetilde{x}| \leq \frac{c-\widehat{c}}{c\widehat{c}}(cR + L(R))$. By the density of $\{x_i\}_{i=1}^{\infty}$, given $\delta > 0$, there exists i such that

$$|x_i - \widehat{x}| < \delta. \tag{5.4}$$

Let $\widehat{a}_i$ be such that

$$\psi(\overline{x}) = \widehat{a}_i - \frac{\widehat{c}}{2}|\overline{x} - x_i|^2. \tag{5.5}$$

Define

$$\psi^0(x) = \begin{cases} \psi(x) + \frac{\overline{c}}{2}|x - \overline{x}|^2 & \text{if } x \in B_R, \\ +\infty & \text{if } x \notin B_R, \end{cases}$$

so that ψ^0 is convex. By the definition of $\overline{x}$,

$$-c(\overline{x} - \widetilde{x})^T \in \partial\psi^0(\overline{x}),$$

and so by (5.3)

$$-\widehat{c}(\overline{x} - \widehat{x})^T \in \partial\psi^0(\overline{x}),$$

which, by the definition of subgradient and that $\psi^0(\overline{x}) = \psi(\overline{x})$, implies

$$\psi(\overline{x}) - \widehat{c}(\overline{x} - \widehat{x})^T(x - \overline{x}) \leq \psi^0(x) \qquad \forall\, x \in \Re^n. \tag{5.6}$$

It will next be shown that the a_i given by (5.1b) is not much smaller than $\widehat{a}_i$ given by (5.5). Define

$$F(x) \doteq \widehat{a}_i - \frac{\widehat{c}}{2}|x - x_i|^2 + \frac{\overline{c}}{2}|x - \overline{x}|^2 - [\psi(\overline{x}) - \widehat{c}(\overline{x} - \widehat{x})^T(x - \overline{x})]. \tag{5.7}$$

The maximum of F occurs at

$$x_0 = \overline{x} + \frac{\widehat{c}}{\widehat{c} - \overline{c}}(x_i - \widehat{x}),$$

and the maximum value is

$$F(x_0) = \frac{\widehat{c}^2}{2(\widehat{c} - \overline{c})}|x_i - \widehat{x}|^2,$$

which by (5.4)

$$< \frac{\widehat{c}^2}{2(\widehat{c} - \overline{c})}\delta^2. \tag{5.8}$$

We choose $\delta > 0$ sufficiently small such that

$$\frac{\widehat{c}^2}{2(\widehat{c} - \overline{c})}\delta^2 < \frac{\epsilon}{2}. \tag{5.9}$$

Then, by (5.7), (5.8), and (5.9),

$$\widehat{a}_i - \frac{\widehat{c}}{2}|x - x_i|^2 + \frac{\overline{c}}{2}|x - \overline{x}|^2 < \psi(\overline{x}) - \widehat{c}(\overline{x} - \widehat{x})^T(x - \overline{x}) + \frac{\epsilon}{2} \qquad \forall\, x \in \Re^n,$$

and so by (5.6),

$$\widehat{a}_i - \frac{\widehat{c}}{2}|x - x_i|^2 + \frac{\overline{c}}{2}|x - \overline{x}|^2 < \psi^0(x) + \frac{\epsilon}{2} \qquad \forall\, x \in \Re^n,$$

which, by the definition of ψ^0, implies

$$\widehat{a}_i - \frac{\widehat{c}}{2}|x - x_i|^2 < \psi(x) + \frac{\epsilon}{2} \qquad \forall\, x \in B_R.$$

Consequently, with a_i given by (5.1b), one has

$$a_i > \widehat{a}_i - \frac{\epsilon}{2}. \tag{5.10}$$

Now it will be shown that $a_i - \frac{\widehat{c}}{2}|x - x_i|^2 = a_i + g_i(x)$ is not more than ϵ below $a_{\widetilde{x}} + \overline{g}_{\widetilde{x}}(x)$ (for sufficiently small δ). Define

$$H(x) = a_{\widetilde{x}} - \frac{c}{2}|x - \widetilde{x}|^2 - \left(a_i - \frac{\widehat{c}}{2}|x - x_i|^2\right). \tag{5.11}$$

By (5.2),

$$H(\overline{x}) = \psi(\overline{x}) - \left(a_i - \frac{\widehat{c}}{2}|\overline{x} - x_i|^2\right),$$

which by (5.10) and (5.5)

$$< \frac{\epsilon}{2}.$$

By (5.11)

$$\frac{dH}{dx}(\overline{x}) = -c(\overline{x} - \widetilde{x})^T + \widehat{c}(\overline{x} - x_i)^T,$$

which by (5.3)

$$= -\widehat{c}(\overline{x} - \widehat{x})^T + \widehat{c}(\overline{x} - x_i)^T,$$

so that

$$\left|\frac{dH}{dx}(\overline{x})\right| \leq \widehat{c}|\widehat{x} - x_i|.$$

Also,

$$\frac{d^2H}{dx^2}(x) = (\widehat{c} - c)I < 0$$

(whereby the last inequality, we mean that $(\widehat{c}-c)I$ is negative definite). Consequently,

$$H(x) < \frac{\epsilon}{2} + \widehat{c}|\widehat{x} - x_i||x - \overline{x}| - \frac{c - \widehat{c}}{2}|x - \overline{x}|^2 \qquad \forall\, x \in \Re^n,$$

and taking the maximum of the right-hand side yields

$$H(x) < \frac{\epsilon}{2} + \frac{\widehat{c}^2}{2(c - \widehat{c})}|\widehat{x} - x_i|^2 \qquad \forall\, x \in \Re^n,$$

which by (5.4)

$$\leq \frac{\epsilon}{2} + \frac{\widehat{c}^2}{2(c - \widehat{c})}\delta^2 \qquad \forall\, x \in \Re^n,$$

and by possibly reducing the size of δ from that in (5.9), one can make this

$$< \epsilon. \tag{5.12}$$

By (5.11) and (5.12),

$$a_i - \frac{\widehat{c}}{2}|x - x_i|^2 \geq a_{\widetilde{x}} - \frac{c}{2}|x - \widetilde{x}|^2 - \epsilon \qquad \forall\, x \in \Re^n.$$

This is the desired right inequality, and the desired left inequality is a direct consequence of (5.1b). □

THEOREM 5.2.

$$\psi(x) = \sup_i [a_i + g_i(x)] = \bigoplus_{i=1}^{\infty} [a_i \otimes g_i(x)] \qquad \forall\, x \in B_R. \tag{5.13}$$

Further, for any $\delta > 0$ it is sufficient to consider only a set of $\{x_i\}$ which is a dense subset of $B_{\delta + L(R)/c}$.

Proof. Let $\overline{x} \in B_R$. By Theorem 4.6, there exists $\widetilde{x} \in \Re^n$ such that

$$\psi(\overline{x}) = \overline{g}_{\widetilde{x}}(\overline{x}) + a_{\widetilde{x}}$$

and by Lemma 5.1, given $\epsilon > 0$, there exists i such that

$$\leq g_i(\overline{x}) + a_i + \epsilon.$$

Since $\epsilon > 0$ was arbitrary,

$$\psi(\overline{x}) \leq \sup_i [g_i(\overline{x}) + a_i],$$

and the reverse follows by the definition of the a_i. Also, the last assertion follows from Theorem 4.6 and the density of the x_i. ∎

For the remainder of the section, let $\{x_i\}$ be dense over all of $\Re^n$ in order to reduce complication. One has the countable basis function representation (5.13) for any such countable dense set $\{x_i\}$ and particular choice of constant $\widehat{c} \in (\overline{c}, \infty)$, where $\overline{c}$ is any constant such that (4.7) holds. Some further guidance on appropriate choices of $\widehat{c}$ will now be obtained. Consider several possible choices of basis function sets indexed by j,

$$\mathcal{B}_j \doteq \left\{ g_{j,i}(x) = -\frac{\widehat{c}_j}{2}|x - x_i|^2 : \{x_i\}_{i=1}^{\infty} \right\}. \tag{5.14}$$

From Theorem 5.2, we know that

$$\psi(x) = \bigoplus_{i=1}^{\infty} [a_i \otimes g_{j,i}(x)] \tag{5.15}$$

for any j such that $\widehat{c}_j > \overline{c}$. Conceptually then, given some $\psi \in \mathcal{P}$ and $R > 0$, one could choose any set of basis functions $\mathcal{B}_j \in \{\mathcal{B}_j\}$ such that

$$\widehat{c}_j > \overline{c}. \tag{5.16}$$

It will now be shown that it is in some sense optimal to choose the set $\mathcal{B}_j$ with the smallest (loosely speaking) $\widehat{c}_j$ which satisfies (5.16). This may be rigorously stated as follows.

THEOREM 5.3. *Suppose $R > 0$ and $\psi \in \mathcal{P}$ with semiconvexity constant $\overline{c}$ (so that (4.7) is satisfied). Let $\mathcal{B}_{j_1}, \mathcal{B}_{j_2} \in \{\mathcal{B}_j\}$ be such that $\widehat{c}_{j_2} > \widehat{c}_{j_1} > \overline{c}$. Then, for any $n < \infty$ and $\epsilon > 0$, there exists $\{k_i\}_{i=1}^n$ such that*

$$\psi(x) \geq \bigoplus_{i=1}^{n} a_{j_1,k_i} \otimes g_{j_1,k_i}(x) \geq \bigoplus_{i=1}^{n} a_{j_2,i} \otimes g_{j_2,i}(x) - \epsilon \qquad \forall\, x \in B_R,$$

where $a_{j_1,i}, a_{j_2,i}$ are given by (5.1b) for all i.

Before proceeding with the proof, note that this implies that one can always do arbitrarily close to better by choosing the smaller choice of $\widehat{c}_j$, but that the particular order of x_i and the function ψ may affect this. Typically, of course, one would expect better performance with the smaller choice of $\widehat{c}_j$ since the result is obviously not true if $\widehat{c}_{j_2} < \widehat{c}_{j_1}$.

Proof. Let $1 \leq i \leq n < \infty$ and $\epsilon > 0$. It is sufficient to show that there exists $k_i < \infty$ such that

$$\psi(x) \geq a_{j_1,k_i} \otimes g_{j_1,k_i}(x) \geq a_{j_2,i} \otimes g_{j_2,i}(x) - \epsilon \qquad \forall\, x \in B_R. \tag{5.17}$$

But this is simply Lemma 5.1. ∎

Finally, we turn to motivation for the choice of coefficients a_i in the expansion. Specifically we motivate the choice given by (5.1b). Note that this theorem will be true independent of the particular choice of basis functions $\{g_i\}$ in (5.1a).

THEOREM 5.4. *Let $\{a_i\}_{i=1}^{\infty}$ be given by* (5.1b) *and let $\{\widetilde{a}_i\}_{i=1}^{\infty}$ be an alternative set of coefficients. If $\widetilde{a}_{i_0} > a_{i_0}$ for some i_0, then*

$$\max_{x\in B_R}\left|\bigoplus_{i=1}^{\infty}\widetilde{a}_i\otimes g_i(x)-\psi(x)\right|>0. \tag{5.18}$$

Alternatively, if $\widetilde{a}_i \le a_i$ for all i, then

$$\max_{x\in B_R}\left|\bigoplus_{i=1}^{n}\widetilde{a}_i\otimes g_i(x)-\psi(x)\right| \ge \max_{x\in B_R}\left|\bigoplus_{i=1}^{n}a_i\otimes g_i(x)-\psi(x)\right| \qquad \forall\, n. \tag{5.19}$$

Remark 5.5. Before proceeding to the proof, let us note that this theorem implies that of the possible choices for $\{a_i\}_{i=1}^{\infty}$ such that (5.13) holds, the choice given by (5.1b) is optimal. However, this does not imply that for any fixed $n<\infty$,

$$\max_{x\in B_R}\left|\bigoplus_{i=1}^{n}a_i\otimes g_i(x)-\psi(x)\right| \le \max_{x\in B_R}\left|\bigoplus_{i=1}^{n}\widetilde{a}_i\otimes g_i(x)-\psi(x)\right|.$$

That is, there may be better choices of $\{a_i\}_{i=1}^{n}$ for finite approximations, but such a choice cannot be extended to yield ψ in the limit without changing the a_i's for $i\le n$.

Proof. For the first part, note that by (5.1a), there exists $x_0\in B_R$ such that

$$\psi(x_0)-a_{i_0}\otimes g_{i_0}(x_0)=0,$$

and consequently,

$$\psi(x_0)-\bigoplus_{i=1}^{\infty}\widetilde{a}_i\otimes g_i(x_0)\le \psi(x_0)-\widetilde{a}_{i_0}\otimes g_{i_0}(x_0),$$

which since $\widetilde{a}_{i_0}>a_{i_0}$

$$<\psi(x_0)-a_{i_0}\otimes g_{i_0}(x_0)=0.$$

For the second part, note that for any $n\ge 1$ and any $x\in B_R$,

$$\begin{aligned}\psi(x)\ge\textstyle\bigoplus_{i=1}^{n}a_i\otimes g_i(x) &= \max_{1\le i\le n}[a_i+g_i(x)]\\ &\ge \max_{1\le i\le n}[\widetilde{a}_i+g_i(x)]=\bigoplus_{i=1}^{n}\widetilde{a}_i\otimes g_i(x).\end{aligned}$$

□

6. Algorithm. Before explicitly describing the computational algorithm, we discuss how one may combine basis representation (5.13) with the max-plus linearity of the solution operator (Theorem 3.1) to update the information state between measurement times. This is actually a general algorithm for solving HJB equations of the form (2.14b) with semiconvex initial conditions, and so is more general than the particular application considered in this paper.

Specific error estimates have not been included in this section. The goal of this paper is to outline a new approach to computations for the Mortensen and robust

nonlinear filters. Since this is a new and unusual approach, and since the paper is already of substantial length, error estimates are delayed to a future paper, where computational comparisons could also be examined.

Recall the discrete-time filter discussed in section 2. Suppose $P(t_l^+, \cdot)$ is semiconvex, and that we approximate it over some ball, B_R, by some finite number of elements of the basis representation (5.13), that is,

$$P(t_l^+, x) = \bigoplus_{i=1}^{\infty} [a_i \otimes g_{j_1,i}(x)] \simeq \bigoplus_{i=1}^{n} [a_i \otimes g_{j_1,i}(x)] \qquad \forall\, x \in B_R,$$

where

$$g_{j_1,i} = -\frac{\widehat{c}_{j_1}}{2}|x - x_i|^2$$

with appropriate choice of $\widehat{c}_{j_1}$ and with a_i given by (5.1b). Then $P(t_{l+1}{}^-, x)$ is given by $\mathcal{S}_\delta[P(t_l^+, \cdot)](x)$, where $\mathcal{S}_\delta$ is the solution operator for (2.14b), where $\delta = t_{l+1} - t_l$.

We know by Theorem 3.1 that

$$\begin{aligned} P(t_{l+1}{}^-, x) &= \mathcal{S}_\delta[P(t_l^+, \cdot)](x) \\ &\simeq \mathcal{S}_\delta \left[\bigoplus_{i=1}^{n} [a_i \otimes g_{j_1,i}(\cdot)] \right](x) \\ &= \bigoplus_{i=1}^{n} \{a_i \otimes \mathcal{S}_\delta[g_{j_1,i}](x)\}. \end{aligned} \tag{6.1}$$

It is assumed that one has precomputed $\mathcal{S}_\delta[g_{j,i}]$ for all appropriate j, i, where the solutions are represented in the form

$$\mathcal{S}_\delta[g_{j,i}](x) \simeq \bigoplus_{k=1}^{n} b_{k,i} \otimes g_{j_2,k}(x), \tag{6.2}$$

where $g_{j_2,k}(x) = -\frac{\widehat{c}_{j_2}}{2}|x - x_k|^2$, $\widehat{c}_{j_2}$ is greater than the largest semiconvexity constant over the set $\{\mathcal{S}_\delta g_{j_1,i}\}_{i=1}^n$, and $b_{k,i}$ is given by (5.1b), i.e., $b_{k,i} = -\max_{x \in B_R}[g_{j_2,k}(x) - \mathcal{S}_\delta[g_i](x)]$. (Note that one might choose to vary the set of interest, B_R, from one time to another, but for simplicity of notation, we do not include that here.) In other words, one has precomputed and stored the matrices $[b_{k,i}]$ for all appropriate j_1, j_2. Then by (6.1), (6.2)

$$\begin{aligned} P(t_{l+1}{}^-, x) &\simeq \bigoplus_{i=1}^{n} \{a_i \otimes \mathcal{S}_\delta[g_{j_1,i}](x)\} \\ &\simeq \bigoplus_{i=1}^{n} \left\{ a_i \otimes \left[\bigoplus_{k=1}^{n} b_{k,i} \otimes g_{j_2,k}(x) \right] \right\} \\ &= \bigoplus_{i=1}^{n} \bigoplus_{k=1}^{n} [b_{k,i} \otimes a_i \otimes g_{j_2,k}(x)] \\ &= \bigoplus_{k=1}^{n} \left[\left(\bigoplus_{i=1}^{n} (b_{k,i} \otimes a_i) \right) \otimes g_{j_2,k}(x) \right]. \end{aligned} \tag{6.3}$$

Let B be the $n \times n$ matrix $[b_{k,i}]$, and let A be the vector $[a_i]$. Define max-plus matrix multiplication in the natural way; that is, let $C = B \otimes A$ be given by $C = [c_k]$, where $c_k = \bigoplus_{i=1}^{n} [b_{k,i} \otimes a_i]$. Then,

$$P(t_{l+1}^-, x) = \bigoplus_{i=1}^{n} [c_i \otimes g_{j_2,i}(x)],$$

where

$$C = B \otimes A. \tag{6.4}$$

In other words, the update $P(t_{l+1}^-, x) = \mathcal{S}_\delta[P(t_l^+, \cdot)](x)$ is performed simply by the max-plus matrix multiplication (6.4).

Now we will outline the structure of the information state propagation numerical algorithm for the discrete-time measurement case. Assume that the measurements occur at fixed time intervals $t_l = l\delta$ for $l = 1, 2, \ldots$. Assume that $P(0^+, \cdot) = \phi(\cdot)$ is semiconvex. (If not, then we know from Theorem 4.3 that $P(\delta, \cdot)$ will be semiconvex for $\delta > 0$, under the nondegeneracy assumption discussed there.)

Let $l = 0$.

(1) Compute an estimate of the semiconvexity constant, $\bar{c}$, for $P(t_l^+, \cdot)$ over B_R. (Since $P(t_l^+, \cdot)$ will be known only on some finite set, this is estimated by second-order differences.)

(2) Choose a set of basis functions $\mathcal{B}_{j_1}$ with $\widehat{c}_{j_1} > \bar{c}$. (In actual operations, one may choose to use more than one set, $\{\mathcal{B}_{j_1}, \mathcal{B}_{j_2}, \ldots, \mathcal{B}_{j_n}\}$, but we require $\max_{j_m} \widehat{c}_{j_m} > \bar{c}$.) Also, determine the particular subset $\{x_i\}_{i=1}^n \subset \{x_i\}_{i=1}^\infty$ to use.

(3) Compute $A^l = [a_i^l]$ for $i = 1, 2, \ldots, n$ from (5.1b). One now has the approximation

$$P(t_l^+, x) \simeq \bigoplus_{i=1}^{n} [a_i^l \otimes g_{j_1,i}(x)],$$

where $g_{j_1,i}(x) = -\frac{\widehat{c}_{j_1}}{2}|x - x_i|^2$, and so the information state $P(t_l^+, \cdot)$ is represented simply by the vector A^l.

(4) Propagate the solution of (2.14b) forward in time to time t_{l+1}. From the above discussion, we see that this takes the form of the max-plus matrix multiplication

$$A^{l+1} = B \otimes A^l.$$

(5) Compute $P(t_{l+1}{}^-, \cdot)$ from

$$P(t_{l+1}{}^-, x) = \bigoplus_{i=1}^{n} \left[a_i^{l+1} \otimes g_{j_2,i}(x)\right].$$

This step takes the representation of the information state in the form A^{l+1} and converts it back to a function of x (actually evaluated only on some grid over B_R).

(6) Given measurement y_{l+1}, perform the measurement update

$$P(t_{l+1}{}^+, x) = P(t_{l+1}{}^-, x) - \frac{1}{2}|\rho^{-1}(x)[y_{l+1} - h(x)]|^2$$

(for all points on the grid over B_R).

(7) Increment l, and return to step 1.

It is interesting to note that in the absence of measurement updates, the solution could be propagated forward in time simply by a series of max-plus matrix multiplications. This would be the case if there were other applications where one was only interested in solution of an HJB equation of the form (2.14b) with semiconvex initial conditions or nondegenerate σ. However, in the robust filter application, the largest portion of the computational effort is likely to be expended in step 3, where one needs to compute the coefficients in the expansion. In many other types of expansions, the coefficients are computed by integration. Here, we see that the integration step is replaced by a maximization over the relevant portion of the state space, B_R.

7. Continuous-time measurements. Let us return to the case of continuous-time measurements satisfying (2.2), and let $P(T,x)$ be the information state given by (2.6). Let $P^\delta(T,x)$ be the discrete-time measurement information state with time-step δ. (To be specific, we take $P^\delta(t,\cdot)$ to be $P(t^+,\cdot)$ at the measurement times, although $P(t^-,\cdot)$ would work as well.) In this section, we will show that P is the limit of P^δ as $\delta \downarrow 0$ (Theorem 7.3). Section 6 described a numerical algorithm for finding $P^\delta(T,x)$ approximately, and hence for approximate solution of the HJB equation (2.9) for $P(T,x)$.

We use the simplest discretization of the problem, letting $t_l = l\delta$ for $l = 0, 1, 2 \ldots$. At each $l \geq 1$, the discrete approximation model receives a measurement y_l given by

$$y_l = \frac{1}{\delta} \int_{t_{l-1}}^{t_l} y(t)\, dt. \tag{7.1}$$

Given T, let N be the largest integer such that $t_N \leq T$, and let

$$J^\delta(T,x,w) = \phi(x_0) - \frac{1}{2}\int_0^T |w(t)|^2\, dt - \frac{\delta}{2}\sum_{l=1}^{N} |\rho^{-1}(x(t_l))[y_l - h(x(t_l))]|^2, \tag{7.2}$$

where $x_0 = x(0)$ given $x(T) = x$ and dynamics (2.1), and

$$P^\delta(T,x) = \sup_{w \in L_2} J^\delta(T,x,w). \tag{7.3}$$

Note that if $T = N\delta$, one has $P^\delta(T,x)$ being what was denoted as $P(T^+,x)$ in (2.13b) rather than $P(T^-,x)$.

It will first be shown that the semiconvexity constants for the approximations P^δ (which we recall affect the choice of basis functions in the algorithm of section 6) do not blow up as $\delta \downarrow 0$.

THEOREM 7.1. *Assume that ϕ is semiconvex. Then for any T^0, R, there exists a constant $\widehat{\Gamma}$ (depending on T^0, R) such that for any $T \in [0, T^0]$, $P^\delta(T,x) + \widehat{\Gamma}|x|^2$ is convex on B_R.*

Proof. Let

$$\begin{aligned} J^{\delta,0}(T,x,w) &= \phi(x_0) - \frac{1}{2}\int_0^T |w(t)|^2\, dt, \\ J^{\delta,1}(T,x,w) &= -\frac{\delta}{2}\sum_{l=1}^{N} |\rho^{-1}(x(t_l))[y_l - h(x(t_l))]|^2, \end{aligned}$$

so that $J^\delta = J^{\delta,0} + J^{\delta,1}$. In the proof of Theorem 4.1, one finds that given T^0, R, there exists Γ such that

$$v^T J^{\delta,0}_{xx}(x,T,w)v \geq -2\Gamma \tag{7.4}$$

for all $|x| \leq R$, $T \in [0,T^0]$, $|v| = 1$, and ϵ_0-optimal w with $\epsilon_0 \leq 1$. In particular, $\Gamma = \Gamma_R(T^0)$ is independent of δ. All that remains is to prove a similar result for $J^{\delta,1}$. It will now be shown that there exists $\overline{\Gamma}$ (depending on T^0, R) such that

$$v^T J^{\delta,1}_{xx}(x,T,w)v \geq -2\overline{\Gamma}$$

for all $|x| \leq R$, $T \in [0,T^0]$, $|v| = 1$, and ϵ_0-optimal w with $\epsilon_0 \leq 1$.

As in the proof of Theorem 4.1, let $\zeta^1(t), \zeta^2(t)$ denote the first and second derivatives of $x(t)$ in the direction v (considered as a function of $x = x(T)$). Then

$$\begin{aligned} v^T J^{\delta,1}_{xx}(x,T,w)v = -\frac{\delta}{2}\sum_{l=1}^{N}\Big\{ & \left(|\rho^{-1}(x)[y_l - h(x)]|^2\right)_x \Big|_{x=x(t_l)} \zeta^2(t_l) \\ & + \left[\left(|\rho^{-1}(x)[y_l - h(x)]|^2\right)_{xx} \Big|_{x=x(t_l)} \zeta^1(t_l)\right] \cdot \zeta^1(t_l)\Big\}. \end{aligned} \tag{7.5}$$

From Lemma 2.1, we know that there exists $R_1 = R_1(T^0, R)$ such that $|x(t)| \leq R_1 \, \forall t \in [0,T]$, which leads, via (A1), to bounds on all x-dependent terms in (7.5). That is, there exists $M_1 = M_1(T^0, R)$ such that

$$v^T J^{\delta,1}_{xx}(x,T,w)v \geq -\frac{\delta}{2}\sum_{l=1}^{N} C_\rho^2 \left\{\left(M_1 + |y_l|^2\right)\left(|\zeta^2(t_l)| + |\zeta^1(t_l)|^2\right)\right\}.$$

But as shown in the proof of Theorem 4.1, $|\zeta^2(t)|, |\zeta^1(t)|$ are bounded with bounds only dependent on T^0, R. Therefore, there exists $M_2 = M_2(T^0, R)$ such that

$$\begin{aligned} v^T J^{\delta,1}_{xx}(x,T,w)v &\geq -\frac{\delta}{2}\sum_{l=1}^{N}\left\{M_2(1 + |y_l|^2)\right\} \\ &= -\frac{\delta}{2}\sum_{l=1}^{N}\left\{M_2\left(1 + \left|\frac{1}{\delta}\int_{t_{l-1}}^{t_l} y(t)\,dt\right|^2\right)\right\} \\ &\geq -\frac{\delta}{2}\sum_{l=1}^{N}\left\{M_2\left(1 + \frac{1}{\delta}\int_{t_{l-1}}^{t_l} |y(t)|^2\,dt\right)\right\} \\ &\geq \frac{-M_2}{2}\left(T + \|y\|^2_{L_2[0,T]}\right), \end{aligned}$$

which we define to be

$$\doteq -2\overline{\Gamma}. \tag{7.6}$$

Combining (7.4) and (7.6), one has

$$v^T J^\delta_{xx}(x,T,w)v \geq -2(\Gamma + \overline{\Gamma})$$

for all $|x| \leq R$, ϵ_0-optimal w with $\epsilon_0 \leq 1$, where $\Gamma, \overline{\Gamma}$ are independent of δ. This implies the desired result. □

Remark 7.2. The P^δ satisfy a uniform (in δ) local Lipschitz condition in x. That is, given $R, T < \infty$, there exists $K = K(T,R)$ such that $|P^\delta(T,x) - P^\delta(T,\widetilde{x})| \le K|x - \widetilde{x}|$ for all $x, \widetilde{x} \in B_R$. This follows directly from the uniform semiconvexity; see [8, p. 111]

Finally, we obtain the convergence result.

THEOREM 7.3. *$P^\delta(T,x) \to P(T,x)$ as $\delta \downarrow 0$ for all $(T,x) \in [0,\infty) \times \Re^n$.*

Proof. Fix $T, \overline{R} < \infty$ and $x \in B_{\overline{R}}$. Let $\epsilon_0 \in [0,1]$. Let J be given by (2.5) and J^δ by (7.2). By Lemma 2.1 (and a similar result for the discrete case), we know ϵ_0-optimal w for either problem satisfies $\|w\|_{L_2[0,T]} \le M_{\overline{R},T}$ for appropriate $M_{\overline{R},T}$.

Let $\|w\|_{L_2[0,T]} \le M_{\overline{R},T}$. It is sufficient to show that

$$|J(T,x,w) - J^\delta(T,x,w)| \to 0$$

as $\delta \downarrow 0$ (with an estimate that is independent of w). We have

$$2|J(T,x,w) - J^\delta(T,x,w)| = \left| \int_0^T |\rho^{-1}(x(t))[y(t) - h(x(t))]|^2 \, dt - \delta \sum_{l=1}^{N} |\rho^{-1}(x(t_l))[y_l - h(x(t_l))]|^2 \right|. \tag{7.7}$$

We now show that given $\epsilon > 0$, the right-hand side of (7.7) can be made $< \epsilon$ for $\delta > 0$ sufficiently small. The proof will involve first an approximation of y by an L_∞ function, and then a mollification.

Let $R > 0$ and

$$y^R(t) \doteq \begin{cases} y(t) & \text{if } |y(t)| \le R, \\ 0 & \text{otherwise.} \end{cases}$$

Let $\Xi_R = \{t \in [0,T] : |y(t)| > R\}$. Then

$$\int_0^T \Big[|\rho^{-1}(x(t))[y(t) - h(x(t))]|^2 - |\rho^{-1}(x(t))[y^R(t) - h(x(t))]|^2 \Big] \, dt = \int_{\Xi_R} |\rho^{-1}(x(t))[y(t) - h(x(t))]|^2 dt,$$

which by (2.3), (A1), and Lemma 2.1

$$\le 2C_\rho^2 \int_{\Xi_R} (M_1 + |y(t)|^2) \, dt$$

for the proper choice of M_1, and then for R sufficiently large,

$$< \frac{\epsilon}{4}. \tag{7.8}$$

Also,

$$\delta\sum_{l=1}^{N}\left|\left|\rho^{-1}(x(t_l))\left[\frac{1}{\delta}\int_{t_{l-1}}^{t_l}y(t)\,dt-h(x(t_l))\right]\right|^2\right.$$
$$\left.-\left|\rho^{-1}(x(t_l))\left[\frac{1}{\delta}\int_{t_{l-1}}^{t_l}y^R(t)\,dt-h(x(t_l))\right]\right|^2\right|$$
$$\le\delta C_\rho^2\delta\sum_{l=1}^{N}\left|\frac{1}{\delta}\int_{t_{l-1}}^{t_l}(y(t)+y^R(t))\,dt-2h(x(t_l))\right|\left|\frac{1}{\delta}\int_{t_{l-1}}^{t_l}(y(t)-y^R(t))\,dt\right|,$$

and using Lemma 2.1 and Cauchy–Schwarz, one finds that for proper choice of M_2,

$$\le\delta C_\rho^2\delta\sum_{l=1}^{N}\left\{\frac{1}{\sqrt{\delta}}\left[\int_{t_{l-1}}^{t_l}|y(t)+y^R(t)|^2\,dt\right]^{\frac12}\frac{1}{\sqrt{\delta}}\left[\int_{t_{l-1}}^{t_l}|y(t)-y^R(t)|^2\,dt\right]^{\frac12}\right.$$
$$\left.+\frac{M_2}{\delta}\int_{t_{l-1}}^{t_l}|y(t)-y^R(t)|\,dt\right\},$$

which, using $2ab\le ca^2+b^2/c$ for any $c>0$,

$$\le C_\rho^2\left\{\frac{c}{2}\int_0^T|y+y^R|^2\,dt+\frac{1}{2c}\int_0^T|y-y^R|^2\,dt+M_2\int_0^T|y-y^R|\,dt\right\},$$

which for c sufficiently small and then R sufficiently large

$$<\frac{\epsilon}{4}. \tag{7.9}$$

From (7.7),

$$2|J(T,x,w)-J^\delta(T,x,w)|$$
$$\le\left|\int_0^T|\rho^{-1}(x(t))[y(t)-h(x(t))]|^2-|\rho^{-1}(x(t))[y^R(t)-h(x(t))]|^2,dt\right|$$
$$+\left|\int_0^T|\rho^{-1}(x(t))[y^R(t)-h(x(t))]|^2\,dt\right.$$
$$\left.-\sum_{l=1}^{N}\int_{t_{l-1}}^{t_l}\left|\rho^{-1}(x(t))\left[\frac{1}{\delta}\int_{t_{l-1}}^{t_l}y^R(r)\,dr-h(x(t))\right]\right|^2dt\right|$$
$$+\left|\sum_{l=1}^{N}\int_{t_{l-1}}^{t_l}\left\{\left|\rho^{-1}(x(t))\left[\frac{1}{\delta}\int_{t_{l-1}}^{t_l}y^R(r)\,dr-h(x(t))\right]\right|^2\right.\right.$$
$$\left.\left.-\left|\rho^{-1}(x(t_l))\left[\frac{1}{\delta}\int_{t_{l-1}}^{t_l}y^R(r)\,dr-h(x(t_l))\right]\right|^2\right\}dt\right|$$
$$+\left|\sum_{l=1}^{N}\delta\left\{\left|\rho^{-1}(x(t_l))\left[\frac{1}{\delta}\int_{t_{l-1}}^{t_l}y^R(r)\,dr-h(x(t_l))\right]\right|^2\right.\right.$$
$$\left.\left.-\left|\rho^{-1}(x(t_l))\left[\frac{1}{\delta}\int_{t_{l-1}}^{t_l}y(r)\,dr-h(x(t_l))\right]\right|^2\right\}dt\right|,$$

which by (7.8), (7.9), and some rearrangement is

$$
\begin{aligned}
&< \left| \sum_{l=1}^{N} \int_{t_{l-1}}^{t_l} \left\{ |\rho^{-1}(x(t))[y^R(t) - h(x(t))]|^2 \right.\right. \\
&\qquad \left.\left. - \left| \rho^{-1}(x(t)) \left[\frac{1}{\delta} \int_{t_{l-1}}^{t_l} y^R(r)\, dr - h(x(t)) \right] \right|^2 \right\} dt \Bigg| \\
&\quad + \left| \sum_{l=1}^{N} \int_{t_{l-1}}^{t_l} \left\{ \left| \rho^{-1}(x(t)) \left[\frac{1}{\delta} \int_{t_{l-1}}^{t_l} y^R(r)\, dr - h(x(t)) \right] \right|^2 \right.\right. \\
&\qquad \left.\left. - \left| \rho^{-1}(x(t_l)) \left[\frac{1}{\delta} \int_{t_{l-1}}^{t_l} y^R(r)\, dr - h(x(t_l)) \right] \right|^2 \right\} dt \Bigg| \\
&\quad + \int_{t_N}^{T} |\rho^{-1}(x(t))[y^R(t) - h(x(t))]|^2\, dt + \frac{\epsilon}{2},
\end{aligned}
$$

which by the continuity of $x(\cdot)$ over $[0,T]$ for $\|w\| < M_{\overline{R},T}$

$$
\begin{aligned}
&< \left| \sum_{l=1}^{N} \int_{t_{l-1}}^{t_l} \left\{ |\rho^{-1}(x(t))[y^R(t) - h(x(t))]|^2 \right.\right. \\
&\qquad \left.\left. - \left| \rho^{-1}(x(t)) \left[\frac{1}{\delta} \int_{t_{l-1}}^{t_l} y^R(r)\, dr - h(x(t)) \right] \right|^2 \right\} dt \right| + M_3\delta + \frac{3\epsilon}{4} \\
&= \left| \sum_{l=1}^{N} \int_{t_{l-1}}^{t_l} \left\{ \left| \rho^{-1}(x(t)) \left[\frac{1}{\delta} \int_{t_{l-1}}^{t_l} y^R(t)\, dr - h(x(t)) \right] \right|^2 \right.\right. \\
&\qquad \left.\left. - \left| \rho^{-1}(x(t)) \left[\frac{1}{\delta} \int_{t_{l-1}}^{t_l} y^R(r)\, dr - h(x(t)) \right] \right|^2 \right\} dt \right| + M_3\delta + \frac{3\epsilon}{4},
\end{aligned}
$$

which for proper choice of $\xi(t) \in B_R$ for all t,

$$
< 2C_\rho^2 \sum_{l=1}^{N} \int_{t_{l-1}}^{t_l} |\xi(t) - h(x(t))| \left| \frac{1}{\delta} \int_{t_{l-1}}^{t_l} (y^R(t) - y^R(r))\, dr \right| dt + M_3\delta + \frac{3\epsilon}{4},
$$

which for proper choice of $M_4 = M_4(R,T)$

$$
< M_4 \sum_{l=1}^{N} \int_{t_{l-1}}^{t_l} \left| \frac{1}{\delta} \int_{t_{l-1}}^{t_l} (y^R(t) - y^R(r))\, dr \right| dt + M_3\delta + \frac{3\epsilon}{4}. \tag{7.10}
$$

Now let $y^{R,\rho}$ be a C^∞ mollification of y^R such that $\|y^{R,\rho} - y^R\|_{L_1[0,T]} \to 0$ as $\rho \downarrow 0$. Rearranging (7.10) yields

$$
\begin{aligned}
&2|J(T,x,w) - J^{\delta}(T,x,w)| \\
&\quad < M_4 \sum_{l=1}^{N} \int_{t_{l-1}}^{t_l} \left| \frac{1}{\delta} \int_{t_{l-1}}^{t_l} (y^{R,\rho}(t) - y^{R,\rho}(r))\, dr \right| dt \\
&\qquad + M_4 \sum_{l=1}^{N} \int_{t_{l-1}}^{t_l} \left| \left| \frac{1}{\delta} \int_{t_{l-1}}^{t_l} (y^{R}(t) - y^{R}(r))\, dr \right| - \left| \frac{1}{\delta} \int_{t_{l-1}}^{t_l} (y^{R,\rho}(t) - y^{R,\rho}(r))\, dr \right| \right| dt \\
&\qquad + M_3\delta + \frac{3\epsilon}{4},
\end{aligned}
$$

and letting $\mu^{\rho}(\cdot)$ be the modulus of continuity of $y^{R,\rho}$, and using the inequality $\big||a-b| - |c-d|\big| \le |a-c| + |b-d|$ on the second sum, one finds that this is

$$
\begin{aligned}
&\le M_4 T \mu^{\rho}(\delta) + M_4 \delta \sum_{l=1}^{N} \left[\int_{t_{l-1}}^{t_l} |y^{R}(t) - y^{R,\rho}(t)|\, dt + \int_{t_{l-1}}^{t_l} |y^{R}(r) - y^{R,\rho}(r)|\, dr \right] \\
&\qquad + M_3\delta + \frac{3\epsilon}{4}, \\
&= M_4 T \mu^{\rho}(\delta) + 2M_4 \|y^{R} - y^{R,\rho}\|_{L_1[0,T]} + M_3\delta + \frac{3\epsilon}{4},
\end{aligned}
$$

which for ρ sufficiently small

$$
< M_4 T \mu^{\rho}(\delta) + M_3\delta + \frac{7\epsilon}{8},
$$

which for δ sufficiently small

$$
< \epsilon. \qquad \square
$$

Remark 7.4. In the case where $y(\cdot)$ is continuous, one can obtain an explicit estimate of the convergence rate. Specifically, using only the bound $\|w\|_{L_2[0,t]} \le M_{\overline{R},T}$, for ϵ_0-optimal disturbances for both continuous- and discrete-time problems with $\delta \le 1$, one can obtain an estimate of the form $|P^{\delta}(T,x) - P(T,x)| \le \mathcal{M}[\sqrt{\delta} + \mu(\delta)]$, where $\mu(\cdot)$ is the modulus of continuity for y. We expect that with additional effort, one can obtain a uniform L_∞ bound on the optimal disturbances for the discrete-time problems (in analogy with the continuous-time bound of Lemma 2.1). This would lead to an estimate of the form $|P^{\delta}(T,x) - P(T,x)| \le \mathcal{M}[\delta + \mu(\delta)]$. However, the effort would likely require more journal space than would seem appropriate for this initial exploration.

Acknowledgments. The authors would like to thank P. Dupuis and K. Ito for helpful suggestions.

REFERENCES

[1] T. Başar and P. Bernhard, *H_∞-Optimal Control and Related Minimax Design Problems: A Dynamic Game Approach*, 2nd ed., Birkhäuser, Basel, 1995.

[2] M. Bardi and F. Da Lio, *On the Bellman equation for some unbounded control problems*, NoDEA Nonlinear Differential Equations Appl., 4, (1997), pp. 491–510.

[3] F. L. Baccelli, G. Cohen, G. J. Olsder, and J.-P. Quadrat, *Synchronization and Linearity*, John Wiley, New York, 1992.

[4] E. N. Barron and R. Jensen, *Generalized viscosity solutions for Hamilton–Jacobi equations with time–measurable Hamiltonians*, J. Differential Equations, 69 (1987), pp. 10–21.

[5] R. K. Boel, M. R. James, and I. R. Petersen, *Robustness and risk-sensitive filtering,* submitted.

[6] M. H. A. Davis, *Lectures on Stochastic Control and Nonlinear Filtering,* Tata Inst. Fund. Res. Lectures on Math. and Phys. 75, Springer-Verlag, New York, 1984.

[7] F. Da Lio and W. M. McEneaney, *Finite time-horizon risk-sensitive control and the robust limit under a quadratic growth assumption,* SIAM J. Control Optim., submitted.

[8] W. H. Fleming, *Functions of Several Variables*, Springer-Verlag, New York, 1977.

[9] W. H. Fleming, *Deterministic nonlinear filtering,* Ann. Scuola Norm. Sup. Pisa Cl. Sci. (4), 24 (1997), pp. 435–454.

[10] W. H. Fleming and W. M. McEneaney, *Risk sensitive and robust nonlinear filtering,* in Proceedings of the 36th IEEE Conference on Decision and Control, San Diego, 1997, pp. 1088–1093.

[11] W. H. Fleming and E. Pardoux, *Optimal control of partially observed diffusions,* SIAM J. Control Optim., 20 (1982), pp. 261–285.

[12] O. J. Hijab, *Minimum Energy Estimation,* Ph.D. thesis, University of California, Berkeley, CA, 1980.

[13] J. W. Helton and M. R. James, *Extending H^∞-Control to Nonlinear Systems: Control of Nonlinear Systems to Achieve Performance Objectives*, SIAM, Philadelphia, 1999.

[14] H. Ishii, *Hamilton–Jacobi equations with discontinuous Hamiltonians on arbitrary open sets,* Bull. Fac. Sci. Engrg. Chuo Univ. Ser. I Math., 28 (1985), pp. 33–77.

[15] H. Ishii, *Comparison results for Hamilton–Jacobi equations without a growth condition on solutions from above,* Appl. Anal., to appear.

[16] K. Ito, *Approximation of the Zakai equation for nonlinear filtering,* SIAM J. Control Optim., 34 (1996), pp. 620–634.

[17] M. R. James and J. S. Baras, *Partially observed differential games, infinite-dimensional Hamilton–Jacobi–Isaacs equations, and nonlinear H_∞ control,* SIAM J. Control Optim., 34 (1996), pp. 1342–1364.

[18] A. J. Krener and A. Duarte, *A hybrid computational approach to nonlinear estimation,* in Proceedings of the 35th IEEE Conference on Decision and Control, Japan, 1996.

[19] H. Kunita, *Stochastic Partial Differential Equations Connected with Nonlinear Filtering,* Lecture Notes in Math. 972, Springer-Verlag, New York, 1982.

[20] G. L. Litvinov and V. P. Maslov, *Correspondence principle for idempotent calculus and some computer applications,* in Idempotency, J. Gunawardena, ed., Publ. Newton Inst. 11, Cambridge University Press, Cambridge, UK, 1998, pp. 420–443.

[21] S. Lototsky, R. Mikulevicius, and B. L. Rozovskii, *Nonlinear filtering revisited: A spectral approach,* SIAM J. Control Optim., 35 (1997), pp. 435–461.

[22] P. L. Lions and B. Perthame, *Remarks on Hamilton–Jacobi equations with measurable time–dependent Hamiltonians,* Nonlinear Anal., 11 (1987), pp. 613–621.

[23] V. P. Maslov, *On a new principle of superposition for optimization problems,* Russian Math. Surveys, 42 (1987), pp. 43–54.

[24] W. M. McEneaney, *Robust/H_∞ filtering for nonlinear systems,* Systems Control Lett., 33 (1998), pp. 315–325.

[25] W. M. McEneaney, *Uniqueness for viscosity solutions of nonstationary Hamilton–Jacobi–Bellman equations under some a priori conditions (with applications),* SIAM J. Control Optim., 33 (1995), pp. 1560–1576.

[26] W. M. McEneaney, *Robust control and differential games on a finite time horizon*, Math. Control Signals Systems, 8 (1995), pp. 138–166.

[27] W. M. McEneaney and M. V. Day, *Characteristic characterization of viscosity supersolutions corresponding to H_∞ control,* in Proceedings of the IFAC 13th World Congress, Vol. E, San Francisco, 1996, pp. 401–406.

[28] S. K. Mitter, *Lectures on Nonlinear Filtering and Stochastic Control,* Lecture Notes in Math. 972, Springer-Verlag, New York, 1982.

[29] R. E. Mortensen, *Maximum likelihood recursive nonlinear filtering,* J. Optim. Theory Appl., 2 (1968), pp. 386–394.
[30] R. T. Rockafellar, *Conjugate Duality and Optimization*, CBMS-NSF Regional Conf. Ser. in Appl. Math. 16, SIAM, Philadelphia, 1974.
[31] R. T. Rockafellar and R. J. Wets, *Variational Analysis*, Springer-Verlag, New York, 1997.
[32] P. Soravia, *Evolution operators and viscosity solutions of Hamilton–Jacobi equations,* Boll. Un. Mat. Ital. B (7), 2 (1988), pp. 571–600.

SIAM J. CONTROL OPTIM.
Vol. 38, No. 3, pp. 711–735

CONTROLLABILITY AND STABILIZATION OF A CANAL WITH WAVE GENERATORS*

STÉPHANE MOTTELET†

Abstract. This paper deals with a physical system that can be represented by an abstract wave equation: a canal with wave generators.

We first study the various degrees of controllability that this system can enjoy, with two kinds of input operators and associated control spaces, corresponding to "flexible" and "rigid" generators. A counterexample to the exact controllability and a positive result for the approximate controllability are given for the flexible generator case. In the rigid case, we show that approximate controllability in finite time does not hold.

We also study the stability of the system when the elevation of the surface is measured at $x = 0$, and a static feedback is used to control a rigid generator. We show that strong stability holds (but with a nonuniform decay), although the perturbation caused by the feedback on the system operator is not dissipative in the natural topology.

Key words. boundary control, hyperbolic system, nonharmonic Fourier series, hydrodynamics

AMS subject classifications. 76B15, 35B37, 93C20, 93D15

PII. S0363012998347134

1. Statement of the problem.

The physical system under consideration is a canal where the behavior of floating bodies is studied from an experimental point of view. In such a canal, waves are created by means of plane generators which can be controlled.

This paper deals with the mathematical analysis of control problems involving this system and, more precisely, of various degrees of controllability and feedback stabilization.

Practical control problems involving this system have been studied, from a numerical and experimental point of view. For these aspects we refer to [14], [15], [7].

The paper is organized as follows : in section 2 we briefly introduce the model of the canal (for a more complete discussion of the model see [14]). In section 3 we reformulate the original model in the framework of Riesz-spectral systems. This will allow us to use the standard tools concerning these systems, especially for controllability tests.

In section 4 we focus on controllability problems, with two types of input operators, concerning two types of generators. For the case where the generator is "flexible," the input space is infinite-dimensional and we give a positive result for the approximate controllability and a counterexample to the exact controllability. For the case where the generator is "rigid," the input space is $\mathbf{R}$ and we will see that the spectral properties of the system operator are such that approximate controllability in finite time does not hold. This result has been obtained by means of results on the nonharmonic Fourier series.

In section 5 we consider a feedback stabilization problem which has been already set in [14], but the only type of stability we obtained was input-output stability. We

*Received by the editors November 9, 1998; accepted for publication (in revised form) July 6, 1999; published electronically February 24, 2000.
http://www.siam.org/journals/sicon/38-3/34713.html
†Division Mathématiques Appliquées, Département de Génie Informatique, Université de Technologie de Compiègne, BP 20529, 60205 Compiegne Cedex, France (stephane.mottelet@utc.fr).

give here a strong stability result, which has been obtained by means of an ad hoc energy functional.

2. Model of the canal. We consider the rectangular canal in Figure 2.1. This canal is supposed to be wide enough to consider that the waves created by generators located at each end are plane waves. This allows us to use a bidimensional model, where the domain Ω is the rectangle $[0, L] \times [0, h]$, represented in Figure 2.1. The boundary $\Gamma = \Gamma_s \cup \Gamma_f \cup \Gamma_1 \cup \Gamma_2$ is represented by

- the free surface $\Gamma_s = \{(x, h) \mid 0 < x < L\}$,
- the bottom of the canal $\Gamma_f = \{(x, 0) \mid 0 < x < L\}$,
- the left end $\Gamma_1 = \{(0, y) \mid 0 < y < h\}$,
- the right end $\Gamma_2 = \{(L, y) \mid 0 < y < h\}$.

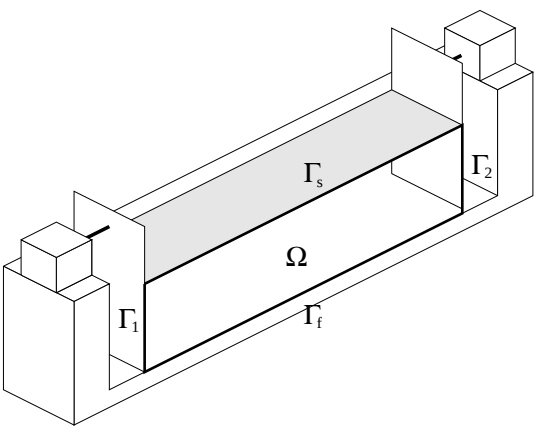


FIG. 2.1. *Definition of domain* Ω.

The fluid is supposed to be perfect, incompressible, and irrotational. Let $\vec{V}(x, y, t)$ be the velocity field at time t. From the hypothesis $\operatorname{curl} \vec{V} = 0$, there exists a potential ψ defined by

$$\vec{V}(x, y, t) = \vec{\nabla}\, \psi(x, y, t).$$

We now study the boundary conditions on Γ. The boundary condition on Γ_s is a dynamic condition. Let us call $\eta(x, t)$ the elevation of a point $M(x, h)$ of Γ_s with respect to its equilibrium position. From the hypothesis we have made the static pressure $P(x, y, t)$ is related to the velocity potential ψ by the Bernoulli condition

$$\frac{P}{\rho} + 1/2|\, \vec{\nabla}\, \psi|^2 + \dot{\psi} + g_0 \eta = \frac{P_a}{\rho}, \text{ on } \Gamma_s,$$

where ρ is the (constant) volumic mass, g_0 is the acceleration of gravity, and P_a is the atmospheric pressure (supposed to be constant). The Bernoulli condition expresses the continuity of pressure across the free surface and takes the following form:

$$1/2|\, \vec{\nabla}\, \psi|^2 + \dot{\psi} + g_0 \eta = 0, \text{ on } \Gamma_s.$$

Once linearized under the hypothesis of small fluid motion, this condition takes the form

$$\dot{\psi} + g_0\eta = 0, \text{ on } \Gamma_s. \tag{2.1}$$

The kinematic condition, which expresses the fact that the vertical component of the velocity of a fluid particle $M(x,h)$ of the free surface is equal to the time derivative of $\eta(x,t)$, takes the form

$$\dot{\eta} = \partial_n\psi, \text{ on } \Gamma_s. \tag{2.2}$$

One can eliminate η between (2.1) and (2.2) and obtain the so-called free surface condition,

$$\ddot{\psi} + g_0\partial_n\psi = 0, \text{ on } \Gamma_s, \tag{2.3}$$

where $\ddot{\psi}$ is the second time derivative of ψ with respect to time, and $\eta(x,t)$ takes the form

$$\eta(x,t) = -\frac{1}{g_0}\dot{\psi}(x,h,t). \tag{2.4}$$

The condition on the bottom Γ_f expresses the fact that the normal velocity of the fluid on Γ_f is zero. If one calls $\vec{n}$ the outer normal to Γ, we have

$$\partial_n\psi \equiv \vec{\nabla}\,\psi.\vec{n} = 0, \text{ on } \Gamma_f. \tag{2.5}$$

Finally, we obtain the following equations:

$$\begin{cases} \Delta\psi &= 0, & \text{in } \Omega\times[0,\tau], \\ \ddot{\psi} + g_0\partial_n\psi &= 0, & \text{on } \Gamma_s\times[0,\tau], \\ \partial_n\psi &= 0, & \text{on } \Gamma_f\times[0,\tau], \\ \partial_n\psi &= v_1, & \text{on } \Gamma_1\times[0,\tau], \\ \partial_n\psi &= v_2, & \text{on } \Gamma_2\times[0,\tau]. \end{cases}$$

As we will see in what follows, we only need to specify initial conditions on Γ_s:

$$\psi(0) = \varphi_0, \dot{\psi}(0) = \varphi_1, \text{ on } \Gamma_s. \tag{2.6}$$

The boundary conditions on Γ_1 and Γ_2 take into account the action of generators which can produce desired velocities $v_1(y,t)$ and $v_2(y,t)$. This kind of control will allow us to consider in section 4.1 an exact controllability problem. The devices to produce these velocities have to be some kind of "flexible" generators.

When the generators have a fixed "shape" described by the function $f(y)$, we consider the following form for the controls:

$$v_1(y,t) = f(y)u_1(t),$$
$$v_2(y,t) = f(y)u_2(t).$$

In this case, we will speak of "rigid" generators. We have scalar controls $u_1(t)$ and $u_2(t)$. When $f(y) = y$ these boundary conditions represent plane generators being able to rotate around an axis located at the bottom of the canal, with small angular velocities $u_1(t)$ and $u_2(t)$.

To simplify the notations in what follows and without loss of generality, we will take $g_0 = 1$, $L = \pi$, $h = 1$ and we will consider only one generator, at the left end of the canal; thus we will note

$$v := v_1,\ u := u_1,$$

and we will take $v_2 = 0$.

3. Mathematical analysis.

3.1. Regularity results. We now consider the following equations, corresponding to a canal of longitudinal section $\Omega =]0, \pi[\times]0, 1[$, with one generator on the left side:

$$\Delta\psi = 0, \text{ in } \Omega \times [0, \tau], \tag{3.1}$$

$$\ddot{\psi} + \partial_n \psi = 0, \text{ on } \Gamma_s \times [0, \tau], \tag{3.2}$$

$$\partial_n \psi = 0, \text{ on } (\Gamma_f \cup \Gamma_2) \times [0, \tau], \tag{3.3}$$

$$\partial_n \psi = v, \text{ on } \Gamma_1 \times [0, \tau]. \tag{3.4}$$

Let D denote the "Dirichlet map," i.e., the continuous map $\mathrm{D} : H^{1/2}(\Gamma_s) \to H^1(\Omega)$ defined by $\mathrm{D}\varphi = \Phi$ where

$$\left\{ \begin{array}{rcll} \Delta\Phi & = & 0, & \text{in } \Omega, \\ \Phi & = & \varphi, & \text{on } \Gamma_s, \\ \partial_n\Phi & = & 0, & \text{on } \Gamma_f \cup \Gamma_1 \cup \Gamma_2. \end{array} \right. \tag{3.5}$$

Let N denote the "Neumann map," i.e., the continuous map $\mathrm{N} : H^{-1/2}(\Gamma_s) \to H^1(\Omega)$ defined by $\mathrm{N}v = \Psi$ where

$$\left\{ \begin{array}{rcll} \Delta\Psi & = & 0, & \text{in } \Omega, \\ \Psi & = & 0, & \text{on } \Gamma_s, \\ \partial_n\Psi & = & 0, & \text{on } \Gamma_f \cup \Gamma_2, \\ \partial_n\Psi & = & v, & \text{on } \Gamma_1. \end{array} \right. \tag{3.6}$$

If we define $\varphi = \psi|_{\Gamma_s}$, then the original problem can be transformed in a one-dimensional problem on Γ_s,

$$\left\{ \begin{array}{rcll} \ddot{\varphi} + \mathcal{A}\varphi & = & \mathcal{B}v, & \text{on } \Gamma_s \times [0, \tau], \\ \varphi(x, 0) & = & \varphi_0(x), & \\ \dot{\varphi}(x, 0) & = & \varphi_1(x), & \end{array} \right. \tag{3.7}$$

where the operators $\mathcal{A}$ and $\mathcal{B}$ are defined by

$$\mathcal{A}\varphi = \partial_n \mathrm{D}\varphi|_{\Gamma_s}, \quad \mathcal{B}v = -\partial_n \mathrm{N}v|_{\Gamma_s}. \tag{3.8}$$

It is easy to see that this abstract formulation is related to the original equations by the fact that the velocity potential ψ and the two functions φ and v verify

$$\psi = \mathrm{D}\varphi + \mathrm{N}v. \tag{3.9}$$

The application of elementary theorems for elliptic problems allows us to claim that $\mathcal{A}$ is a linear unbounded operator in $H^{-1/2}(\Gamma_s)$ with domain $H^{1/2}(\Gamma_s)$. If we apply the results of Grisvard (see [5]) for $\varphi \in H^{3/2}(\Gamma_s)$ and the additional compatibility conditions

$$x^{-1/2}\varphi'(x) \in L^2(\Gamma_s), \ (\pi - x)^{-1/2}\varphi'(x) \in L^2(\Gamma_s), \tag{3.10}$$

we have $\mathrm{D}\varphi \in H^2(\Omega)$ (we will note this space $H_c^{3/2}(\Gamma_s)$). Thus we can consider $\mathcal{A}$ as a linear unbounded operator in $H^{1/2}(\Gamma_s)$ with domain $H_c^{3/2}(\Gamma_s)$.

As far as the operator $\mathcal{B}$ is concerned, the application of classical trace theorems (see [11]) and some basic inequalities allow us to claim that $\mathcal{B}$ is a bounded operator

from $H^{-1/2}(\Gamma_1)$ into $H^{-1/2}(\Gamma_s)$. We have the same kind of compatibility requirement for the boundary conditions of (3.6) : for $v \in H^{1/2}(\Gamma_1)$ with the additional condition (we will use the notation $H_c^{1/2}(\Gamma_1)$ to denote this space)

$$(1-y)^{-1/2}v(y) \in L^2(\Gamma_1), \tag{3.11}$$

we have $\mathrm{N}v \in H^2(\Omega)$, and so $\mathcal{B}$ is also bounded from $H_c^{1/2}(\Gamma_1)$ into $H^{1/2}(\Gamma_s)$.

Results of interpolation theory (see [11]) allow us to finally obtain the following theorem.

THEOREM 3.1.

i. *The operator $\mathcal{A}$ is a linear unbounded operator in $L^2(\Gamma_s)$ with domain $H^1(\Gamma_s)$.*

ii. *The operator $\mathcal{B}$ is a linear bounded operator from $L^2(\Gamma_1)$ into $L^2(\Gamma_s)$.*

When the generator is rigid with a shape $f \in L^2(\Gamma_1)$, the control $u(t)$ is a scalar and the state equation (3.7) takes the form

$$\ddot{\varphi} + \mathcal{A}\varphi = \beta(x)u(t),$$

where $\beta = \mathcal{B}f$. In this case the control operator is of rank one and bounded. In the particular case of rotating plane generators, i.e., $f(y) = y$, we can show from its Fourier coefficients that $\beta \in H^{1/2-\varepsilon}(\Gamma_s)\ \forall\ \varepsilon > 0$.

For some reasons that will appear clearly in what follows, we will work with zero mean functions. We define the Hilbert space

$$H = \left\{ \varphi \in L^2(\Gamma_s) \;\middle|\; \int_{\Gamma_s} \varphi(x)\,dx = 0 \right\} \equiv \tilde{L}^2(\Gamma_s)$$

and the domain of operator $\mathcal{A}$

$$D(\mathcal{A}) = \left\{ \varphi \in H^1(\Gamma_s) \;\middle|\; \int_{\Gamma_s} \varphi(x)\,dx = 0 \right\} \equiv \tilde{H}^1(\Gamma_s).$$

The fact that $D(\mathcal{A})$ is actually $\tilde{H}^1(\Gamma_s)$ results from the fact that the results of interpolation theory we used before remain valid for quotiented spaces (see [11]).

We have the following results on $\mathcal{A}$.

PROPOSITION 3.2. *The operator $\mathcal{A} : D(\mathcal{A}) \subset H \to H$ is strictly positive, self-adjoint, and $R(\lambda I + \mathcal{A}) = H$ for $\lambda > 0$.*

Proof. We have

$$\langle \mathcal{A}\varphi\,,\,\varphi\rangle = \|\nabla \mathrm{D}\varphi\|^2_{L^2(\Omega)} \geq 0,$$

where $\varphi \in D(\mathcal{A})$. Moreover $\mathcal{A}\xi = 0 \Rightarrow \xi = 0$ because of the particular choice of $D(\mathcal{A})$. This shows that $\mathcal{A}$ is strictly positive (the symmetry of $\mathcal{A}$ easily results from its definition).

For $\lambda > 0$, the operator $\lambda I + \mathcal{A}$ is trivially a bijection from $\tilde{H}^{1/2}(\Gamma_s)$ to $\tilde{H}^{-1/2}(\Gamma_s)$, and with the results of [5], we show that $\lambda I + \mathcal{A}$ is also a bijection from $\tilde{H}_c^{3/2}(\Gamma_s)$ to $\tilde{H}^{1/2}(\Gamma_s)$. We obtain the final result by applying interpolation theory (see [11]). □

This result ensures that the operator $\mathcal{A}^{1/2}$ is also well defined and we have

$$D(\mathcal{A}^{1/2}) = [D(\mathcal{A}), H]_{1/2} = \tilde{H}^{1/2}(\Gamma_s).$$

Unfortunately we don't have an explicit representation of $\mathcal{A}^{1/2}$, but we have for φ and w in $\tilde{H}^{1/2}(\Gamma_s)$

$$\langle \mathcal{A}\varphi\,,\,w\rangle_{\tilde{H}^{-1/2},\tilde{H}^{1/2}} = \left\langle \mathcal{A}^{1/2}\varphi\,,\,\mathcal{A}^{1/2}w \right\rangle = \int_\Omega \nabla \mathrm{D}\varphi.\nabla \mathrm{D}w.$$

3.2. Spectral analysis. The eigenvalues and associated eigenfunctions of $\mathcal{A}$, i.e., the nonzero functions $w_k(x) \in D(\mathcal{A})$ and the numbers λ_k such that $\mathcal{A}\, w_k = \lambda_k\, w_k$ for $k > 0$ integer, are obtained by solving the *Steklov problem*

$$\left\{\begin{array}{rcll} \Delta W_k &=& 0, & \text{in } \Omega, \\ \partial_n W_k &=& \lambda_k W_k, & \text{on } \Gamma_s, \\ \partial_n W_k &=& 0, & \text{on } \Gamma_f \cup \Gamma_1 \cup \Gamma_2, \end{array}\right. \tag{3.12}$$

where $w_k = W_k|_{\Gamma_s}$. By separation of variables one easily obtains the following form for W_k:

$$W_k = \alpha \cosh ky \cos kx,$$

where $\alpha \neq 0$ is an arbitrary constant. Thus we can choose

$$w_k(x) = \cos kx,\ x \in [0, \pi],\ k = 1, 2, \ldots,$$

and the eigenvalues λ_k are given by

$$\lambda_k = k \tanh k,\ k = 1, 2, \ldots.$$

In what follows, we will also consider the pair $(w_0 = 1, \lambda_0 = 0)$ which is not an eigenpair of $\mathcal{A}$ due to the choice of $D(\mathcal{A})$.

3.3. Formulation as a first-order system. We adopt the following abstract formulation of the original system:

$$\left\{\begin{array}{rcl} \dot{\xi} &=& A\xi + Bv, \\ \eta &=& C\xi, \\ \xi(0) &=& \xi_0 \in X. \end{array}\right. \tag{3.13}$$

The variable ξ is related to the original variables $\varphi = \psi|_{\Gamma_s}$ and $\dot{\varphi}$ in (3.7) by

$$\xi = \begin{pmatrix} \Pi_H \varphi \\ \Pi_H \dot{\varphi} \\ \Pi_{\mathbf{R}} \dot{\varphi} \end{pmatrix}, \tag{3.14}$$

where Π_H denotes the orthogonal projection on H and $\Pi_{\mathbf{R}}$ the orthogonal projection on the space of constant functions of x, i.e.,

$$\Pi_{\mathbf{R}} \varphi = \frac{1}{\pi} \int_0^\pi \varphi\, dx, \quad \Pi_H \varphi = \varphi - \Pi_{\mathbf{R}} \varphi.$$

The state space X is the following:

$$\begin{aligned} X &= D(\mathcal{A}^{1/2}) \times H \times \mathbf{R} \\ &= \tilde{H}^{1/2}(\Gamma_s) \times \tilde{L}^2(\Gamma_s) \times \mathbf{R}. \end{aligned}$$

The operators A, B, and C are defined as follows:

$$A\xi = \begin{pmatrix} \xi_2 \\ -\mathcal{A}\xi_1 \\ 0 \end{pmatrix},\ Bv = \begin{pmatrix} 0 \\ \Pi_H \mathcal{B} v \\ \Pi_{\mathbf{R}} \mathcal{B} v \end{pmatrix}, \tag{3.15}$$

$$C\xi = -(\xi_2 + \xi_3), \tag{3.16}$$

and the domain of A is given by

$$\begin{aligned} D(A) &= D(\mathcal{A}) \times D(\mathcal{A}^{1/2}) \times \mathbf{R} \\ &= \tilde{H}^1(\Gamma_s) \times \tilde{H}^{1/2}(\Gamma_s) \times \mathbf{R}. \end{aligned}$$

We define the inner product in X by

$$\begin{aligned} \langle \xi \, , \, \zeta \rangle_X &= \left\langle \mathcal{A}^{1/2}\xi_1 \, , \, \mathcal{A}^{1/2}\zeta_1 \right\rangle + \langle \xi_2 + \xi_3 \, , \, \zeta_2 + \zeta_3 \rangle \\ &= \int_\Omega \nabla \mathrm{D}\xi_1 . \nabla \mathrm{D}\eta_1 + \int_{\Gamma_s} (\xi_2 + \xi_3)(\zeta_2 + \zeta_3). \end{aligned} \tag{3.17}$$

We adopt the notation $\|.\|_X$ for the associated norm, which is defined by

$$\begin{aligned} \|\xi\|_X^2 &= |\mathcal{A}^{1/2}\xi_1|^2 + |\xi_2 + \xi_3|^2 \\ &= \int_\Omega |\nabla \mathrm{D}\xi_1|^2 + \int_{\Gamma_s} |\xi_2 + \xi_3|^2. \end{aligned}$$

This norm is related in some sense to the natural energy, in terms of the original potential ψ,

$$E(\psi, \dot{\psi}) = \frac{1}{2}\int_\Omega |\nabla \psi|^2 + \frac{1}{2}\int_{\Gamma_s} \dot{\psi}^2,$$

which is equal to the sum of kinetic and potential energies of the fluid.

Indeed, if we consider the definition of ξ given by (3.14) and the relation $\psi = \mathrm{D}\varphi + \mathrm{N}v$, then we can show that

$$\int_\Omega |\nabla \psi|^2 = \int_\Omega |\nabla \mathrm{D}\xi_1|^2 + \int_\Omega |\nabla \mathrm{N}v|^2.$$

Thus we will have $\frac{1}{2}\|\xi\|_X^2 = E(\psi, \dot{\psi})$, only if $v = 0$.

The main advantage of our abstract formulation is that the chosen energy is coercive on X, i.e., the only element such that $\|\xi\|_X^2 = 0$ is $\xi = 0$, whereas the natural energy is not coercive on the original energy space, since all solutions of the type

$$\psi = c, \ t \in [0, \infty)$$

verify $E(\psi, \dot{\psi}) = 0$.

3.4. Spectral analysis and semigroup generation. One can show that the eigenpairs $(\mu_k, \phi_k)_{k \in \mathbf{Z}}$ of operator A are given by $\mu_k = i\omega_k$, where $\omega_k = \sqrt{\lambda_k}$ for $k > 0$, $\omega_k = -\sqrt{\lambda_{-k}}$ for $k < 0$, and

$$\phi_k = \frac{1}{\sqrt{\pi}} \begin{pmatrix} \frac{1}{\mu_k} w_k \\ w_k \\ 0 \end{pmatrix}$$

for $k \in \mathbf{Z}^*$ and $\mu_0 = 0$,

$$\phi_0 = \begin{pmatrix} 0 \\ 0 \\ \frac{1}{\sqrt{\pi}} \end{pmatrix}.$$

The family ϕ_k can be shown to be an orthonormal basis of X, with $\|\phi_k\|_X = 1$. This means that A is a Riesz-spectral operator (for a complete theory of Riesz-spectral systems see [3]).

Remark 3.3. This property of spectrality would have been lost if we had considered the system $\dot{\xi} = A_0\xi$ with $\xi = (\dot{\varphi}, \varphi) \in H^{1/2}(\Gamma_s) \times L^2(\Gamma_s)$, where

$$A_0 = \begin{pmatrix} 0 & I \\ -\mathcal{A} & 0 \end{pmatrix},$$

with domain $D(A_0) = H^1(\Gamma_s) \times H^{1/2}(\Gamma_s)$, since in this case the eigenfunctions of A_0 do not form a basis of $H^{1/2}(\Gamma_s) \times L^2(\Gamma_s)$. Another possibility could have been to consider $\tilde{A}$ with the domain $\tilde{H}^1(\Gamma_s) \times \tilde{H}^{1/2}(\Gamma_s)$, but in this case, the mean value of $\dot{\varphi}$ is constrained to be zero, in spite of its contribution to the surface elevation η. In the system theoretic terminology, this corresponds to remove an observable part of the state ξ, which results in an alteration of the input-output map (from the control v to the surface elevation η). Thus, this other choice is not admissible either.

We have the following proposition.

PROPOSITION 3.4. *The operator A is the infinitesimal generator of a strongly continuous semigroup of contractions $T(t)$ on $X = \tilde{H}^{1/2}(\Gamma_s) \times \tilde{L}^2(\Gamma_s) \times \mathbf{R}$, given by the formula*

$$T(t)\xi = \sum_{k\in\mathbf{Z}} e^{i\omega_k t} \langle \xi\,,\, \phi_k\rangle_X \,\phi_k,$$

where $\omega_k^2 = \lambda_k$.

Proof. Since A is skew-adjoint for the inner product $\langle .\,,\, .\rangle_X$, i.e., $A^* = -A$, we have for $\xi \in D(A)$

$$\mathrm{Re}\ \langle A\xi\,,\, \xi\rangle_X = 0,$$

and the semigroup generation easily follows from the fact that A is Riesz-spectral (see [3, Theorem 2.3.5]). □

In the next two paragraphs we focus on controllability problems.

4. Controllability problems. We focus on the theoretical study of the following problem: given a control space U and a time $\tau > 0$, is it possible to find a control $v \in L^2(0,\tau;U)$ such that the surface of the canal is in a given state at time τ? We will consider the strong version of this question, where the state is to be reached exactly, and the weak version, where the state is reached approximately. The first version is an exact controllability problem. We will see that exact controllability can eventually be obtained if we consider flexible generators. In the case where the generators are rigid, only approximate controllability can eventually be obtained.

Let us first recall some definitions. Consider the system

$$\left\{ \begin{array}{rcl} \dot{\xi} & = & A\xi + Bv, \\ \xi(0) & = & \xi_0, \end{array} \right. \tag{4.1}$$

where A is an infinitesimal generator of a strongly continuous semigroup $T(t)$ on a Hilbert space X, and B is a bounded operator from U to X. At a given time t, the control $v(t)$ belongs to an infinite or finite dimensional space U.

DEFINITION 4.1. *The controllability map of system* (4.1) *on* $[0,\tau]$ *(for a finite* $\tau > 0$*) is the bounded map* $\mathcal{H}_\tau : L_2(0,\tau;U) \to X$ *defined by*

$$\mathcal{H}_\tau v = \int_0^\tau T(\tau - s)Bv(s)\,ds.$$

The natural definition of exact controllability follows.

DEFINITION 4.2. *The system* (4.1) *is exactly controllable on* $[0,\tau]$ *if every element of* X *can be reached from the origin at time* τ*, equivalently*

$$ran\ \mathcal{H}_\tau = X.$$

There are various methods to obtain this result. The classical approach works by showing an inverse inequality [3], [10].

THEOREM 4.3. *The system* (4.1) *is exactly controllable on* $[0,\tau]$ *if there exists* $\gamma > 0$ *such that*

$$\int_0^\tau \|B^*T^*(s)\xi\|_U^2\, ds \geq \gamma\|\xi\|_X^2 \tag{4.2}$$

for every $\xi \in X$.

When U is a finite dimensional space and B is bounded, exact controllability cannot hold, because in this case the map $\mathcal{H}_\tau$ is compact (see [3]).

The concept of exact controllability is often too strong, and sometimes the concept of approximate controllability is more adequate. It is defined as follows.

DEFINITION 4.4. *The system* (4.1) *is approximately controllable on* $[0,\tau]$ *(for a finite* $\tau > 0$*) if for* $\varepsilon > 0$*, it is possible to steer from the origin at a distance* ε *from all elements of* X *in a finite time* τ*, say*

$$\overline{ran\ \mathcal{H}_\tau} = X. \tag{4.3}$$

To show (4.3) one usually tries to show that H_τ^* is one-to-one [3, Theorem 4.1.7]:

$$B^*T(t)^*\xi = 0 \text{ on } [0,\tau] \Rightarrow \xi = 0. \tag{4.4}$$

In our application, we are first going to focus on the case where the generators are flexible.

4.1. Controllability with a flexible generator. In [14] and [6], we tried to attack the exact controllability problem by separating it in two subproblems. The first problem treated the same system with a distributed control on Γ_s, which was shown to be exactly controllable for any $\tau > 0$. The second problem is an elliptic problem, which was not exactly controllable, but only approximately (see Lemma 4.6).

This just allowed us to prove the approximate controllability of the original system in a rather complicated way. In this section we show that in fact exact controllability does not hold. We also give a simpler proof for the approximate controllability result.

In the following we will need to use the operator $\mathcal{B}^*$, the adjoint of $\mathcal{B}$. Let us recall that for $v \in L^2(\Gamma_1)$, $\mathcal{B}v$ is defined by $\mathcal{B}v = -\partial_n \mathrm{N}v|_{\Gamma_s}$, where N is the Neumann map defined by (3.6). In order to define $\mathcal{B}^*z$ for some $z \in L^2(\Gamma_s)$, let us take the auxiliary function Θ solution of

$$\left\{\begin{array}{rcll} \Delta\Theta & = & 0, & \text{in } \Omega, \\ \Theta & = & z, & \text{on } \Gamma_s, \\ \partial_n\Theta & = & 0, & \text{on } \Gamma_f \cup \Gamma_1 \cup \Gamma_2. \end{array}\right. \tag{4.5}$$

We write that $\langle \Delta \mathrm{N}v\,,\, \Theta\rangle_{L^2(\Omega)} = 0$, and applying Green's formula twice, we obtain that $\mathcal{B}^*z = \Theta|_{\Gamma_1}$, and if we consider the definition of the Dirichlet map given by (3.5), we have in fact

$$\mathcal{B}^*z = \mathrm{D}z|_{\Gamma_1}\,.$$

4.1.1. Counterexample to the exact controllability on $[0,\tau]$. We have the following theorem.

THEOREM 4.5. *Consider the system*

$$\dot{\xi} = A\xi + Bv, \tag{4.6}$$

where the definition of A and B has been given in section 3.3*, and $v \in L^2(0,\tau;L^2(\Gamma_1))$. Exact controllability on $[0,\tau]$ does not hold in X for any $\tau > 0$.*

Proof. The classical approach consists in using Theorem 4.3. We first need to identify B^* : using the definition of B given by (3.15) and the definition of the inner product $\langle .\,,\,. \rangle_X$ given by (3.17), we have for some $\xi \in X$ and some $v \in L^2(\Gamma_1)$

$$\begin{aligned}
\langle Bv\,,\,\xi\rangle_X &= \left\langle \begin{pmatrix} 0 \\ \Pi_H \mathcal{B}v \\ \Pi_{\mathbf{R}}\mathcal{B}v \end{pmatrix}\,,\, \begin{pmatrix} \xi_1 \\ \xi_2 \\ \xi_3 \end{pmatrix} \right\rangle_X \\
&= \langle \Pi_H \mathcal{B}v + \Pi_{\mathbf{R}}\mathcal{B}v\,,\,\xi_2+\xi_3\rangle \\
&= \langle \mathcal{B}v\,,\,\xi_2+\xi_3\rangle \\
&= \langle v\,,\,\mathcal{B}^*(\xi_2+\xi_3)\rangle_{L^2(\Gamma_1)}.
\end{aligned}$$

Thus the operator $B^* : L^2(\Gamma_s) \to U = L^2(\Gamma_1)$ is given by

$$B^*\xi = \mathcal{B}^*(\xi_2+\xi_3)$$

for $\xi \in X$. One considers the sequence

$$\xi^{(n)} = \phi_n,\ n > 0,$$

where ϕ_n is the nth eigenfunction of A. This sequence has the property $\|\xi^{(n)}\|_X^2 = 1$, and one can easily show that

$$\begin{aligned}
B^*T^*(t)\xi^{(n)} &= \sum_{k\in\mathbf{Z}} e^{-i\omega_k t}\left\langle \xi^{(n)}\,,\,\phi_k\right\rangle_X B^*\phi_k \\
&= B^*\phi_n e^{-i\omega_n t} \\
&= \frac{1}{\sqrt{\pi}}\mathcal{B}^* w_n e^{-i\omega_n t},
\end{aligned}$$

where w_n is the nth eigenfunction of $\mathcal{A}$. Using the definition of $\mathcal{B}^*$, one can express $\mathcal{B}^*w_n$ by using the problem (3.12) that has been used to obtain the eigenpairs of $\mathcal{A}$: this gives

$$\mathcal{B}^* w_n = \frac{\cosh ny}{\cosh n}. \tag{4.7}$$

It is easy to see that $\mathcal{B}^*w_n \to 0$ almost everywhere (a.e.) in $L^2(\Gamma_1)$, and a simple calculation applied to (4.7) gives

$$\|\mathcal{B}^*w_n\|^2_{L^2(\Gamma_1)} = \mathcal{O}(n^{-1}) \ \text{ as } n \to \infty.$$

Hence for fixed $\tau > 0$ one has

$$\lim_{n\to\infty} \frac{\int_0^\tau \|B^*T^*(s)\xi^{(n)}\|_U^2\,ds}{\|\xi^{(n)}\|_X^2} = 0.$$

So there does not exist $\gamma > 0$ such that (4.2) is verified. ∎

It is easy to see that the result $\mathcal{B}^* w_n \to 0$ would have been lost if we had considered convergence in $H^1(\Gamma_1)$. This could give some directions of search to identify the exactly controllable subspace. Similar problems have been addressed in [13].

4.1.2. Approximate controllability on $[0,\tau]$. We first show the following lemma.

LEMMA 4.6. *The range of $\mathcal{B}$ is dense in $L^2(\Gamma_s)$, or equivalently,*

$$\mathcal{B}^* z = 0 \Rightarrow z = 0, \text{ in } L^2(\Gamma_1).$$

Proof. Let us take $z \in L^2(\Gamma_s)$ such that $\mathcal{B}^* z = 0$. From (4.5), we have that

$$\left\{ \begin{array}{rcll} \Delta\Theta & = & 0, & \text{in } \Omega, \\ \Theta & = & z, & \text{on } \Gamma_s, \\ \Theta & = & 0, & \text{on } \Gamma_1, \\ \partial_n\Theta & = & 0, & \text{on } \Gamma_f \cup \Gamma_1 \cup \Gamma_2. \end{array} \right. \tag{4.8}$$

Since Θ is smooth enough (we have taken z in $L^2(\Gamma_s)$), the Holmgren's uniqueness theorem implies that $\Theta = 0$, since $\Theta = \partial_n\Theta = 0$ on Γ_1. So $z = \Theta|_{\Gamma_s} = 0$. This ends the proof. □

We are now able to show the following theorem.

THEOREM 4.7. *The system*

$$\dot{\xi} = A\xi + Bv, \tag{4.9}$$

with $v \in L^2(0,\tau; L^2(\Gamma_1))$, is approximately controllable on $[0,\tau]$ for any $\tau > 0$.

Proof. To show approximate controllability one has to show that the operator $\mathcal{H}_\tau^*$ is one-to-one, i.e.,

$$B^* T^*(s)\xi_0 = 0, \ s \in [0,\tau] \Rightarrow \xi_0 = 0. \tag{4.10}$$

Since $T^*(s) = T(-s)$, this is the same as showing that the solution of system

$$\left\{ \begin{array}{rcll} -\dot{\xi} & = & A\xi, \ t \in [0,\tau], \\ B^*\xi & = & 0, \\ \xi(0) & = & \xi_0 \in X, \end{array} \right. \tag{4.11}$$

with the supplementary condition

$$B^*\xi = 0, \ t \in [0,\tau], \tag{4.12}$$

is identically zero, i.e., $\xi_0 = 0$.

In order to clarify the abstract formulation, the system (4.11)–(4.12) can be rewritten in terms of the original equations (3.2)–(2.6) under the following form:

$$\left\{ \begin{array}{rcll} \Delta\psi & = & 0, \text{ in } \Omega \times [0,\tau], \\ \ddot{\psi} + \partial_n\psi & = & 0, \text{ on } \Gamma_s \times [0,\tau], \\ \partial_n\psi & = & 0, \text{ on } \Gamma_f \cup \Gamma_1 \cup \Gamma_2 \times [0,\tau], \end{array} \right. \tag{4.13}$$

with initial conditions

$$\psi(0) = \varphi_0 \in \tilde{H}^{1/2}(\Gamma_s), \ \dot{\psi}(0) = \varphi_1 \in L^2(\Gamma_s), \text{ on } \Gamma_s,$$

and the supplementary condition

$$\dot{\psi} = 0, \text{ on } \Gamma_1 \times [0, \tau].$$

We recall that ξ and ψ are related by $\varphi = \psi|_{\Gamma_s}$ and

$$\xi = \begin{pmatrix} \Pi_H \varphi \\ \Pi_H \dot{\varphi} \\ \Pi_{\mathbf{R}} \dot{\varphi} \end{pmatrix}, \ \xi_0 = \begin{pmatrix} \Pi_H \varphi_0 \\ \Pi_H \varphi_1 \\ \Pi_{\mathbf{R}} \varphi_1 \end{pmatrix}.$$

For the rest of the proof, we will use the abstract formulation (4.11)–(4.12). We first check what kind of information on ξ is given by (4.12), which is equivalent to

$$\mathcal{B}^*(\xi_2 + \xi_3) = 0.$$

By Lemma 4.6 this implies that $\xi_2 + \xi_3 = 0$. Since ξ_2 and ξ_3 belong to orthogonal subspaces (ξ_2 has zero mean and ξ_3 is a constant), we must have $\xi_2 = 0$ and $\xi_3 = 0$. From (4.11) we have that $-\dot{\xi}_1 = \xi_2 = 0$, so (4.11) reduces to

$$\mathcal{A}\xi_1 = 0, \ t \in [0, \tau],$$

with ξ_1 only depending of x. Since the kernel of $\mathcal{A}$ is reduced to $\{0\}$, this implies that $\xi_1 = 0 \ \forall t \in [0, \tau]$, and we can conclude that $\xi = 0 \ \forall t \in [0, \tau]$. □

As it is proved in [12, p. 522, Appendix B], the following theorem is an immediate corollary of Theorem 4.7 (the result in [12] is very general and can be directly applied to our case).

THEOREM 4.8. *Let E be any finite dimensional subspace of X. For any $\tau > 0$, for any $\varepsilon > 0$, and for any $\xi_0, \xi_d \in X$, there exists a control $v \in L^2(0, \tau; L^2(\Gamma_1))$ such that the solution of*

$$\left\{ \begin{array}{rcl} \dot{\xi} & = & A\xi + Bv, \ t \in [0, \tau], \\ \xi(0) & = & \xi_0 \end{array} \right.$$

satisfies

$$\left\{ \begin{array}{rcl} \Pi_E(\xi(\tau) - \xi_d) & = & 0, \\ ||\xi(\tau) - \xi_d||_X & < & \varepsilon, \end{array} \right.$$

where $\Pi_E \xi$ denotes the orthogonal projection of ξ over E.

4.2. Controllability with a rigid generator. We consider again the first order form of our system, but this time with a scalar control $u(t)$,

$$\dot{\xi} = A\xi + Bu. \tag{4.14}$$

We take this new definition for B :

$$Bu = bu, \quad b = \begin{pmatrix} 0 \\ \Pi_H \beta \\ \Pi_{\mathbf{R}} \beta \end{pmatrix} \in X, \tag{4.15}$$

where, as before, $\beta = \mathcal{B}f$. In what follows we will consider the set of generators with a "strategic" shape function f. This set is defined as follows.

DEFINITION 4.9. *We say that a shape function $f \in L^2(\Gamma_1)$ is strategic if the corresponding operator B, defined by ways of* (3.8)–(3.6) *and* (4.15)*, is such that* $B^*\phi_k \neq 0 \ \forall\, k \in \mathbf{Z}$.

As it is shown in the following lemma, the positivity of the shape function f on $[0,1]$, which includes the case of $f(y) = y$ (plane rotating generators), is a sufficient condition for f to be strategic.

LEMMA 4.10. *If the shape function $f(y)$ is positive for $y \in [0,1]$, then*

$$B^*\phi_k > 0 \quad \forall k \in \mathbf{Z}.$$

Proof. One has for $k \geq 0$

$$\begin{aligned} B^*\phi_k &= \langle \phi_k \,,\, b \rangle_X \\ &= \frac{1}{\sqrt{\pi}} \langle w_k \,,\, \beta \rangle \,, \end{aligned}$$

and $B^*\phi_{-k} = B^*\phi_k$. From the definition of β we have

$$\begin{aligned} \langle w_k \,,\, \beta \rangle &= \langle w_k \,,\, \mathcal{B} f \rangle \\ &= \langle \mathcal{B}^* w_k \,,\, f \rangle_{L^2(\Gamma_1)} \,, \end{aligned}$$

where $\mathcal{B}^* w_k$ is given by (4.7). Thus we have

$$B^*\phi_k \quad = \quad \tfrac{1}{\sqrt{\pi}\cosh k} \int_0^1 \cosh ky \, f(y)\, dy, \ k \geq 0. \tag{4.16}$$

Since $f(y)$ is a positive function on $[0,1]$, then it is obvious from (4.16) that $B^*\phi_k > 0 \ \forall k$. ∎

Remark 4.11. The strategic set is not limited to positive shapes. For example, if we consider a plane rotating generator with an axis located at $y = a$, $a \in [0,1]$, we can verify by means of (4.16) that the corresponding shape $f_a(y) = y - a$ will be strategic for $a \in [0,1] \setminus \left(\left\{ \tfrac{1}{2} \right\} \cup \left\{ \ 1 - \tfrac{1}{k} \tanh \tfrac{k}{2}, \ k \in \mathbf{N}^* \right\} \right)$.

In the following, we will need to know how f influences the regularity of β. For this purpose we define for an integer $p > 0$ the space

$$H_c^p(\Gamma_1) = \{ f \in H^p(\Gamma_1), \ f^{(q)}(1) = 0, \ q = 0 \ldots p-1 \},$$

and we take the convention $H_c^0(\Gamma_1) = L^2(\Gamma_1)$. We have the following lemma.

LEMMA 4.12. *If $f \in H_c^p(\Gamma_1)$, then there exists $C > 0$ such that the operator B has the property*

$$|B^*\phi_k| \leq \frac{C}{k^{p+\frac{1}{2}}}.$$

If $f \in H^p(\Gamma_1) \backslash H_c^p(\Gamma_1)$ for some $p > 0$ then there exists $C > 0$ and an integer q with $0 < q \leq p$ such that

$$k^q |B^*\phi_k| \to C$$

as $k \to \infty$.

Proof. We obtain the two results by means of repeated integration by parts on (4.16) and the Cauchy–Schwarz inequality. ∎

4.2.1. Lack of approximate controllability on an interval $[0,\tau]$**.** In fact the system (4.14) is not approximately controllable on $[0,\tau]$ for any $\tau>0$. This result was claimed in [6], but the proof was incomplete. The following lemma is crucial in the proof of our main result.

LEMMA 4.13.

1. *For any* $\tau>0$*, the system* $\left\{e^{iw_kt}\right\}_{k\in\mathbf{Z}}$ *is complete and linked in* $L^2(0,\tau)$.
2. *For any* $\tau>0$*, there exists a subset* $\mathcal{S}_\tau\subset\mathbf{Z}$ *such that* $\left\{e^{iw_kt}\right\}_{k\in\mathcal{S}_\tau}$ *is a Riesz basis of* $L^2(0,\tau)$.

For $k>0$, ω_k is given by $\omega_k=\sqrt{\lambda_k}$, where $\lambda_k>0$ are the eigenvalues of $\mathcal{A}$, given by $\lambda_k=k\tanh k$. Notice that we have taken $\omega_k=-\omega_{-k}$ for $k<0$ and $\omega_0=0$. We recall that the terminology "linked system" means that every element of the system is in the closed subspace spanned by the other elements.

Before giving the proof, let us recall the following theorem which is due to Young [18].

THEOREM 4.14 ($\frac{1}{4}$-theorem). *If* $\{\omega_k\}$ *is a sequence of real numbers for which*

$$|\omega_k-k|<\frac{1}{4},\ k\in\mathbf{Z},$$

then the system $\{e^{i\omega_kt}\}$ *is a (Riesz) basis of* $L^2(0,2\pi)$.

This result can be applied in $L^2(0,\tau)$, in which case the inequality to obtain is

$$\left|\omega_k-\frac{2k\pi}{\tau}\right|<\frac{\pi}{2\tau}.$$

We also need the following theorem.

THEOREM 4.15. *Let* $\{f_n\}_{n\in\mathbf{Z}}$ *be a Riesz basis of* $L^2(0,\tau)$. *Consider the family* $\{\tilde f_n\}$ *where* $\tilde f_n=g_n$ *for* $n<n_0$ *and* $\tilde f_n=f_n$ *otherwise, where the* g_n *are such that* $\{\tilde f_n\}$ *remains a basis of* $L^2(0,\tau)$. *Then* $\{\tilde f_n\}$ *remains a Riesz basis of* $L^2(0,\tau)$.

Proof. Since $\{f_n\}_{n\in\mathbf{Z}}$ is a Riesz basis of $L^2(0,\tau)$, there exists an isomorphism T and an orthonormal basis $\{e_n\}$ of $L^2(0,\tau)$ such that $Te_n=f_n,\ n\in\mathbf{Z}$. Since T is an isomorphism there exists $\rho>0$ such that $\|Tf\|\geq\rho\|f\|\ \forall\, f\in L^2(0,\tau)$. Then we have

$$\infty>\sum\|f_n-\tilde f_n\|^2=\sum\|Te_n-\tilde f_n\|^2\geq\rho^2\sum\|e_n-T^{-1}\tilde f_n\|^2,$$

showing that $\{e_n\}$ and $\{T^{-1}\tilde f_n\}$ are quadratically close. Then by Bari's theorem (see [18], p. 45), since $\{T^{-1}\tilde f_n\}$ is a basis, it is also a Riesz basis, and so is $\{\tilde f_n\}$ since T^{-1} is an isomorphism. □

Proof of Lemma 4.13. We take the sequence of integers $\{c_k\}$ with

$$c_k=\operatorname{Int}\left(\frac{4k^2\pi^2}{\tau^2}\right),k\geq0,$$

(the notation Int (.) denotes the integer part function) and we define $\mu_k=\sqrt{c_k}$ for $k\geq0$ and $\mu_k=-\sqrt{c_k}$ for $k<0$. It can be easily shown that this sequence has the property

$$\left|\mu_k-\frac{2k\pi}{\tau}\right|<\frac{1}{2|\mu_k|}\quad\forall k\in\mathbf{Z},$$

which shows that for a sufficiently large k, say k_0, one has $\left|\mu_k-\frac{2k\pi}{\tau}\right|<\frac{\pi}{2\tau}\ \forall|k|\geq k_0$, since $\{|\mu_k|\}$ is an increasing sequence. Since $\omega_p=\sqrt{p\tanh p}$, one has

$$\frac{\omega_p}{\sqrt{p}}\to1,\quad\text{as } p\to\infty,$$

which means that there exists $k_1 \geq k_0$ such that

$$\left|\omega_{c_k} - \frac{2k\pi}{\tau}\right| < \frac{\pi}{2\tau},\ \forall |k| \geq k_1.$$

Let us define the sequence $\{\mu_k\}_{k\in\mathbf{Z}}$ by

$$\mu_k = \begin{cases} \frac{2k\pi}{\tau}\ , & |k| < k_1, \\ \omega_{c_k}\ , & k \geq k_1, \\ -\omega_{c_k}\ , & k \leq -k_1. \end{cases}$$

One has by construction $|\mu_k - \frac{2k\pi}{\tau}| < \frac{\pi}{2\tau}\ \forall k \in \mathbf{Z}$, so by Theorem 4.14 the system $\{e^{i\mu_k t}\}$ is a Riesz basis of $L^2(0,\tau)$. If one replaces a finite number of μ_k by other numbers distinct from μ_k [16, Theorem 7, p. 135] the system remains a basis, and from Theorem 4.15 it is also a Riesz basis.

So one replaces μ_k by ω_{d_k} for $|k| < k_1$, where each d_k, $|k| < k_1$ are distinct and can be chosen such that $d_k \neq c_l$ for any l with $|l| \geq k_1$. We then define the set

$$\mathcal{S}_\tau = \{d_k, |k| < k_1\} \cup \{c_k, |k| \geq k_1\}.$$

By construction, the system $\{e^{i\omega_k t}\}_{k\in\mathcal{S}_\tau}$ is a Riesz basis. Hence, the system $\{e^{i\omega_k t}\}_{k\in\mathbf{Z}}$ cannot be linearly independent in $L^2(0,\tau)$ since it strictly contains a basis, and by a result of Schwartz [16, Theorem 4, p. 102], this system is linked. □

As announced above, we have the following controllability theorem.

THEOREM 4.16. *Suppose that the shape function $f \in H^p(\Gamma_1)\backslash H^p_c(\Gamma_1)$ is strategic. Then, the system* (4.14)–(4.15) *is not approximately controllable on* $[0,\tau]$ *for any* $\tau > 0$.

Proof. Since $f \in H^p(\Gamma_1)\backslash H^p_c(\Gamma_1)$ there exists an integer q and a constant $C > 0$ such that $k^q |B^*\phi_k| \to C$. Moreover, we have shown in Lemma 4.13 that the system $\{e^{i\omega_k t}\}_{k\in\mathcal{S}_\tau}$ is a Riesz basis of $L^2(0,\tau)$ and the set $\mathbf{Z}\backslash\mathcal{S}_\tau$ is not finite. Hence we can apply a result of Avdonin and Ivanov (see [1, Theorem II.6.6, p. 141], to claim that there exists a nonzero sequence $\{a_k\}_{k\in\mathbf{Z}}$ such that $\sum_{k\in\mathbf{Z}} |a_k|^2 |\omega_k|^{4q} < \infty$ and

$$\sum_{k\in\mathbf{Z}} a_k e^{i\omega_k t} = 0,\ \text{in } L^2(0,\tau).$$

Since $\omega_k = (k \tanh k)^{\frac{1}{2}}$ is equivalent to $k^{\frac{1}{2}}$, the series $\sum_{k\in\mathbf{Z}} |a_k|^2 |k|^{2q}$ is convergent. Hence, if we consider the vector ξ defined by

$$\xi = \sum_{k\in\mathbf{Z}} \frac{a_k}{B^*\phi_k}\phi_k,$$

we have $\xi \in X$ since the series

$$\sum_{k\in\mathbf{Z}} |\langle \xi\ ,\ \phi_k\rangle|^2 = \sum_{k\in\mathbf{Z}} \left|\frac{a_k}{B^*\phi_k}\right|^2$$

is convergent. Finally, we have by construction

$$B^*T^*(t)\xi = \sum_{k\in\mathbf{Z}} B^*\phi_k\, \langle \xi\ ,\ \phi_k\rangle\, e^{i\omega_k t} = 0,\ \text{in } L^2(0,\tau). \tag{4.17}$$

This ends the proof. □

This result is essentially due to the fact that there is no asymptotic gap between ω_k and ω_{k+1}, i.e., one has $\lim_{k\to\infty} |\omega_{k+1} - \omega_k| = 0$. Nevertheless, the hypothesis of "maximal regularity" for f is fundamental because the result of Avdonin and Ivanov cannot be applied if $\{a_k\}_{k\in\mathbf{Z}}$ is expected to have an exponential decay (see Theorem II.6.5, page 140 in [1]). Such a decay would be required if $|B^*\phi_k|$ already had an exponential decay. Lemma 4.12 shows that if $f \in H_c^p(\Gamma_1)$ for any $p > 0$, then the decay rate of $|B^*\phi_k|$ will be faster than any polynomial, but it is not clear whether this decay rate can be exponential or not.

In the next section we show that the system (4.14)–(4.15) is approximately controllable in the sense of a weaker definition.

4.2.2. Generic approximate controllability. One can define a generic concept of approximate controllability on $[0,\infty)$. This is the concept used by Curtain and Zwart [3]. Its definition follows.

DEFINITION 4.17. *Let us call* $\mathcal{R}$ *the reachable subspace*

$$\mathcal{R} = \bigcup_{\tau\in\mathbf{R}^+} Ran\,\mathcal{H}^\tau.$$

The system is approximately controllable on $[0,\infty)$ *if* $\overline{\mathcal{R}} = X$. □

As for the case where the controllability time is finite, this type of approximate controllability corresponds to an observability property on the dual system, i.e., approximate controllability on $[0,\infty)$ will hold if

$$B^*T^*(t)\xi = 0 \quad \forall t > 0 \Longrightarrow \xi = 0. \tag{4.18}$$

Since we have shown that A is a Riesz-spectral operator, we can use Theorem 4.2.3 in [3] which claims that (4.18) will hold if

$$B^*\phi_k \neq 0 \quad \forall k \in \mathbf{Z},$$

where ϕ_k, $k \in \mathbf{Z}$ are the eigenfunctions of operator A. The following proposition follows directly from Definition 4.9.

PROPOSITION 4.18. *Suppose that the shape function* f *is strategic. Then the system* (4.14)–(4.15) *is approximately controllable on* $[0,\infty)$.

4.3. Interpretation of controllability results. Since we have established some controllability results on an abstract formulation of the original problem, it is necessary to explain the meaning of these results in terms of the original system, posed in the bidimensional domain Ω,

$$\left\{\begin{array}{rcll} \Delta\psi &=& 0, & \text{in } \Omega\times[0,\tau],\\ \ddot{\psi}+\partial_n\psi &=& 0, & \text{on } \Gamma_s\times[0,\tau],\\ \partial_n\psi &=& v, & \text{on } \Gamma_1\times[0,\tau],\\ \partial_n\psi &=& 0, & \text{on } \Gamma_f\cup\Gamma_2\times[0,\tau]. \end{array}\right. \tag{4.19}$$

It must be well understood that the unknown of this system is not the pair $(\psi,\dot{\psi})$, but the pair $(\psi,\dot{\psi}|_{\Gamma_s})$, as it is clearly expressed in the natural energy norm

$$E(\psi,\dot{\psi}) = \frac{1}{2}\int_\Omega |\nabla\psi|^2 + \frac{1}{2}\int_{\Gamma_s}\dot{\psi}^2, \tag{4.20}$$

and we will take this into account in the following.

We have obtained an approximate controllability result for the abstract system, posed on the free surface Γ_s,

$$\left\{ \begin{array}{rcl} \dot{\xi} & = & A\xi + Bv, \\ \xi(0) & = & \xi_0, \end{array} \right. \tag{4.21}$$

where ξ is related to the original velocity potential ψ by

$$\xi = \begin{pmatrix} \Pi_H \psi|_{\Gamma_s} \\ \Pi_H \dot{\psi}|_{\Gamma_s} \\ \Pi_{\mathbf{R}} \dot{\psi}|_{\Gamma_s} \end{pmatrix}.$$

This abstract system is well posed, and for $v \in L^2(0,\tau;L^2(\Gamma_1))$ and $\xi_0 \in X$ we have $\xi \in C(0,\tau;X)$. This means that the pair $(\psi|_{\Gamma_s}, \dot{\psi}|_{\Gamma_s})$ is continuous in time for the norm of X while in general the pair $(\psi, \dot{\psi}|_{\Gamma_s})$ is not continuous for the energy norm since it would require that v is also continuous. We will see that it is not a problem in the context of approximate controllability.

Let us consider a target

$$(\psi_d, \eta_d) \in H^1(\Omega) \times L^2(\Gamma_s),$$

with $\Delta\psi_d = 0$, which is to be reached approximately in the sense of the energy norm, at $t = \tau$ by $(\psi, \dot{\psi}|_{\Gamma_s})$, i.e., we seek for a v such that

$$E(\psi(\tau) - \psi_d, \dot{\psi}(\tau)|_{\Gamma_s} - \eta_d) \leq \varepsilon. \tag{4.22}$$

In terms of the abstract system the target (ψ_d, η_d) corresponds to the abstract target

$$\xi_d = \begin{pmatrix} \Pi_H \psi_d|_{\Gamma_s} \\ \Pi_H \eta_d \\ \Pi_{\mathbf{R}} \eta_d \end{pmatrix}.$$

Hence, the meaning of Theorem 4.7 is that, given $\varepsilon > 0$, we can find a $v \in L^2(0,\tau;L^2(\Gamma_1))$ such that $\frac{1}{2}\|\xi(\tau) - \xi_d\|_X \leq \varepsilon$, i.e.,

$$\frac{1}{2} \left\| \begin{pmatrix} \Pi_H(\psi(\tau) - \psi_d)|_{\Gamma_s} \\ \Pi_H(\dot{\psi}(\tau)|_{\Gamma_s} - \eta_d) \\ \Pi_{\mathbf{R}}(\dot{\psi}(\tau)|_{\Gamma_s} - \eta_d) \end{pmatrix} \right\|_X \leq \varepsilon. \tag{4.23}$$

We have shown that the approximately reachable space for $\xi(\tau)$ is

$$\tilde{H}^{1/2}(\Gamma_s) \times \tilde{L}^2(\Gamma_s) \times \mathbf{R},$$

and thus

(4.24) *the approximately reachable space for* $\Pi_H \psi(\tau)|_{\Gamma_s}$ *is* $\tilde{H}^{1/2}(\Gamma_s)$,

and since the approximately reachable space for the pair $(\Pi_H \dot{\psi}(\tau)|_{\Gamma_s}, \Pi_{\mathbf{R}} \dot{\psi}(\tau)|_{\Gamma_s})$ is $\tilde{L}^2(\Gamma_s) \times \mathbf{R}$, and $\dot{\psi}(\tau)|_{\Gamma_s} = \Pi_H \dot{\psi}(\tau)|_{\Gamma_s} + \Pi_{\mathbf{R}} \dot{\psi}(\tau)|_{\Gamma_s}$, then

(4.25) *the approximately reachable space for* $\dot{\psi}(\tau)|_{\Gamma_s}$ *is* $L^2(\Gamma_s)$.

We can note that the final value of $\psi|_{\Gamma_s}$ is only reached up to an additive constant. This is related to the fact that in the whole domain Ω, the velocity potential ψ is itself defined up to an additive constant.

We must be very careful in extending the results (4.24) and (4.25) to the pair $(\psi, \dot{\psi}|_{\Gamma_s})$ in the whole domain Ω. Since we are only concerned with approximate controllability, by a density argument it is always possible to find

$$v \in \tilde{U} = \{u \in C(0,\tau; L^2(\Gamma_1)),\ u(\tau) = 0\},$$

such that the solution of (4.19) verifies (4.23). Moreover, when $v \in \tilde{U}$ then the energy norm and the abstract norm of the pair $(\psi(\tau), \dot{\psi}(\tau)|_{\Gamma_s})$ coincide, i.e.,

$$v \in \tilde{U} \Rightarrow \frac{1}{2}\|\xi(\tau) - \xi_d\|_X = E(\psi(\tau) - \psi_d, \dot{\psi}(\tau)|_{\Gamma_s} - \eta_d).$$

Thus, we can finally claim that when $v \in \tilde{U}$, the approximately reachable space for $\psi(\tau)$ is the set

$$\left\{\psi_d \in H^1(\Omega) \text{ such that } \Delta\psi_d = 0 \text{ in } \Omega,\ \partial_n\psi_d|_{\Gamma_1\cup\Gamma_2\cup\Gamma_f} = 0,\ \int_{\Gamma_s} \psi_d = 0\right\}.$$

5. Stabilization. As in the previous sections, we take $g_0 = 1$, $L = \pi$, $h = 1$, and we will consider only one generator at the left of the canal. We consider a "rigid" generator: we recall that the corresponding boundary condition is

$$\partial_n\psi = f(y)u, \quad \text{on } \Gamma_1,$$

where u is the velocity of the generator and $f \in L^2(\Gamma_1)$ is the "shape" of the generator. For the moment we will only require that f is a positive function on Γ_1.

It makes sense to measure the elevation of the surface at $x = 0$, i.e.,

$$\eta(0,t) = -\dot{\psi}(0,1,t),$$

and practically speaking, this requires a sensor installed on the generator itself. We use this measurement to construct a very simple feedback, under the form

$$u(t) = \eta(0,t).$$

This kind of feedback will require enough regularity for $\dot{\psi}$ on Γ_s, and this point will be clarified in section 5.1. Anyway, we are interested by the choices of $f(y)$ which will make the system (at least) strongly stable.

In terms of the original system (3.2)–(3.4), this feedback leads to the following equations:

$$\left\{\begin{array}{rcl} \Delta\psi &=& 0, \text{ in } \Omega\times[0,\infty),\\ \partial_n\psi + \ddot{\psi} &=& 0, \text{ on } \Gamma_s\times[0,\infty),\\ \partial_n\psi + f(y)\dot{\psi}(0,1) &=& 0, \text{ on } \Gamma_1\times[0,\infty),\\ \partial_n\psi &=& 0, \text{ on } \Gamma_f\cup\Gamma_2\times[0,\infty), \end{array}\right. \tag{5.1}$$

with initial conditions

$$\psi(0) = \varphi_0, \dot{\psi}(0) = \varphi_1, \text{ on } \Gamma_s.$$

We can already see that initial data of the form

$$\psi(0) = c, \dot{\psi}(0) = 0, \text{ on } \Gamma_s,$$

will stay invariant, i.e., we will have

$$\psi = c, \text{ in } \Omega \times [0, \infty).$$

Moreover, if we consider the natural energy of the system, i.e.,

$$E(\psi, \dot{\psi}) = \frac{1}{2} \int_\Omega |\nabla \psi|^2 + \frac{1}{2} \int_{\Gamma_s} \dot{\psi}^2,$$

we have $E(\psi, \dot{\psi}) = 0$ for such a solution. In order to have a correct framework for the stability analysis of the system (5.1), we will eliminate such initial data by choosing spaces adequately.

In the following we will show that for some choices of $f(y)$ the solution of the system (5.1) is strongly stable, although the natural energy is not decreasing: multiplying by $\dot{\psi}$ the first equation in (5.1) and integrating in Ω, we formally get that

$$\dot{E}(\psi, \dot{\psi}) = -\dot{\psi}(0,1) \int_{\Gamma_1} f(y) \dot{\psi}.$$

For the case of a plane rotating generator, we have $f(y) = y$ and if we take for example ψ such that for some t, $\dot{\psi}(t)|_{\Gamma_1} = \varepsilon - 1 + y$ for a sufficiently small $\varepsilon > 0$, it is easy to show that $\dot{E}(\psi, \dot{\psi}) > 0$ (for any $f(y)$ it is obviously possible to construct such a $\psi(t)$).

As a first intermediate step, and with the same notations as in section 3, we can reformulate (5.1) on Γ_s with a single unknown

$$\varphi = \psi|_{\Gamma_s},$$

which is the solution of

$$\ddot{\varphi} + \mathcal{A}\varphi + \beta \dot{\varphi}(0) = 0 \text{ on } \Gamma_s \times [0, \infty), \tag{5.2}$$

with initial conditions

$$\varphi(0) = \varphi_0, \dot{\varphi}(0) = \varphi_1, \tag{5.3}$$

where the operator $\mathcal{A}$ has the same definition as in section 3, i.e., $\mathcal{A}\varphi = \partial_n \mathrm{D}\varphi|_{\Gamma_s}$, where D is the Dirichlet map defined by (3.5). The function $\beta \in L^2(\Gamma_s)$ is defined by $\beta = \mathcal{B}f = \partial_n N f|_{\Gamma_s}$, where N is the Neumann map defined by (3.6). Finally, the original potential ψ is related to φ and f by the relation

$$\psi = \mathrm{D}\varphi + \dot{\varphi}(0)\mathrm{N}f,$$

and of course we still have $\varphi = \psi|_{\Gamma_s}$.

The next step consists in using the first order formulation of (5.2): as in section 3.3 we define a new unknown ξ, related to φ and $\dot{\varphi}$ by

$$\xi = \begin{pmatrix} \Pi_H \varphi \\ \Pi_H \dot{\varphi} \\ \Pi_{\mathbf{R}} \dot{\varphi} \end{pmatrix}, \ \xi_0 = \begin{pmatrix} \Pi_H \varphi_0 \\ \Pi_H \varphi_1 \\ \Pi_{\mathbf{R}} \varphi_1 \end{pmatrix}. \tag{5.4}$$

In order to exhibit the very general aspect of the method we will use to obtain our main result, we need to define an observation operator C such that $\eta(0)$, the elevation of the free surface at $x = 0$ (which is a feedback to the control), is obtained by $\eta(0) = C\xi$. Since we have $\eta(0) = -\dot{\varphi}(0)$, then (5.4) implies that C is defined by

$$C\xi = -(\xi_2(0) + \xi_3).$$

If we also consider the control operator B defined for $u \in \mathbf{R}$ by

$$Bu = bu, \text{ with } b = \begin{pmatrix} 0 \\ \Pi_H \beta \\ \Pi_{\mathbf{R}} \beta \end{pmatrix} \in X,$$

and the definition of the operator A given by

$$A\xi = \begin{pmatrix} \xi_2 \\ -\mathcal{A}\xi_1 \\ 0 \end{pmatrix},$$

then the problem (5.2), in terms of the new unknown ξ, is equivalent to the problem

$$\left\{ \begin{array}{rcl} \dot{\xi} & = & (A + BC)\xi, \ t > 0, \\ \xi(0) & = & \xi_0, \end{array} \right. \tag{5.5}$$

which appears as a perturbation of the original open-loop system. In the following, we will use the notation $A_f = A + BC$.

The stability of system (5.5) has been already studied in [14], for the case $f(y) = y$ (plane rotating generator), where it is shown that the eigenvalues of A_f have strictly negative real parts (in fact it is easily shown that this results will always hold if $f(y)$ is a positive function). But this reference does not give any result on the eventual generation of a semigroup.

The result on the real parts of A_f itself is not sufficient to conclude to strong stability, because we don't know if the eigenfunctions of A_f form a Riesz basis. One could apply a result of Lasiecka and Triggiani [9], [17] but the criterion is hard to apply, since one has to know the eigenvalues of A_f explicitly.

In the following we show that the solution of (5.5) is strongly stable by means of an ad hoc energy, and we will have to restrict ourselves to a particular choice of f, since the regularity of β will directly determine the energy space in which the stability result will hold.

5.1. Ad hoc energy. The ad hoc energy we propose is based on the bilinear form

$$B(\xi, \zeta) = \sum_{k \in \mathbf{Z}} d_k \, \langle \xi \, , \, \phi_k \rangle_X \, \overline{\langle \zeta \, , \, \phi_k \rangle}_X, \tag{5.6}$$

where the bar denotes the complex conjugate, ϕ_k denotes the kth eigenfunction of A, and

$$d_k = -\frac{C\phi_k}{\langle \phi_k \, , \, b \rangle_X}.$$

The definitions of C and b give

$$d_k = d_{-k} = \frac{w_k(0)}{\langle \beta \, , \, w_k \rangle} \text{ for } k \geq 0.$$

Since we only consider the case where $f(y)$ is positive, the positivity of d_k follows directly from Lemma 4.10.

Remark 5.1. The bilinear form $B(\xi,\zeta)$ is a reweighted form of the classical inner product associated with the open-loop system $\dot{\xi} = A\xi$, i.e.,

$$\begin{aligned}\langle \xi\,,\,\zeta\rangle_X &= \left\langle \mathcal{A}^{1/2}\xi_1\,,\,\mathcal{A}^{1/2}\zeta_1\right\rangle + \langle \xi_2+\xi_3\,,\,\zeta_2+\zeta_3\rangle \\ &= \sum_{k\in\mathbf{Z}} \langle \xi\,,\,\phi_k\rangle_X\,\overline{\langle \zeta\,,\,\phi_k\rangle}_X. \qquad \square\end{aligned}$$

The positivity of d_k allows us to claim that the bilinear form $B(.,.)$ defines a scalar product and

$$B(\xi,\xi) = \sum_{k\in\mathbf{Z}} d_k |\,\langle \xi\,,\,\phi_k\rangle_X\,|^2 \tag{5.7}$$

can be used as a norm defined on the associated energy space, which is to be identified.

Remark 5.2. The regularity of f will condition the asymptotic behavior of d_k, which will obviously determine the energy space. From an engineering point of view, it could be interesting to consider the case $f(y) = y$, since this corresponds to the most widely used device for wave generators. In this case the energy space can be identified as

$$H^1 \times H^{1/2} \times \mathbf{R},$$

but $b \notin H^1 \times H^{1/2} \times \mathbf{R}$, since for $f(y) = y$ we only have $\beta \in H^{1/2-\varepsilon}(\Gamma_s)\ \forall\ \varepsilon > 0$. This remark explains the particular choice we will make for f in what follows.

In the following, we will focus on the particular case when the "shape" of the generator is given by

$$f_\varepsilon(y) = \begin{cases} \frac{1}{1-\varepsilon}y, & \text{if} \quad 1-\varepsilon \geq y \geq 0, \\ \frac{1}{\varepsilon} - \frac{1}{\varepsilon}y, & \text{if} \quad 1 \geq y > 1-\varepsilon, \end{cases}$$

where $\varepsilon > 0$. We can see that $f_\varepsilon(y)$ is arbitrary close to the original shape, since $f_\varepsilon(y) \to y$ in $L^2(\Gamma_1)$, as $\varepsilon \to 0$. For this particular choice we have $f_\varepsilon \in H_c^{1/2}(\Gamma_1)$, and one can show from formulas (4.16) that

$$B^*\phi_k = \langle \phi_k\,,\,b\rangle_X = \frac{1}{\sqrt{\pi}}\,\langle w_k\,,\,\beta\rangle \tag{5.8}$$

$$= \frac{1}{\varepsilon\sqrt{\pi}\,k^2} + \frac{\cosh k(\varepsilon-1) - \varepsilon}{\varepsilon(\varepsilon-1)\sqrt{\pi}\,k^2\cosh k},\ k>0, \tag{5.9}$$

which gives

$$k^2 B^*\phi_k \to \frac{1}{\varepsilon\sqrt{\pi}},$$

when $k \to \infty$. Thus $\beta \in D(\mathcal{A})$.

PROPOSITION 5.3. *The energy space defined by the convergence of the series in* (5.7) *is equal to*

$$X_f = \tilde{H}_c^{3/2}(\Gamma_s) \times \tilde{H}^1(\Gamma_s) \times \mathbf{R}.$$

Proof. We recall that the natural energy space was

$$X = D(\mathcal{A}^{1/2}) \times H \times \mathbf{R},$$

and that the eigenvalues of $\mathcal{A}$ are given by $\lambda_k = k\tanh k$, for $k > 0$. If we use the expression of the eigenfunctions of A we have

$$B(\xi,\xi) = \frac{1}{\pi} d_0 \xi_3^2 + \frac{2}{\pi} \sum_{k>0} d_k \left(\lambda_k \langle \xi_1 \,,\, w_k \rangle^2 + \langle \xi_2 \,,\, w_k \rangle^2 \right), \tag{5.10}$$

and since we can easily show that

$$\lim_{k\to\infty} \frac{d_k}{\lambda_k^2} = \frac{1}{\varepsilon},$$

the series (5.10) is equivalent to the series

$$\frac{1}{\pi} d_0 \xi_3^2 + \frac{2}{\pi} \sum_{k>0} \lambda_k^3 \langle \xi_1 \,,\, w_k \rangle^2 + \lambda_k^2 \langle \xi_2 \,,\, w_k \rangle^2 .$$

Thus (5.10) will be convergent provided $\xi \in D(\mathcal{A}^{3/2}) \times D(\mathcal{A}) \times \mathbf{R}$, and we have

$$D(\mathcal{A}^{3/2}) = \left\{ \varphi \in D(\mathcal{A}),\ \mathcal{A}\varphi \in D(\mathcal{A}^{1/2}) \right\} = \tilde{H}_c^{3/2}(\Gamma_s). \qquad \square$$

PROPOSITION 5.4. *The domain of A_f is*

$$D(A_f) = \tilde{H}_c^2(\Gamma_s) \times \tilde{H}_c^{3/2}(\Gamma_s) \times \mathbf{R},$$

where the space $H_c^2(\Gamma_s)$ is defined by

$$\tilde{H}_c^2(\Gamma_s) = \left\{ \varphi \in \tilde{H}^2(\Gamma_s),\ \varphi'(0) = \varphi'(\pi) = 0 \right\}.$$

Proof. The domain of A_f is by definition

$$\begin{aligned} D(A_f) &= \{ \xi \in X_f,\ A_f \xi \in X_f \} \\ &= D(\mathcal{A}^2) \times D(\mathcal{A}^{3/2}) \times \mathbf{R}, \end{aligned}$$

where $D(\mathcal{A}^2) = \{ \varphi \in D(\mathcal{A}),\ \mathcal{A}\varphi \in D(\mathcal{A}) \}$. We can identify $D(\mathcal{A}^2)$ by making the following analysis: we will have $\varphi \in D(\mathcal{A}^2)$ if

$$\sum_{k>0} \lambda_k^4 \langle \varphi \,,\, \cos kx \rangle^2 < +\infty. \tag{5.11}$$

It is well known that the operator $-\frac{d}{dx^2}$ in $H = \tilde{L}^2(\Gamma_s)$ with the boundary conditions $\varphi'(0) = \varphi'(\pi) = 0$ has the eigenpairs $(k^2, \cos kx)$ for $k > 0$. Since λ_k^4 is equivalent to k^4, we can claim that the series in (5.11) will be convergent provided that φ is in the domain of this latter operator, which is exactly $\tilde{H}_c^2(\Gamma_s)$. $\square$

Hence, in the following, the space X_f will be endowed with the norm

$$\|\xi\|_f^2 \equiv B(\xi,\xi),$$

where $B(.,.)$ is defined by (5.6), and for $\xi, \zeta \in X_f$, the associated inner product will be

$$\langle \xi\,,\,\zeta\rangle_f \equiv B(\xi,\zeta),$$

and the superscript * will denote adjoint operators with respect to $\langle .\,,\,.\rangle_f$.

The last part of the paper relies on the following observation.

LEMMA 5.5. *The operators B and C verify*

$$B^* = -C,$$

where B^* is the adjoint operator of B, with respect to the inner product $\langle .\,,\,.\rangle_f$.

Proof. Let us take $\xi \in X_f$. We have

$$B^*\xi = \langle \xi\,,\,b\rangle_f, \text{ with } b = \begin{pmatrix} 0 \\ \Pi_H\beta \\ \Pi_{\mathbf{R}}\beta \end{pmatrix}.$$

But we have also

$$\begin{aligned}\langle \xi\,,\,b\rangle_f &= \sum_{k\in\mathbf{Z}} d_k \langle \xi\,,\,\phi_k\rangle \overline{\langle b\,,\,\phi_k\rangle} \\ &= -\sum_{k\in\mathbf{Z}} \frac{C\phi_k}{\langle b\,,\,\phi_k\rangle} \langle \xi\,,\,\phi_k\rangle \overline{\langle b\,,\,\phi_k\rangle} \\ &= -\sum_{k\in\mathbf{Z}} C\phi_k \langle \xi\,,\,\phi_k\rangle \\ &= -C\xi,\end{aligned}$$

where we have used the fact that $\langle b\,,\,\phi_k\rangle$ is a real quantity. □

The situation where $B^* = -C$ is often denoted by "collocation" of the sensor and the actuator, and there are many examples in the literature where such a property corresponds to a realizable actuator/sensor device (see e.g. [4] for the case of a beam with a piezoelectric actuator/sensor). It is interesting to see that the change of inner product has revealed a rather favorable situation, which was hidden in the original topology.

5.2. Strong stability. We note that the result given by Lemma 5.5 shows that C is bounded for the topology of X_f, although this operator was unbounded in the "natural" topology. We will use this fact in the proof of our main result.

THEOREM 5.6. *The system*

$$\left\{\begin{array}{rcl} \dot{\xi} &=& A_f\xi,\ t>0, \\ \xi(0) &=& \xi_0, \end{array}\right. \tag{5.12}$$

with an initial data ξ_0 in X_f, is strongly stable, i.e.,

$$\lim_{t\to\infty} \|\xi(t)\|_f = 0,$$

moreover, the decay rate of $\|\xi(t)\|_f$ cannot be uniform.

Proof. We first recall that $A_f = A + BC = A - BB^*$, and we can easily show that $A^* = -A$ for the inner product $\langle .\,,\,.\rangle_f$. Hence we have for $\xi \in D(A_f)$

$$\begin{aligned}\text{Re}\,\langle A_f\xi\,,\,\xi\rangle_f &= \text{Re}\,\langle A\xi\,,\,\xi\rangle_f - \langle BB^*\xi\,,\,\xi\rangle_f \\ &= -|B^*\xi|^2 \le 0,\end{aligned} \tag{5.13}$$

so that $A - BB^*$ is dissipative.

We then apply the result given in [2]: if A generates a contraction semigroup in X_f and has a compact resolvent, then $A - BB^*$ generates a strongly stable semigroup provided that the pair (A, B) is approximately controllable in the sense of Definition 4.14, i.e., $B^*\phi_k \neq 0 \ \ \forall k \in \mathbf{Z}$.

First, we see that $A - \lambda I$ is maximal in X_f for $\lambda > 0$, if $\mathcal{A} + \lambda^2 I$ is onto from $D(\mathcal{A})$ to $D(\mathcal{A}^2)$. Proposition 3.2 together with the definition of $D(\mathcal{A}^2)$ show that this last result is true. Hence, the resolvent $(\lambda I - A)^{-1}$: $X_f \to D(A_f)$ is bounded. It is also compact, since the injection of $D(A_f) = D(\mathcal{A}^2) \times D(\mathcal{A}^{3/2}) \times \mathbf{R}$ into $X_f = D(\mathcal{A}^{3/2}) \times D(\mathcal{A}) \times \mathbf{R}$ can be easily shown to be compact. Thus A generates a contraction semigroup in X_f and has a compact resolvent in X_f.

Second, for any $k \in \mathbf{Z}$ we have $B^*\phi_k = -C\phi_k = \frac{1}{\sqrt{\pi}}$.

Finally, the decay rate cannot be uniform since BB^* is compact in X_f (bounded and one-dimensional range in our case). □

Remark 5.7. The approach we have used to show the stability result cannot be applied if we use the real "shape" of the plane rotating generator, i.e., $f(y) = y$. In this case the perturbation operator BC is not A-bounded (see [8]) in the natural topology. This may suggest that uniform stability could eventually hold, but in this case, even the well-posedness of the feedback system seems not to be a trivial issue.

Acknowledgment. The author would like to thank Professor Hans Zwart for helpful discussions concerning the controllability problem in section 4.2.1.

REFERENCES

[1] S. Avdonin and S. Ivanov, *Families of Exponentials*, Cambridge University Press, Cambridge, UK, 1995.

[2] C. D. Benchimol, *A note on weak stabilizability of contraction semigroups*, SIAM J. Control Optim., 16 (1978), pp. 373–379.

[3] R. Curtain and H. Zwart, *An Introduction to Infinite-Dimensional Linear Systems Theory*, Springer-Verlag, Berlin, 1995.

[4] J. Dosch, D. Inman, and E. Garcia, *A self-sensing piezoelectric actuator for collocated control*, J. Intell. Mater. Systems Struct., 3 (1992), pp. 166–185.

[5] P. Grisvard, *Elliptic Problems in Non-Smooth Domains*, Pitman, Boston, 1985.

[6] G. Joly, S. Mottelet, and J. Yvon, *Analysis of the control of wave generators in a canal*, in Control of Partial Differential Equations and Applications (Laredo, 1994), Lecture Notes in Pure and Appl. Math. 174, Marcel Dekker, New York, 1996, pp. 119–134.

[7] G. Joly, S. Mottelet, and J. Yvon, *Application of H_∞ control to wave generators in a canal*, in Control Problems in Industry, Progr. Systems Control Theory 21, I. Lasiecka and B. Morton, eds., Birkhäuser, Basel, Boston, 1995, pp. 179–204.

[8] T. Kato, *Perturbation Theory for Linear Operators*, Springer-Verlag, Berlin, 1980.

[9] I. Lasiecka and R. Triggiani, *Finite rank, relatively bounded perturbations of c-semi-groups, part* II: *Spectrum allocation and Riesz basis in parabolic and hyperbolic feedback systems*, Ann. Mat. Pura Appl., CXLIII (1986), pp. 47–100.

[10] J. Lions, *Contrôlabilité exacte, perturbation et stabilisation de systèmes distribués* 2, Collection Recherches en Mathématiques Appliquées 9, Masson, Paris, 1988.

[11] J. Lions and E. Magenes, *Problèmes aux limites non homogènes et applications* 3, Coll. Travaux et Recherches Mathématiques 20, Dunod, Paris, 1968.

[12] J. Lions and E. Zuazua, *The cost of controlling unstable systems: Time irreversible systems*, Rev. Mat. Univ. Complut. Madrid, 10 (1997), pp. 481–523.

[13] S. Micu and E. Zuazua, *Boundary controllability of a linear hybrid system arising in the control of noise*, SIAM J. Control Optim., 35 (1997), pp. 1614–1637.

[14] S. Mottelet, *Quelques aspects théoriques et numériques du contrôle d'un bassin de carènes*, Ph.D. thesis, Université de Technologie de Compiègne, Compiègne, France, 1994.

[15] S. Mottelet, G. Joly, and J. Yvon, *Design of a feedback controller for wave generators in a canal using H_∞ methods*, in System Modelling and Optimization (Compiègne, 1993), Lecture Notes in Control and Inform. Sci. 197, Springer-Verlag, London, UK, 1994,

pp. 716–726.
[16] L. Schwartz, *Etude des sommes d'exponentielles*, Hermann, Paris, 1959.
[17] R. Triggiani, *Finite rank, relatively bounded perturbations of semi-groups generators, part* III: *A sharp result on the lack of uniform stabilization*, Differential Integral Equations, 3 (1990), pp. 503–522.
[18] R. M. Young, *An Introduction to Nonharmonic Fourier Series*, Academic Press, New York, 1980.

SIAM J. CONTROL OPTIM.
Vol. 38, No. 3, pp. 736–750

EXACT BOUNDARY CONTROLLABILITY OF A MAXWELL PROBLEM*

N. WECK†

Abstract. We consider the problem of steering some initial state of the time-dependent Maxwell equations to rest by controlling the lateral boundary condition. Depending on the topology (Betti numbers) of the given domain Ω, we identify completely those subspaces which can be steered to rest and those which cannot. In the positive case it is shown that each time interval which is longer than diam Ω suffices. This result is almost sharp for the class of regions Ω which contain a segment of length diam Ω.

Key words. exact boundary controllability, Maxwell

AMS subject classifications. 93B05, 35B37, 35Q60, 78A25, 93C20

PII. S0363012998347559

1. Introduction. Consider a bounded domain $\Omega \subset \mathbb{R}^3$. An initial boundary value problem for Maxwell's equations may be written (formally) as follows:

$$\left.\begin{aligned} \varepsilon \partial_t E + \operatorname{curl}_x H &= 0 \\ \mu \partial_t H - \operatorname{curl}_x E &= 0 \end{aligned}\right\} \quad \text{in} \quad \mathbb{R}_+ \times \Omega,$$

$$\nu \wedge H = -J \quad \text{in} \quad \mathbb{R}_+ \times \partial\Omega,$$

$$E(0,\cdot) = E^0, \quad H(0,\cdot) = H^0 \quad \text{in} \quad \Omega.$$

Here ε, μ are given positive constants and ν is the exterior unit normal of Ω.

We want to investigate the problem whether it is possible to choose the "boundary control" J such that the given "initial state" (E^0, H^0) is driven to some desired "final state" (E^1, H^1) in a time interval $[0, T]$. This problem has been discussed for special regions Ω with restricted controls by D. L. Russell [23] and K. A. Kime [11] as well as for general regions and general controls by J. E. Lagnese [14], A. Bensoussan [5], O. Nalin [17], V. Komornik [12], and B. V. Kapitonov [10]. (Restricting controls gives a stronger result; therefore the last five results do not imply the first two.) For related results on corresponding stabilization problems see also [3] and [4].

However, the first three general results ([14], [5], and [17]) are based on an inequality in [13, Thm. 7.1] which does not hold in general. Nevertheless, the first two results are essentially true because by control-theoretic considerations the authors are led to consider mainly the special case of star-shaped domains where it *does* hold. The fourth [12] and fifth [10] address only the star-shaped and so-called "sub-star-shaped" cases trying to find an optimal time for exact controllability. On the other hand, spherical shells and tori are counterexamples, and these lead to interesting control-theoretic problems.

In the present paper we shall completely identify those initial states which can be controlled and those which cannot. Furthermore we shall show that in the positive

*Received by the editors November 12, 1998, accepted for publication (in revised form) June 28, 1999; published electronically February 29, 2000.
http://www.siam.org/journals/sicon/38-3/34755.html
†Fachbereich 6-Mathematik und Informatik, Universität GH Essen, Universitätsstraße 2, D-45141 Essen, Germany (weck@uni-essen.de).

case any time interval of length larger than diam Ω suffices. Even in the star-shaped case this is at least as good as the result in [12], improves it in a variety of cases, and is almost sharp for a class of regions (see Remark 5 below).

We shall make extensive use of the theory of the Maxwell system as exhibited in [15] (and the literature therein). Though we shall not dwell upon this point, we remark that this theory would enable us to treat more general problems (nonsmooth boundaries, inhomogeneous anisotropic media, H. Weyl's [29] generalization of the Maxwell system to differential forms of arbitrary rank in $\Omega \subset \mathbb{R}^N$). For another suitable source of reference for our present simple setting see [6].

In order to keep the analytical efforts low we choose states and controls from convenient spaces. In particular, we strive to use only standard elliptic theory and standard results of the spectral calculus for the self-adjoint time-independent Maxwell operator (or semigroup theory [1, Thms. 4.8.1 and 4.8.2]). Concerning control theory, we use only the analogue of D. L. Russell's trick to obtain suitable boundary controls for the wave equation [20], [22].

Compared to [14], [5], [17], and [12] we need a slightly larger space of controls (with respect to regularity) or, conversely, have to impose stronger regularity upon the states. However, the simplicity of our argument as well as the more precise nature of our results seems to justify our approach. Furthermore, recent regularity results (see [24] and the literature cited there) give hope that this gap can be closed.

We introduce the following modifications and simplifications of our problem:

(i) $(E^1, H^1) = (0, 0)$ (without loss of generality because the family of evolution operators is a group);

(ii) $\varepsilon = \mu = 1$ (for the case of constant scalar dielectricity ε and permeability μ—which is considered in the previous literature—this may be achieved by a standard transformation);

(iii) as in [14] and [5], we assume Ω to be a bounded open set "lying on one side" of its C^∞-boundary $\partial\Omega =: \Gamma$ with exterior unit normal $\nu : \Gamma \to S^2$ (although C^2-regularity would be sufficient).

Hence (interchanging the roles of E and H) we say that (E, H) solves $\mathcal{P}(E^0, H^0; J)$ in the "time interval" $I := [\,0\,, T\,]$ or $I := [\,0\,, \infty\,)$ if (in a sense to be defined below)

$$\text{(IBVP)} \qquad \left.\begin{array}{lll} \partial_t E = \nabla_x \wedge H, \quad \partial_t H = -\nabla_x \wedge E & \text{in} & I \times \Omega \\ (E(0,\cdot), H(0,\cdot)) = (E^0, H^0) & & \\ \nu \wedge E = J & \text{on} & I \times \Gamma \end{array}\right\}.$$

2. Notation and a solution theory for the direct problem. For the convenience of the reader we want to exhibit an easily accessible theory of the direct problem tailored to fit the need of having a convenient description of the map "control to state." So we strive to use only the fact that the Maxwell operator is self-adjoint plus the corresponding spectral calculus. As mentioned in the introduction, this approach could be generalized in various directions.

We shall use the notation of [15]. In particular, spaces of $\mathbb{C}$-valued functions will be denoted by roman letters whereas script letters indicate spaces of $\mathbb{C}^3$-valued functions. Usually Ω will be fixed, so we shall indicate the dependence of our function spaces on the domain S only if it is different from Ω. Hence

$$L^2 := \{\, f : \Omega \to \mathbb{C} \;\mid\; f \quad \text{measurable and square integrable} \,\},$$
$$\mathcal{L}^2 := \left\{\, v : \Omega \to \mathbb{C}^3 \;\mid\; v_1, \ldots, v_3 \in L^2 \right\}.$$

Similarly, for $s \in \mathbb{R}$ we denote L^2-Sobolev spaces by H^s and $\mathcal{H}^s$ in the scalar, respectively, vector valued case and in the same spirit

$$C^\infty(S) := C^\infty(S, \mathbb{C}), \quad \mathcal{C}^\infty(S) := C^\infty(S, \mathbb{C}^3).$$

Furthermore, following [15] we introduce

$$\begin{aligned}
\mathcal{R} &:= \left\{ v \in \mathcal{L}^2 \quad \middle| \quad \nabla \wedge v := \operatorname{curl} v \in \mathcal{L}^2 \right\}, \\
\mathcal{D} &:= \left\{ v \in \mathcal{L}^2 \quad \middle| \quad \langle \nabla, v \rangle := \operatorname{div} v \in L^2 \right\}, \\
\mathcal{R}_0 &:= \{ v \in \mathcal{R} \quad | \quad \nabla \wedge v = 0 \}, \\
\mathcal{D}_0 &:= \{ v \in \mathcal{D} \quad | \quad \langle \nabla, v \rangle = 0 \}.
\end{aligned}$$

The following spaces, $\overset{0}{\mathcal{R}}$, $\overset{0}{\mathcal{D}}$, and $\overset{0}{H}{}^1$, generalize the boundary conditions $\nu \wedge v|_\Gamma = 0$, $\langle \nu,\ v|_\Gamma \rangle = 0$, and $f|_\Gamma = 0$, respectively:

$$\begin{aligned}
\overset{0}{\mathcal{R}} &:= \{ v \in \mathcal{R} \quad | \quad \langle \nabla \wedge v, \varphi \rangle = \langle v, \nabla \wedge \varphi \rangle \quad \text{for all} \quad \varphi \in \mathcal{R} \}, \\
\overset{0}{\mathcal{D}} &:= \left\{ v \in \mathcal{D} \quad | \quad \langle \operatorname{div} v, \varphi \rangle = -\langle v, \nabla \varphi \rangle \quad \text{for all} \quad \varphi \in H^1 \right\}, \\
\overset{0}{H}{}^1 &:= \left\{ f \in H^1 \quad \middle| \quad \langle \nabla f, \varphi \rangle = -\langle f, \operatorname{div} \varphi \rangle \quad \text{for all} \quad \varphi \in \mathcal{D} \right\}.
\end{aligned}$$

Here the natural scalar product both in L^2 and $\mathcal{L}^2$ is denoted by $\langle \cdot, \cdot \rangle$ (and by $\langle \cdot, \cdot \rangle_S$ if the domain S needs to be specified).

The space $\overset{0}{H}{}^1$ coincides with the closure of the test functions $\overset{0}{C}{}^\infty$ in H^1 and this is used to *define* $\overset{0}{H}{}^s(S)$ for an arbitrary $s \in \mathbb{R}$.

LEMMA 2.1 (see [15, pp. 144 ff.]). *The operator*

$$\begin{array}{rccc}
A & : & \overset{0}{\mathcal{R}} \times \mathcal{R} \subset \mathcal{L}^2 \times \mathcal{L}^2 & \longrightarrow & \mathcal{L}^2 \times \mathcal{L}^2, \\
 & & (E, H) & \longmapsto & (i\nabla \wedge H, -i\nabla \wedge E)
\end{array}$$

is self-adjoint with respect to the natural scalar product

$$\langle\langle (E^1, H^1), (E^2, H^2) \rangle\rangle := \langle E^1, E^2 \rangle + \langle H^1, H^2 \rangle.$$

For each $(\Phi^0, \Psi^0) \in D(A) = \overset{0}{\mathcal{R}} \times \mathcal{R}$ *the family*

$$(\Phi(t), \Psi(t)) := \exp(-itA)(\Phi^0, \Psi^0), \quad t \in \hat{I} := [\,0\,,\,\infty\,)$$

defines a strong solution

$$(\Phi, \Psi) \in C^0(\hat{I}, D(A)) \cap C^1(\hat{I}, \mathcal{L}^2 \times \mathcal{L}^2)$$

of (IBVP) *with* $J = 0$.

The strong solutions (Φ, Ψ) defined in Lemma 2.1 will be used as test functions in order to define weak solutions to (IBVP) with $J \neq 0$ in some finite interval $I := [\,0\,,\,T\,]$. We introduce

$$M := M(T) := (\,0\,,\,T\,) \times \Gamma, \quad Z := Z(T) := (\,0\,,\,T\,) \times \Omega$$

and for simplicity take the control from

$$\mathcal{U} := \mathcal{U}(T) := \mathcal{H}_\tau^{1/2}(M(T)).$$

Here (and in what follows) the index τ denotes spaces consisting of "tangential" vector fields, i.e., we require

$$\langle \nu(y),\, J(t,y)\rangle = 0 \quad \text{for} \quad (t,y)\in M, \quad J\in\mathcal{U},$$

in view of the boundary condition in (IBVP). Each $J\in\mathcal{U}$ may be represented by

$$\hat{J}\in \mathcal{H}^1(Z)$$

in the sense that $\mathcal{T}_\tau \hat{J} := \nu\wedge\mathcal{T}\hat{J} = J$ (on M), where here and in what follows $\mathcal{T}$ denotes the ordinary trace operator [25, Prop. 4.5].

LEMMA 2.2. *Let $(E^0,H^0)\in\mathcal{L}^2\times\mathcal{L}^2$ and $J\in\mathcal{U}$. Then there exists a unique $(E,H)\in C^0(I,\mathcal{L}^2\times\mathcal{L}^2)$ such that*

$$\begin{aligned}\langle\!\langle (E(t),H(t)),(\Phi(t),\Psi(t))\rangle\!\rangle - \langle\!\langle (E^0,H^0),(\Phi^0,\Psi^0)\rangle\!\rangle \\ = -\int_0^t\int_\Gamma \langle J(s,y),\,\Psi(s,y)\rangle d\mathbf{o}(y)ds\end{aligned} \tag{1}$$

for all $t\in I$ and all (Φ,Ψ) defined by

$$(\Phi(t),\Psi(t)) := \exp(-itA)(\Phi^0,\Psi^0), \quad (\Phi^0,\Psi^0)\in D(A).$$

Proof. First of all, we need an interpretation of the right-hand side of (1). This may be achieved with the aid of [8, Lem. VII.4.2] saying that there exists a unique continuous trace operator

$$\mathcal{T}_\tau \;:\; \mathcal{R} \;\longrightarrow\; \mathcal{H}_\tau^{-1/2}(\Gamma)$$

which extends the map $\psi\mapsto\nu\wedge\psi|_\Gamma$ from $\mathcal{C}^\infty(\overline{\Omega})$ into $\mathcal{C}^\infty_\tau(\Gamma)$, say. Therefore, from Lemma 2.1,

$$t\mapsto \mathcal{T}_\tau\Psi(t)\in C^0(I,\mathcal{H}_\tau^{-1/2}(\Gamma))\subset\mathcal{H}_\tau^{-1/2}(M),$$

and the right-hand side of (1) may be interpreted as $\langle \nu\wedge\mathcal{T}_\tau\overline{\Psi},\, J\rangle_{X',X}$, where $\langle\cdot,\cdot\rangle$ denotes the duality between the Banach space $X := \mathcal{H}_\tau^{1/2}(M)$ and its topological dual. Furthermore, $\mathcal{H}^{-1/2}(M)$ is identified with the dual of $\overset{0}{\mathcal{H}}{}^{1/2}(M) = \mathcal{H}^{1/2}(M)$ (see [25, Ex. 4.1.16]). Finally, the index τ in $\mathcal{H}_\tau^{-1/2}$ is defined in the weak sense:

$$\Psi\in\mathcal{H}_\tau^{-1/2}(M) :\Leftrightarrow \langle\Psi,\,\varphi\cdot\nu\rangle = 0 \quad \text{for all} \quad \varphi\in H^{1/2}(M).$$

Uniqueness of (E,H) is clear because $D(A)$ is dense in $\mathcal{L}^2\times\mathcal{L}^2$ and $t\mapsto\exp(-itA)$ is a group.

As to existence, we introduce a representation $\hat{J}\in\mathcal{H}^1(Z)\subset C^0(I,\mathcal{L}^2)$ of J (i.e., $J=\mathcal{T}_\tau\hat{J}$) and make the ansatz

$$(E,H) = (\tilde{E}+\hat{J},\tilde{H}),$$

which (formally) leads to

$$\begin{aligned}\partial_t\tilde{E} &= \nabla_x\wedge\tilde{H} - \partial_t\hat{J},\\ \partial_t\tilde{H} &= -\nabla_x\wedge\tilde{E} - \nabla_x\wedge\hat{J},\\ \nu\wedge\tilde{E}\Big|_\Gamma &= 0,\\ (\tilde{E}(0,\cdot),\tilde{H}(0,\cdot)) &\overset{!}{=} (E^0-\hat{J}(0,\cdot),H^0) =: (\tilde{E}^0,\tilde{H}^0).\end{aligned} \tag{2}$$

So we try

$$
\begin{aligned}
(E(t),H(t)) := {} & (\hat{J}\,(t,\cdot),0) \\
& + \exp(-itA)(\tilde{E}^0,\tilde{H}^0) \\
& + \int_0^t \exp(-i(t-s)A)(F(s),G(s))ds
\end{aligned} \tag{3}
$$

with

$$
(F,G) := (-\partial_t\,\hat{J}, -\nabla_x\wedge\,\hat{J}) \in L^2(I,\mathcal{L}^2\times\mathcal{L}^2). \tag{4}
$$

The spectral calculus (or semigroup theory) shows that indeed

$$
(E,H)\in C^0(I,\mathcal{L}^2\times\mathcal{L}^2).
$$

It remains to prove (1). To this end we put

$$
(\Phi(t),\Psi(t)) := \exp(-itA)(\Phi^0,\Psi^0), \qquad (\Phi^0,\Psi^0)\in D(A)
$$

and note that

$$
\langle\!\langle \exp(-itA)(\tilde{E}^0,\tilde{H}^0),(\Phi(t),\Psi(t))\rangle\!\rangle = \langle\!\langle(\tilde{E}^0,\tilde{H}^0),(\Phi^0,\Psi^0)\rangle\!\rangle \tag{5}
$$

because $\exp(-itA)$ is unitary. Furthermore,

$$
\begin{aligned}
&\left\langle\!\!\left\langle \int_0^t \exp(-i(t-s)A)(F(s),G(s))ds,(\Phi(t),\Psi(t))\right\rangle\!\!\right\rangle \\
&\quad = \left\langle\!\!\left\langle \int_0^t \exp(isA)(F(s),G(s))ds,(\Phi^0,\Psi^0)\right\rangle\!\!\right\rangle \\
&\quad = \int_0^t \langle\!\langle (F(s),G(s)),\exp(-isA)(\Phi^0,\Psi^0)\rangle\!\rangle ds \\
&\quad = \int_0^t \left(\langle -\partial_s\,\hat{J}\,(s,\cdot),\,\Phi(s,\cdot)\rangle - \langle\nabla_x\wedge\,\hat{J}\,(s,\cdot),\,\Psi(s,\cdot)\rangle\right) ds \\
&\quad = \int_0^t \left(\langle\hat{J}\,(s,\cdot),\,\partial_s\Phi(s,\cdot)\rangle - \langle\nabla_x\wedge\,\hat{J}\,(s,\cdot),\,\Psi(s,\cdot)\rangle\right) \\
&\qquad - \langle\hat{J}\,(t,\cdot),\,\Phi(t,\cdot)\rangle + \langle\hat{J}\,(0,\cdot),\,\Phi(0,\cdot)\rangle.
\end{aligned}
$$

Because of $\partial_s\Phi = \nabla_x\wedge\Psi$, the last integral yields the right-hand side of (1). Therefore, combining our last computation with (3) and (5) gives (1). □

In the following we shall refer to the solution (E,H) just introduced (defined by (3) and (2),(4)) as "the solution to $\mathcal{P}(E^0,H^0;J)$ in I." Two simple applications of the spectral calculus will be useful later on.

LEMMA 2.3. *Let $(\Phi^0,\Psi^0)\in N(A)$ (the null-space of A). Then*

$$
(\Phi(t),\Psi(t)) := \exp(-itA)(\Phi^0,\Psi^0) = (\Phi^0,\Psi^0)
$$

does not depend on t, and we have

$$
\begin{aligned}
&\langle\!\langle (E(t),H(t)),(\Phi^0,\Psi^0)\rangle\!\rangle - \langle\!\langle (E^0,H^0),(\Phi^0,\Psi^0)\rangle\!\rangle \\
&\qquad = -\int_0^t\int_\Gamma \langle J(\tau,y),\,\Psi^0(y)\rangle d\mathbf{o}(y)d\tau
\end{aligned} \tag{6}
$$

for all $t \in I$ and all solutions (E,H) of $\mathcal{P}(E^0,H^0;J)$ in I.

LEMMA 2.4. *Let $(F,G) \in C^1(I,\mathcal{L}^2\times\mathcal{L}^2)$. Then $(E(T),H(T)) \in D(A)$ for*

$$(E(T),H(T)) := \int_0^T \exp(itA)(F(t),G(t))dt.$$

In particular, if (E,H) solves $\mathcal{P}(E^0,H^0;J)$ in I with $(E^0,H^0)\in D(A)$ and J which is represented by $\hat{J}\in C^2(I,\mathcal{L}^2)\cap C^1(I,\mathcal{R})$ with $\hat{J}(T)\in\overset{0}{\mathcal{R}}$, then $(E(T),H(T))\in D(A)$.

3. Hodge–Helmholtz decompositions. Results on exact controllability are related to the topological nature of Ω via Hodge–Helmholtz decompositions. Furthermore, there are also some decompositions on $\partial\Omega$ which are needed to give a precise account of the space of controllable states. So in this section we shall provide these decompositions in classical terms. But as mentioned in the introduction, H. Weyl's generalization of the Maxwell system could be investigated in the same spirit. For references from the vast literature on this topic, see, e.g., [18], [19], or [6] and their references.

Here and in what follows, the symbols $\oplus$, $\ominus$, and $\perp$ indicate orthogonality in $\mathcal{L}^2$. We recall the following.

LEMMA 3.1 (Hodge–Helmholtz decomposition). *With $\mathcal{H}_D := \mathcal{R}_0\cap\mathcal{D}_0\cap\overset{0}{\mathcal{R}}$ and $\mathcal{H}_N := \mathcal{R}_0\cap\mathcal{D}_0\cap\overset{0}{\mathcal{D}}$ we have the following orthogonal decompositions:*

$$\mathcal{L}^2 = (\nabla\wedge\mathcal{R})\oplus\mathcal{H}_D\oplus(\nabla\overset{0}{H}{}^1) = \mathcal{D}_0\oplus(\nabla\overset{0}{H}{}^1), \tag{7}$$

$$\mathcal{L}^2 = (\nabla\wedge\overset{0}{\mathcal{R}})\oplus\mathcal{H}_N\oplus(\nabla H_1) = (\mathcal{D}_0\cap\overset{0}{\mathcal{D}})\oplus\nabla H_1. \tag{8}$$

Proof. Theorem 8.4 of [15] says that these decompositions hold in an arbitrary open set if we replace the intervening subspaces $\nabla\wedge\mathcal{R}$, etc. by their respective closures $\overline{\nabla\wedge\mathcal{R}}$, etc. However, if Ω enjoys the "Maxwell compactness property" (which is certainly true if $\partial\Omega$ is smooth [15, Thm. 8.6]), then $\nabla\wedge\mathcal{R} = \overline{\nabla\wedge\mathcal{R}}$, etc. (cf. [16]). □

Let us also mention that there are larger classes of (nonsmooth) regions for which this property holds [28], [26], [15], [30] and that Lemma 3.1 is a special case of a more general theory (cf., e.g., [25, section 5.9]).

The orthogonal projectors onto $\nabla\wedge\mathcal{R}$ and $\nabla\wedge\overset{0}{\mathcal{R}}$ will be denoted by P_D and P_N, respectively. They may be constructed by solving appropriate coercive boundary value problems which enjoy regularity properties (cf. [16]).

LEMMA 3.2. *The following boundary value problem is coercive:*

$$(P_D)\qquad \left.\begin{array}{l} \text{Find } v\in\mathcal{V}_D := (\nabla\wedge\overset{0}{\mathcal{R}})\cap\mathcal{R} \text{ such that} \\ \qquad\qquad \langle \operatorname{curl} v, \operatorname{curl}\varphi\rangle = \langle E, \operatorname{curl}\varphi\rangle \\ \text{for all } \varphi\in\mathcal{V}_D. \end{array}\right\}$$

Let

$$\begin{array}{rccc} G_D & : & \mathcal{L}^2 & \longrightarrow & \mathcal{V}_D, \\ & & E & \longmapsto & v \end{array}$$

denote its solution operator. Then $P_DE = \operatorname{curl} G_DE$.

The problem (P_D) has the regularity property: $E \in \mathcal{H}^k$ implies $G_D E \in \mathcal{H}^{k+1}$.

LEMMA 3.3. *The following boundary value problem is coercive:*

$$(P_N)\qquad \left.\begin{array}{l} \text{Find } w \in \mathcal{V}_N := (\nabla \wedge \mathcal{R}) \cap \overset{0}{\mathcal{R}} \text{ such that} \\ \qquad\qquad \langle \operatorname{curl} w, \operatorname{curl} \psi\rangle = \langle E, \operatorname{curl} \psi\rangle \\ \text{for all } \psi \in \mathcal{V}_N. \end{array}\right\}$$

Let

$$\begin{array}{rccc} G_N & : & \mathcal{L}^2 & \longrightarrow & \mathcal{V}_N, \\ & & E & \longmapsto & w \end{array}$$

denote its solution operator. Then $P_N E = \operatorname{curl} G_N E$.

The problem (P_N) has the regularity property: $E \in \mathcal{H}^k$ implies $G_N E \in \mathcal{H}^{k+1}$.

In order to obtain precise controllability results it is necessary to recall the construction and properties of the spaces $\mathcal{H}_D$ and $\mathcal{H}_N$ as well as to decompose ∇H^1 even further.

Standard elliptic theory (see, e.g., [25, Prop. 1.7]) shows the following.

LEMMA 3.4. *For $s \in [1/2, \infty)$ and $g \in H^s(\Gamma)$, define u as the solution of the Dirichlet problem $\Delta u = 0$, $\mathcal{T} u = g$. Then $u \in H^{s+1/2}$ and the operator $\hat{G}$ defined by*

$$\begin{array}{rccc} \hat{G} & : & H^s(\Gamma) & \longrightarrow & \mathcal{H}^{s-1/2} \cap \nabla H^1, \\ & & g & \longmapsto & \nabla u \end{array}$$

is continuous. Its null-space consists of $\operatorname{Lin}\{\mathbf{1}\}$*—the constant functions—and its range equals $\mathcal{H}^{s-1/2} \cap (\nabla H^1 \ominus \nabla \overset{0}{H}{}^1)$. Therefore, the operator*

$$\begin{array}{rccccc} G & : & \overset{\bullet}{H}{}^s(\Gamma) := \{u \in H^s(\Gamma) \colon \langle u, \mathbf{1}\rangle_\Gamma = 0\} & \longrightarrow & \mathcal{H}^{s-1/2} \cap (\nabla H^1 \ominus \nabla \overset{0}{H}{}^1), \\ & & g & \longmapsto & \nabla u \end{array}$$

is a topological isomorphism.

Let $K(\Gamma)$ denote the *locally* constant functions on Γ. Then $\mathcal{H}_D$ may be described in terms of G (cf. [16], [6]).

LEMMA 3.5. *We have $\mathcal{H}_D = G(K(\Gamma))$. More precisely (in the notation of Lemma* 3.4*),*

$$\begin{array}{rccc} G_0 & : & \overset{\bullet}{H}{}^s(\Gamma) \cap K(\Gamma) & \longrightarrow & \mathcal{H}_D, \\ & & g & \longmapsto & \nabla u \end{array}$$

is a topological isomorphism. Therefore, the space $\mathcal{H}_D$ of "Dirichlet fields" is finite-dimensional ($\dim \mathcal{H}_D = \beta_1$ *if there are $\beta_1 + 1$ connected components of Γ) and contained in $\mathcal{C}^\infty(\overline{\Omega})$.*

Similar results are known for $\mathcal{H}_N$. We only need the following lemma which may be deduced from [27, Thm. 2.2] (or cf. [9], [16], [6]).

LEMMA 3.6. *The space $\mathcal{H}_N$ of "Neumann fields" is finite-dimensional and contained in $\mathcal{C}^\infty(\overline{\Omega})$.*

REMARK 1. *There are elementary examples of regions Ω where $\mathcal{H}_D$ or $\mathcal{H}_N$ are nontrivial:*

(i) *If $\partial\Omega$ has two connected components Γ_0, Γ_1 (e.g., in the case of a spherical shell $\Omega := \{x \in \mathbb{R}^3 \colon 1 < |x| < 2\}$), then $\nabla u \in \mathcal{H}_D \setminus \{0\}$ if u is a solution to Dirichlet's boundary value problem for $\Delta u = 0$ with $u|_{\Gamma_0} = 0$ and $u|_{\Gamma_1} = 1$.*

(ii) *If Ω is invariant under rotations around the x_3-axis A and $A \cap \overline{\Omega} = \emptyset$, e.g., in the case of a torus*

$$\Omega := \left\{ x = \begin{bmatrix} x' \\ x_3 \end{bmatrix} \in \mathbb{R}^3 \colon \left| |x'|\, x - R x' \right| < r\, |x'| \right\}, \qquad 0 < r < R,$$

then $v \in \mathcal{H}_N \setminus \{0\}$ for

$$v(x) := (x_1^2 + x_2^2)^{-1} \begin{bmatrix} -x_2 \\ x_1 \\ 0 \end{bmatrix}.$$

These examples disprove the claims made in [13, *Thm.* 7.1], [14, *section* 2], [5, *section* 1.2], *and* [17, *section* 1].

DEFINITION 3.7. *For $\varphi \in C^\infty(\Gamma)$ we define the "surface gradient" $\nabla_o \varphi \in \mathcal{C}^\infty(\Gamma)$ by extending φ arbitrarily to some $\overset{\wedge}{\varphi} \in C^\infty(\mathbb{R}^3)$ and using the tangential components of $\nabla \overset{\wedge}{\varphi}$:*

$$\nabla_o \varphi := -\nu \wedge (\nu \wedge \nabla \overset{\wedge}{\varphi}).$$

The continuous extension of ∇_o to $H^1(\Gamma)$ will also be denoted by ∇_o and its adjoint (with respect to $\int_\Gamma \dots d\mathbf{o}$) by div_o.

REMARK 2. *Of course, both ∇_o and* div_o *may be defined intrinsically without reference to the embedding into $\mathbb{R}^3$. In fact, identifying scalar functions and tangential vector fields with differential forms of rank* 0 *and* 1, *respectively, we see that ∇_o and* div_o *correspond to the differential d and codifferential $*d*$.*

A compactness argument shows that $\nabla_o H^1(\Gamma)$ is closed in $\mathcal{L}^2(\Gamma)$. This proves the following lemma.

LEMMA 3.8. *With $\mathcal{D}_0(\Gamma) := \{U \in \mathcal{L}^2_\tau(\Gamma) \colon \operatorname{div}_o U = 0\}$ we have the following orthogonal decomposition:*

$$\mathcal{L}^2_\tau(\Gamma) = \mathcal{D}_0(\Gamma) \oplus \nabla_o H^1(\Gamma).$$

For $u \in H^t$ $(t > 3/2)$, we may define

$$\overset{\bullet}{\partial}\, u := \langle \mathcal{T} \nabla u,\, \nu \rangle \in H^{t-3/2}(\Gamma).$$

In particular,

$$\mathcal{Y} := \mathcal{Y}_s := \overset{\bullet}{H}{}^s(\Gamma) \cap \{\overset{\bullet}{\partial}\, u \colon \nabla u \in \mathcal{H}_D\}^\perp \qquad \text{(orthogonal complement in } L^2(\Gamma))$$

is a well-defined closed subspace of $\overset{\bullet}{H}{}^s(\Gamma)$ for $s \geq 0$.

LEMMA 3.9. *Let $s \in [\,1/2\,,\, \infty\,)$. Then*

$$\begin{array}{rccl} G_1 & : & \mathcal{Y}_s & \longrightarrow \quad \mathcal{X}_s := \mathcal{H}^{s-1/2} \cap \left(\nabla H^1 \ominus (\nabla \overset{0}{H}{}^1 \oplus \mathcal{H}_D) \right), \\ & & g & \longmapsto \qquad\qquad \nabla u \end{array}$$

is a topological isomorphism. In particular,

$$\nabla H^1 = \mathcal{X}_{1/2} \oplus \mathcal{H}_D \oplus \nabla \overset{0}{H}{}^1. \tag{9}$$

Proof. Let $g \in \overset{\bullet}{H}{}^s(\Gamma)$ and $c \in K(\Gamma)$ and define $\nabla v := Gg$ and $\nabla u := Gc$. By Green's formula,

$$\langle \nabla v, \nabla u \rangle = \langle g, \overset{\bullet}{\partial}\, u \rangle_\Gamma,$$

which implies (using Lemma 3.4)

$$G\mathcal{Y}_s \subset \mathcal{X}_s \quad \text{and} \quad G^{-1}\mathcal{X}_s \subset \mathcal{Y}_s$$

and thus our assertion. □

4. Boundary controllability. We now turn to the question of boundary controllability:

Given $(E^0, H^0), (E^1, H^1) \in \mathcal{L}^2 \times \mathcal{L}^2$ *and* $T \in \mathbb{R}^+$, *can we find* $J \in \mathcal{U}$ *such that* $(E(T), H(T)) = (E^1, H^1)$ *for the solution* (E, H) *of* $\mathcal{P}(E^0, H^0; J)$ *in* $I := [\,0\,,\,T\,]$*?*

If this is possible we say that "the state (E^0, H^0) can be steered into (E^1, H^1) in time T." In the special case $(E^1, H^1) = (0, 0)$ we say that "the state (E^0, H^0) can be steered to rest." It will be decisive for the analysis of this problem to decompose (E^0, H^0) according to the material of the preceding section. So we write

$$E^0 = \nabla \wedge D + \underline{U} + \underline{\nabla e}, \tag{10}$$

$$H^0 = \nabla \wedge B + W + \nabla v + \underline{V} + \underline{\nabla h} \tag{11}$$

with uniquely determined

$$D \in (\nabla \wedge \overset{0}{\mathcal{R}}) \cap \mathcal{R}, \quad U \in \mathcal{H}_D, \quad e \in \overset{0}{H}{}^1,$$
$$B \in (\nabla \wedge \mathcal{R}) \cap \overset{0}{\mathcal{R}}, \quad W \in \mathcal{H}_N$$

and (cf. Lemma 3.9)

$$\nabla v = Gg \subset \mathcal{X}_{1/2}(\Omega) \quad (g \in H^{1/2}(\Gamma)), \quad V \in \mathcal{H}_D, \quad h \in \overset{0}{H}{}^1.$$

We start with a negative answer.

THEOREM 4.1. *Let* $\hat{U} \in \mathcal{H}_D$. *Then* $\langle E, \hat{U} \rangle$ *and* $\langle H, \hat{U} \rangle$ *as well as* $\operatorname{div} E$ *and* $\operatorname{div} H$ *are not influenced by our boundary controls. More precisely: Let* $J \in \mathcal{U}$ *and* $(E^0, H^0) \in \mathcal{L}^2 \times \mathcal{L}^2$. *Then, for all* t,

$$\langle E(t), \hat{U} \rangle = \langle E^0, \hat{U} \rangle,$$
$$\langle H(t), \hat{U} \rangle = \langle H^0, \hat{U} \rangle,$$
$$\operatorname{div} E(t) = \operatorname{div} E^0,$$
$$\operatorname{div} H(t) = \operatorname{div} H^0$$

for the solution (E, H) *to* $\mathcal{P}(E^0, H^0; J)$ *in* I.

In particular, no state (E^0, H^0) *for which any of the underlined components in* (10), (11) *is different from zero can be steered to rest in any time.*

Proof. The states $(\hat{U}, 0)$ and $(0, \hat{U})$ belong to $N(A)$. Therefore, by Lemma 2.3,

$$\langle E(t), \hat{U}\rangle - \langle E^0, \hat{U}\rangle = 0,$$
$$\langle H(t), \hat{U}\rangle - \langle H^0, \hat{U}\rangle = \int_0^t \int_\Gamma \langle J(s,y), \nu(y) \wedge (\nu(y) \wedge \hat{U}(y))\rangle d\mathbf{o}(y) ds.$$

But the integral vanishes because $\nu \wedge \hat{U} = 0$. This proves the first two assertions. The next two are proved analogously with the aid of Lemma 2.3 by looking at

$$(\nabla\varphi, 0), (0, \nabla\varphi) \in N(A) \quad \text{for} \quad \varphi \in C_0^\infty.$$

Finally, the negative conclusion follows from (10),(11) and Lemma 3.1. □

REMARK 3. *This result disproves previous claims in the literature which do not assume star-shapedness* [14, *section* 4.3], [17].

For a positive result it is necessary to require that the underlined components (which in what follows will be called "noncontrollable components") in (10), (11) should vanish. Furthermore, in order to keep the analytical efforts low, we require some additional regularity of the initial states. Thus we introduce the following spaces of states:

$$\mathcal{S}_2 := (\nabla \wedge \mathcal{R}) \times \left((\nabla \wedge \overset{0}{\mathcal{R}}) \oplus \mathcal{H}_N \oplus \mathcal{X}_{1/2}(\Omega) \right),$$
$$\mathcal{S}_1 := (\nabla \wedge \mathcal{R}) \times \left((\nabla \wedge \overset{0}{\mathcal{R}}) \oplus \mathcal{H}_N \right),$$
$$\mathcal{S}_0 := (\nabla \wedge \mathcal{R}) \times (\nabla \wedge \overset{0}{\mathcal{R}}),$$
$$\mathcal{S}_k^1 := \left(\mathcal{H}^1 \times \mathcal{H}^1\right) \cap \mathcal{S}_k, \quad k = 0, 1, 2,$$
$$\mathcal{S}_k^D := D(A) \cap \mathcal{S}_k^1, \quad k = 0, 1, 2.$$

We are ready to formulate our main result.

MAIN THEOREM. *Let $T > \operatorname{diam}(\Omega)$. Then each state in $\mathcal{S}_2^1$ can be steered to rest in time T. In particular, each divergence-free state can be steered to rest in time T if $\partial\Omega$ is connected.*

This will be proved in four steps: The initial state $(E^0, H^0) \in \mathcal{S}_2^1$ may not be an element of $D(A)$. First we show that it takes arbitrarily short time to steer it into a state which *is* in $\mathcal{S}_2^D$. In the second and third steps we remove extra components of H^0 by steering into $\mathcal{S}_1^D$ and then into $\mathcal{S}_0^D$, again in arbitrarily short time. Finally, granted time $T > \operatorname{diam}(\Omega)$, the resulting state $(E^0, H^0) \in \mathcal{S}_0^D$ may be steered to rest by using the trick invented by D. L. Russell for the wave equation.

THEOREM 4.2. *Each state $(E^0, H^0) \in \mathcal{S}_2^1$ may be steered into a state in $\mathcal{S}_2^D$ in arbitrarily short time.*

Proof. Given $T \in \mathbb{R}^+$ choose $\chi \in C^\infty(\mathbb{R})$ such that $\operatorname{supp}\chi \subset (-\infty, T)$ and $\chi(0) = 1$. Then

$$J(s,y) := \chi(s) \cdot \mathcal{T}_\tau E^0(y), \quad (s,y) \in M,$$

defines an element of $\mathcal{U}$. Furthermore, J is represented by $\hat{J} \in C^\infty(I, \mathcal{H}^1)$, where

$$\hat{J}(s,x) = \chi(s) \cdot E^0(x), \quad (s,x) \in Z.$$

For $t = T$, the components of the solution formula (3) either vanish by construction or belong to $D(A)$ by Lemmas 2.1 and 2.4. Furthermore, according to Theorem 4.1, no noncontrollable components have been introduced by this control process.

It remains to be shown that the state stays in $\mathcal{H}^1 \times \mathcal{H}^1$: From $E(T) \in \overset{0}{\mathcal{R}} \cap \mathcal{D}_0$ we obtain $E(T) \in \mathcal{H}^1$ [15, Thm. 8.6].

Concerning $H(T)$, we decompose

$$H^0 = \nabla \wedge B + V + \nabla h, \quad B \in \overset{0}{\mathcal{R}}, \quad V \in \mathcal{H}_N, \quad h \in H^1.$$

Lemma 3.3 shows that additionally $B \in \mathcal{H}^2$ and therefore $h \in H^2$ (because $V \in \mathcal{C}^\infty(\overline{\Omega})$ by Lemma 3.6). We apply Lemma 2.3 with $(\Phi^0, \Psi^0) := (0, \nabla\psi)$ and an arbitrary $\psi \in H^1$ to obtain

$$\langle H(T) - H^0, \nabla\psi\rangle = -\int_0^T \chi(s) \left(\int_\Gamma \langle \mathcal{T}_\tau E^0(y), \nabla\psi(y)\rangle d\mathbf{o}(y) \right) ds.$$

Assuming additionally—as we may—that $\int_0^T \chi(s)ds = 0$, we see that in the decomposition

$$H(T) = \nabla \wedge \tilde{B} + \tilde{V} + \nabla h, \quad \tilde{B} \in \overset{0}{\mathcal{R}}, \quad \tilde{V} \in \mathcal{H}_N,$$

the third component remains unchanged in $\mathcal{H}^1$, whereas the second is still in $\mathcal{C}^\infty(\overline{\Omega})$ by Lemma 3.6. The first belongs to $\mathcal{D}_0 \cap \overset{0}{\mathcal{D}}$ and also to $\mathcal{R}$ (because of $(E(T), H(T)) \in D(A)$). This implies $\nabla \wedge \tilde{B} \in \mathcal{H}^1$ (by the analogue of [15, Thm. 8.6] or cf. [9, Lem. 4.2 and section 8]) and hence $H(T) \in \mathcal{H}^1$. □

THEOREM 4.3. *Each state $(E^0, H^0) \in \mathcal{S}_2^D$ may be steered into a state in $\mathcal{S}_1^D$ in arbitrarily short time.*

Proof. We have

$$H^0 = \nabla \wedge B + W + X, \quad B \in \overset{0}{\mathcal{R}}, \quad W \in \mathcal{H}_N, \quad X = \nabla v \in \mathcal{X}_{1/2}.$$

The sesquilinear form

$$B(h, \varphi) := \langle \nabla_o h, \nabla_o \varphi\rangle_\Gamma$$

is continuous and positive semidefinite on $H^1(\Gamma)$ with null-space $K(\Gamma)$. Furthermore, $\mathcal{Y}_1$ and $K(\Gamma)$ are complementary closed subspaces of $H^1(\Gamma)$ by Lemmas 3.4, 3.5, and 3.9. A standard compactness argument (using the compact embedding of $\overset{\bullet}{H}{}^1(\Gamma)$ in $L^2(\Gamma)$) shows that B is strictly coercive on $\mathcal{Y}_1$. Therefore, the problem to find $h \in \mathcal{Y}_1$ such that

$$\langle \nabla_o h, \nabla_o \varphi\rangle_\Gamma = -\langle X, G\varphi\rangle \quad \text{for all} \quad \varphi \in \mathcal{Y}_1$$

is uniquely solvable. Lemma 3.4 shows that the right-hand side defines a continuous antilinear functional on $H^{1/2}(\Gamma)$ and thus may be considered as an element of $H^{-1/2}(\Gamma)$. Therefore, elliptic regularity theory tells us that the solution h belongs to $H^{3/2}(\Gamma)$ and thus to $\mathcal{Y}_{3/2}$. (In fact, a closer look using $H^0 \in \mathcal{H}^1$ would show that $h \in H^{5/2}(\Gamma)$.) Let $\chi \in C^\infty(\mathbb{R})$ with

$$\operatorname{supp}\chi \subset (\,0\,,\,T\,) \quad \text{and} \quad \int_0^T \chi(t)dt = 1.$$

Then

$$J(t,y) := -\chi(t)\nabla_o h(y), \quad (t,y) \in M,$$

defines an element of $\mathcal{U}$. By the trace theorem, we may extend $\nu \wedge \nabla_o h \in \mathcal{H}^{1/2}(\Gamma)$ to some $j \in \mathcal{H}^1$ such that J is represented by

$$\hat{J}(t,x) := \chi(t) j(x), \quad (t,x) \in Z,$$

which satisfies

$$\hat{J} \in C^\infty([\,0\,,\,T\,],\mathcal{R}), \quad \hat{J}(0,\cdot) = \hat{J}(T,\cdot) = 0. \tag{12}$$

Let (E,H) be the solution to $\mathcal{P}(E^0,H^0;J)$. Theorem 4.1 implies that $(E(T),H(T))$ does not contain any noncontrollable components and (12) and Lemma 2.4 show that $(E(T),H(T)) \in D(A)$.

Let us apply Lemma 2.3 with $(\Phi^0,\Psi^0) := (0,G\varphi)$,$\varphi \in \mathcal{Y}_{3/2}$ (hence $G\varphi \in \mathcal{H}^1$). We compute

$$\begin{aligned}
&\langle H(T),\, G\varphi\rangle \\
&\quad = \langle H^0,\, G\varphi\rangle + \int_0^T \chi(t)\langle \nabla_o h,\, \mathcal{T}G\varphi\rangle_\Gamma dt \\
&\quad = \langle X,\, G\varphi\rangle - \langle \nu\wedge(\nu\wedge\nabla_o h),\, \mathcal{T}G\varphi\rangle_\Gamma \\
&\quad = \langle X,\, G\varphi\rangle - \langle \nabla_o h,\, \nu\wedge(\nu\wedge\mathcal{T}G\varphi)\rangle_\Gamma \\
&\quad = \langle X,\, G\varphi\rangle + \langle \nabla_o h,\, \nabla_o\varphi\rangle_\Gamma = 0.
\end{aligned}$$

We conclude (using Lemma 3.9 as well as the fact that $\mathcal{Y}_{3/2}$ is dense in $\mathcal{Y}_{1/2}$)

$$H(T) = \nabla\wedge\tilde{B} + \tilde{W} \in \mathcal{R}, \quad \tilde{B} \in \overset{0}{\mathcal{R}}, \quad \tilde{W} \in \mathcal{H}_N.$$

Thus $H(T) \in \mathcal{H}^1$ from Lemma 3.6 and [27, Thm. 2.2] and $(E(T),H(T)) \in \mathcal{S}_1^D$. ∎

THEOREM 4.4. *Each state $(E^0,H^0) \in \mathcal{S}_1^D$ may be steered into a state in $\mathcal{S}_0^D$ in arbitrarily short time by a control $J \in \mathcal{C}_\tau^\infty(\Gamma)$.*

Proof. Let π denote the orthogonal projector in $\mathcal{L}_\tau^2(\Gamma)$ onto $\mathcal{D}_0(\Gamma)$ along $\nabla_o H^1(\Gamma)$. Then

$$\sigma(V,W) := \langle \pi\mathcal{T}V,\, \pi\mathcal{T}W\rangle_\Gamma$$

is a true scalar product on $\mathcal{H}_N$. Namely, if $\pi\mathcal{T}V = 0$, then we have

$$\mathcal{T}V = \nabla_o g, \quad g \in C^\infty(\Gamma).$$

We solve Dirichlet's problem

$$\Delta u = 0, \quad u|_\Gamma = g$$

and find by direct computation

$$V - \nabla u \in \mathcal{H}_D.$$

Therefore, Lemma 3.5 implies $V \in \nabla H^1$ and thus $V = 0$ by Lemma 3.1.

So we can choose a basis $\{W_k\colon\ k=1,\dots,\beta_2\}$ of $\mathcal{H}_N$ which is orthonormal with respect to σ. We choose the control $J\in\mathcal{C}^\infty_\tau(M)$ defined by

$$J(t,y):=\chi(t)\cdot\sum_{k=1}^{\beta_2}\langle H^0,\,W_k\rangle\pi\mathcal{T}W_k(y),\qquad (t,y)\in M,$$

with χ as in the preceding proof and denote the solution to $\mathcal{P}(E^0,H^0;J)$ by (E,H). Again (by Theorem 4.1 and Lemma 2.4) no noncontrollable components are introduced and $(E(T),H(T))\in D(A)$. So $(E(T),H(T))\in\mathcal{S}_0^D$ as desired if we can show (recalling Lemma 3.9 and the regularity result [27, Thm. 2.2])

$$\langle H(T),\,W_l\rangle=0\quad\text{for all}\quad l\in\{1,\dots,\beta_2\},\tag{13}$$

$$\langle H(T),\,G\varphi\rangle=0\quad\text{for all}\quad \varphi\in\mathcal{Y}_{3/2}.\tag{14}$$

For (13) we apply Lemma 2.3 with $(\Phi^0,\Psi^0):=(0,W_l)$ and compute

$$\begin{aligned}&\langle H(T)-H^0,\,W_l\rangle\\&\quad=-\sum_k\langle H^0,\,W_k\rangle\langle\pi\mathcal{T}W_k,\,\mathcal{T}W_l\rangle_\Gamma\\&\quad=-\sum_k\langle H^0,\,W_k\rangle\langle\pi\mathcal{T}W_k,\,\pi\mathcal{T}W_l\rangle_\Gamma=-\langle H^0,\,W_l\rangle.\end{aligned}$$

For (14) we use Lemma 2.3 with $(\Phi^0,\Psi^0):=(0,G\varphi)$ and compute

$$\begin{aligned}&\langle H(T),\,G\varphi\rangle\\&\quad=\langle H^0,\,G\varphi\rangle-\sum_k\langle H^0,\,W_k\rangle\langle\pi\mathcal{T}W_k,\,\mathcal{T}G\varphi\rangle_\Gamma\\&\quad=0+\sum_k\langle H^0,\,W_k\rangle\langle\pi\mathcal{T}W_k,\,\nu\wedge(\nu\wedge\mathcal{T}G\varphi)\rangle_\Gamma\\&\quad=-\sum_k\langle H^0,\,W_k\rangle\langle\pi\mathcal{T}W_k,\,\nabla_o\varphi\rangle_\Gamma=0\end{aligned}$$

because $\pi\mathcal{T}W_k\in\big(\nabla_oH^1(\Gamma)\big)^\perp$. □

THEOREM 4.5. *Let $T>\operatorname{diam}(\Omega)$ and $I:=[\,0\,,\,T\,]$. Then each $(E^0,H^0)\in\mathcal{S}_0^D$ can be steered to rest in time T by a control $J\in C^0(I,\mathcal{H}_\tau^{1/2}(\Gamma))\cap\mathcal{U}$.*

Proof. We have

$$(E^0,H^0)=(\nabla\wedge D,\nabla\wedge B)\in\mathcal{H}^1\times\mathcal{H}^1,\qquad (D,B)\in\mathcal{R}\times\overset{0}{\mathcal{R}},$$

and additionally (in view of Lemmas 3.2 and 3.3) $D,B\in\mathcal{H}^2$. We use Calderon's extension theorem to extend D,B to $\hat{D},\hat{B}\in\mathcal{H}^2(\mathbb{R}^3)$ and with a cutoff technique we may assume that $\hat{D}$, $\hat{B}$ are supported in

$$\Omega_\rho:=\Omega+U(0,\rho),\qquad \rho:=T-\operatorname{diam}(\Omega).$$

Let $(\hat{E},\hat{H})$ be the solution of the Cauchy problem

$$\left.\begin{aligned}\partial_t\hat{E}&=+\nabla_x\wedge\hat{H}\\ \partial_t\hat{H}&=-\nabla_x\wedge\hat{E}\end{aligned}\right\}\quad\text{in}\quad[\,0\,,\,\infty\,)\times\mathbb{R}^3,$$

$$(\hat{E}(0,\cdot),\hat{H}(0,\cdot))=(\nabla\wedge\hat{D},\nabla\wedge\hat{B}).$$

We have $\operatorname{div} E(t,\cdot) = 0$ because $\operatorname{div} E(0,\cdot) = 0$. Using $\Delta = -\mathrm{curlcurl} + \nabla\mathrm{div}$, we may look upon $\hat{E}$ as the solution to a Cauchy problem for the wave equation:

$$\begin{aligned} \partial_t^2 \hat{E} - \Delta_x \hat{E} &= 0, \\ \hat{E}(0,\cdot) &= \nabla\wedge \hat{D} \in \mathcal{H}^1, \\ \partial_t \hat{E}(0,\cdot) &= \nabla \wedge (\nabla\wedge \hat{B}) \in \mathcal{H}^0. \end{aligned}$$

The regularity of the initial data implies that $\hat{E}$ is a "solution with finite energy" (cf. [31, p. 20]) and therefore $\hat{E} \in C^0([\,0\,,\infty\,), \mathcal{H}^1(\mathbb{R}^3)) \cap \mathcal{H}^1(Z)$. An application of the trace theorem shows

$$J := \mathcal{T}_\tau \hat{E} \in C^0(I, \mathcal{H}^{1/2}(\Gamma)) \cap \mathcal{U}.$$

Furthermore, by Huygens' principle [15, section 5.3],

$$\operatorname{supp}(\hat{E}(T,\cdot), \hat{H}(T,\cdot)) \cap \Omega = \emptyset.$$

Restricting everything to $[\,0\,,T\,] \times \overline{\Omega}$ shows that J steers (E^0, H^0) into rest. □

REMARK 4. *Replacing the trace theorem by the more sophisticated regularity result of* [24] *would enable us to work with controls in the space* $\mathcal{H}_\tau^{3/4}$.

REMARK 5. *Consider the class of regions* Ω *for which there exists an open segment*

$$S := \{\hat{x} + \tau p\colon\ \tau \in (\,0\,,L\,)\}, \quad \hat{x} \in \overline{\Omega}, \quad p \in S^2,$$

such that $S \subset \Omega$ *and length*$(S) = L = \operatorname{diam}\Omega$. *Then one can show that our Main Theorem is almost optimal in the sense that for* $T < \operatorname{diam}(\Omega)$, *even in* $\mathcal{S}_0^D$ *there are states which cannot be steered to rest in time* T. *This may be shown using the argument of* [7, *Thm.* 2.1] *and the sequence* u_k *from* [2, *Thm.* 3.2] *on a ray* $\gamma \subset [\,0\,,T\,] \times \Omega$ *corresponding to* S. *Namely, defining*

$$W_k := \nabla_x \wedge \begin{bmatrix} u_k \\ 0 \\ 0 \end{bmatrix}, \quad E_k := \nabla_x \wedge W_k, \quad H_k := \partial_t W_k$$

yields a sequence which contradicts exact controllability of $\mathcal{S}_0^D$.

REMARK 6. *Our results show that more precise topological criteria ("Betti numbers") rather than star-shapedness decide about null-controllability of divergence-free states. In the case of Remark* 1(i) *("spherical shell") the initial state* $(\nabla u, 0)$ *is not null-controllable whereas in case* (ii) *("torus") each divergence-free initial state is null-controllable because for this region* $\mathcal{H}_D = \{0\}$.

REMARK 7. *Our methods also allow us to obtain "partial controllability" results for problems where the control* J *is only allowed to be supported on part of the boundary (cf.* [21], [2], [17]*).*

REFERENCES

[1] A. V. BALAKRISHNAN, *Applied Functional Analysis*, Springer, New York, 1976.

[2] C. BARDOS, G. LEBEAU, AND J. RAUCH, *Sharp sufficient conditions for the observation, control and stabilization of waves from the boundary*, SIAM J. Control Optim., 30 (1992), pp. 1024–1065.

[3] H. Barucq and B. Hanouzet, *Asymptotic behavior of solutions to Maxwell's system in bounded domains with absorbing Silver–Müller's condition on the exterior boundary*, Asymptotic Anal., 15 (1997), pp. 25–40.
[4] H. Barucq, F. Delaurens, and B. Hanouzet, *Method of absorbing boundary conditions: phenomena of error stabilization*, SIAM J. Numer. Anal., 35 (1998), pp. 1113–1129.
[5] A. Bensoussan, *Some remarks on the exact controllability of Maxwell's equations*, Acta Appl. Math., 20 (1990), pp. 197–229
[6] M. Cessenat, *Mathematical Methods in Electromagnetism: Linear Theory and Applications*, Ser. Adv. Math. Appl. Sci. 41, World Scientific, Singapore, 1996.
[7] S. Dolecki and D.L. Russell, *A general theory of observation and control*, SIAM J. Control Optim., 158 (1977), pp. 185–220.
[8] G. Duvaut and J.L. Lions, *Inequalities in Mechanics and Physics*, Springer, New York, 1976.
[9] K.O. Friedrichs, *Differential forms on Riemannian manifolds*, Comm. Pure Appl. Math., 8 (1955), pp. 551–590.
[10] B.V. Kapitonov, *Stabilization and exact boundary controllability for Maxwell's equations*, SIAM J. Control Optim., 32 (1994) pp. 408–420.
[11] K.A. Kime, *Boundary controllability of Maxwell's equations in a spherical region*, SIAM J. Control Optim., 28 (1990), pp. 294–319.
[12] V. Komornik, *Boundary stabilization, observation and control of Maxwell's equations*, Panamer. Math. J., 4 (1994), pp. 47–61.
[13] O.A. Ladyzhenskaya and V.A. Solonnikov, *The linearization principle and invariant manifolds for problems of magnetohydrodynamics*, J. Soviet Math. 8 (1977), pp. 384–422.
[14] J.E. Lagnese, *Exact boundary controllability of Maxwell's equations in a general region*, SIAM J. Control Optim., 27 (1989), pp. 374-388.
[15] R. Leis, *Initial Boundary Value Problems in Mathematical Physics*, Teubner, Stuttgart, 1986.
[16] A. Milani and R. Picard, *Decomposition theorems and their application to non-linear electro- and magneto-static boundary value problems*, in Partial Differential Equations and Calculus of Variations, St. Hildebrandt and R. Leis, eds., Lecture Notes in Math. 1357, Springer, New York, 1988, pp. 317–340.
[17] O. Nalin, *Contrôlabilité exacte sur une partie du bord des équations de Maxwell*, C. R. Acad. Sci. Paris Sér. I Math., 309 (1989), pp. 811–815.
[18] R. Picard, *Zur Theorie der harmonischen Differentialformen*, Manuscripta Math., 27 (1979), pp. 31–45.
[19] R. Picard, *Randwertaufgaben in der verallgemeinerten Potentialtheorie*, Math. Methods Appl. Sci., 3 (1981), pp. 218–228.
[20] D.L. Russell, *Boundary value control of the higher dimensional wave equation*, SIAM J. Control Optim., 9 (1971), pp. 29–42.
[21] D.L. Russell, *Boundary value controllability of wave and heat processes in star-complemented regions*, in Proc. Conference on Differential Games and Control Theory, E.O. Roxin, P.-T. Liu, and R.L. Sternberg, eds., Marcel Dekker, New York, 1974.
[22] D.L. Russell, *A unified boundary controllability theory for hyperbolic and parabolic partial differential equations*, Stud. Appl. Math., 52 (1973), pp. 189–211.
[23] D.L. Russell, *The Dirichlet–Neumann boundary control problem associated with Maxwell's equations in cylindrical region*, SIAM J. Control Optim., 24 (1986), pp. 199–229.
[24] D. Tataru, *On the regularity of boundary traces for the wave equation*, Ann. Scuola Norm. Sup. Pisa, 26 (1998), pp. 185–206.
[25] M.E. Taylor, *Partial Differential Equations* I—*Basic Theory*, Springer, New York, 1996.
[26] C. Weber, *A local compactness theorem for Maxwell's equations*, Math. Methods Appl. Sci., 2 (1980), pp. 12–25.
[27] C. Weber, *Regularity theorems for Maxwell's equations*, Math. Methods Appl. Sci., 3 (1981), pp. 523–536.
[28] N. Weck, *Maxwell's boundary value problem on Riemannian manifolds with nonsmooth boundaries*, J. Math. Anal. Appl., 46 (1974), pp. 410–437.
[29] H. Weyl, *Die natürlichen Randwertaufgaben im Außenraum für Strahlungsfelder beliebiger Dimension und beliebigen Ranges*, Math. Z., 56 (1952), pp. 105–119.
[30] K.J. Witsch, *A remark on a compactness result in electromagnetic theory*, Math. Methods Appl. Sci., 16 (1993), pp. 123–129.
[31] C.H. Wilcox, *Scattering for the d'Alembert Equation in Exterior Domains*, Springer, New York, 1975.

SIAM J. CONTROL OPTIM.
Vol. 38, No. 3, pp. 751-766

FINITE-TIME STABILITY OF CONTINUOUS AUTONOMOUS SYSTEMS*

SANJAY P. BHAT† AND DENNIS S. BERNSTEIN‡

Abstract. Finite-time stability is defined for equilibria of continuous but non-Lipschitzian autonomous systems. Continuity, Lipschitz continuity, and Hölder continuity of the settling-time function are studied and illustrated with several examples. Lyapunov and converse Lyapunov results involving scalar differential inequalities are given for finite-time stability. It is shown that the regularity properties of the Lyapunov function and those of the settling-time function are related. Consequently, converse Lyapunov results can only assure the existence of continuous Lyapunov functions. Finally, the sensitivity of finite-time-stable systems to perturbations is investigated.

Key words. stability, finite-time stability, non-Lipschitzian dynamics

AMS subject classifications. 34D99, 93D99

PII. S0363012997321358

1. Introduction.

The object of this paper is to provide a rigorous foundation for the theory of finite-time stability of continuous autonomous systems and motivate a closer examination of finite-time stability as a possible objective in control design.

Classical optimal control theory provides several examples of systems that exhibit convergence to the equilibrium in finite time [17]. A well-known example is the double integrator with bang-bang time-optimal feedback control [2]. These examples typically involve dynamics that are discontinuous. Discontinuous dynamics, besides making a rigorous analysis difficult (see [9]), may also lead to chattering [10] or excite high frequency dynamics in applications involving flexible structures. Reference [8] considers finite-time stabilization using time-varying feedback controllers. However, it is well known that the stability analysis of time-varying systems is more complicated than that of autonomous systems. Therefore, with simplicity as well as applications in mind, we focus on continuous autonomous systems.

Finite-settling-time behavior of systems with continuous dynamics is considered in [3], [4], [11], [19], [21]. However, a detailed analysis of such systems has not been carried out. In particular, a precise formulation of finite-time stability is lacking, while little is known about the settling-time function. Furthermore, while references [3], [4], [11], [19] present Lyapunov conditions for finite-time stability, neither rigorous proofs nor converse results can be found. Reference [21] suggests, based on a scalar example, that systems with finite-settling-time dynamics possess better disturbance rejection and robustness properties. However, no precise results exist for multidimensional systems. This paper attempts to fill these gaps.

In section 2, we define finite-time stability for equilibria of continuous autonomous systems that have unique solutions in forward time. Continuity and forward uniqueness render the solutions continuous functions of the initial conditions, so that the

*Received by the editors May 12, 1997; accepted for publication (in revised form) March 30, 1999; published electronically February 29, 2000.

http://www.siam.org/journals/sicon/38-3/32135.html

†Department of Aerospace Engineering, Indian Institute of Technology, Powai, Mumbai 400076, India (bhat@aero.iitb.ernet.in).

‡Department of Aerospace Engineering, The University of Michigan, Ann Arbor, MI 48109-2140 (dsbaero@engin.umich.edu). The research of this author was supported in part by Air Force Office of Scientific Research grant F49620-95-1-0019.

solutions define a continuous semiflow on the state space. Uniqueness also makes it possible to define the settling-time function. Certain useful properties of the settling time function are established. It is shown by example that it is possible for the settling time to be unbounded in every neighborhood of the origin even if all solutions converge to the origin in finite time. A different example shows that the settling-time function may be continuous without being Hölder continuous at the origin.

In section 3 we define finite-time repellers (called terminal repellers in [7], [23]), which are a special class of unstable equilibria that arise only in non-Lipschitzian systems. We show that a system having a finite-time repeller possesses multiple solutions starting at the finite-time repeller.

In section 4, we give a Lyapunov theorem for finite-time stability. Dini derivatives are used since Lyapunov functions are assumed to be only continuous. A converse result is shown to hold under the assumption that the settling-time function is continuous. In general, the converse result cannot be strengthened in its conclusion regarding the regularity of the Lyapunov function; that is, a system with a finite-time-stable equilibrium may not admit a Hölder continuous Lyapunov function. This is because Hölder continuity of the Lyapunov function necessarily implies Hölder continuity of the settling-time function at the origin. On the other hand, as mentioned above, there exist finite-time-stable systems with settling-time functions that are not Hölder continuous at the origin.

The existence of a Hölder continuous Lyapunov function assumes importance in section 5 where we investigate the sensitivity of stability properties to perturbations of systems with a finite-time-stable equilibrium under the assumption of the existence of a Lipschitz continuous Lyapunov function. Both persistent and vanishing perturbations are considered. It is shown that under certain conditions, finite-time-stable systems may exhibit better rejection of bounded persistent disturbances than Lipschitzian exponentially stable systems. It is also shown that finite-time stability is preserved under perturbations that are Lipschitz in the state.

2. Finite-time stability. Let $\|\cdot\|$ denote a norm on $\mathbb{R}^n$. The notions of openness, convergence, continuity, and compactness that we use refer to the topology generated on $\mathbb{R}^n$ by the norm $\|\cdot\|$. We use $\overline{\mathbb{R}}$, $\mathbb{R}_+$, and $\overline{\mathbb{R}}_+$ to denote the extended, nonnegative, and extended nonnegative, real numbers, respectively. We also use $\overline{\mathcal{A}}$ and $\operatorname{bd}\mathcal{A}$ to denote the closure and the boundary of the set $\mathcal{A}$, respectively. We will call a set $\mathcal{A} \subset \mathbb{R}^n$ bounded if $\overline{\mathcal{A}}$ is compact. Finally, we denote the composition of two functions $U : \mathcal{A} \to \mathcal{B}$ and $V : \mathcal{B} \to \mathcal{C}$ by $V \circ U : \mathcal{A} \to \mathcal{C}$.

Consider the system of differential equations

$$\dot{y}(t) = f(y(t)), \tag{2.1}$$

where $f\colon \mathcal{D} \to \mathbb{R}^n$ is continuous on an open neighborhood $\mathcal{D} \subseteq \mathbb{R}^n$ of the origin and $f(0) = 0$. A continuously differentiable function $y\colon I \to \mathcal{D}$ is said to be a *solution* of (2.1) on the interval $I \subset \mathbb{R}$ if y satisfies (2.1) for all $t \in I$. The continuity of f implies that, for every $x \in \mathcal{D}$, there exist $\tau_0 < 0 < \tau_1$ and a solution $y(\cdot)$ of (2.1) defined on (τ_0, τ_1) such that $y(0) = x$ [12, Thm. I.1.1]. A solution y is said to be *right maximally defined* if y cannot be extended on the right (either uniquely or nonuniquely) to a solution of (2.1). Every solution of (2.1) has an extension that is right maximally defined [12, Thm. I.2.1]. For later use, we state the following result on bounded solutions of (2.1). For a proof, see [12, pp. 17–18] or [22, Thm. 3.3, p. 12].

PROPOSITION 2.1. *If $y : [0, \tau) \to \mathcal{D}$ is a right maximally defined solution of* (2.1) *such that $y(t) \in \mathcal{K}$ for all $t \in [0, \tau)$, where $\overline{\mathcal{K}} \subset \mathcal{D}$ is compact, then $\tau = \infty$.*

We will assume that (2.1) possesses unique solutions in forward time for all initial conditions except possibly the origin in the following sense: for every $x \in \mathcal{D}\backslash\{0\}$ there exists $\tau_x > 0$ such that, if $y_1 : [0, \tau_1) \to \mathcal{D}$ and $y_2 : [0, \tau_2) \to \mathcal{D}$ are two right maximally defined solutions of (2.1) with $y_1(0) = y_2(0) = x$, then $\tau_x \leq \min\{\tau_1, \tau_2\}$ and $y_1(t) = y_2(t)$ for all $t \in [0, \tau_x)$. Without any loss of generality, we may assume that for each x, τ_x is chosen to be the largest such number in $\overline{\mathbb{R}}_+$. In this case, we denote by $\psi(\cdot, x)$ or, alternatively, $\psi^x(\cdot)$ the unique solution of (2.1) on $[0, \tau_x)$ satisfying $\psi(0, x) = x$. Note that ψ^x cannot be extended on the right uniquely to a solution of (2.1) because if $\tau_x < \infty$, then as $t \to \tau_x$, either $\psi(t, x)$ approaches bd $\mathcal{D}$ [12, Thm. I.2.1], in which case ψ^x cannot be extended on the right to a solution of (2.1), or $\psi(t, x)$ approaches 0 with (2.1) having nonunique solutions starting at 0, in which case ψ^x can be extended on the right to a solution of (2.1) in more than one way. If (2.1) has nonunique solutions in forward time for the initial condition 0, then ψ is defined on a relatively open subset of $\mathbb{R}_+ \times \mathcal{D}\backslash\{0\}$ onto $\mathcal{D}\backslash\{0\}$. If (2.1) possesses a unique solution in forward time for the initial condition 0, then ψ is defined on a relatively open subset of $\mathbb{R}_+ \times \mathcal{D}$ onto $\mathcal{D}$ and for each $x \in \mathcal{D}$, $\psi^x : [0, \tau_x) \to \mathcal{D}$ is the unique right maximally defined solution of (2.1) for the initial condition x. Uniqueness in forward time and the continuity of f imply that ψ is continuous on its domain of definition [12, Thm. I.3.4] and defines a *local semiflow* [6], [20, Ch. 2] on $\mathcal{D}\backslash\{0\}$ or $\mathcal{D}$, as the case may be. Various sufficient conditions for forward uniqueness in the absence of Lipschitz continuity can be found in [1], [9, sect. 10], [14], [22, sect. 1].

DEFINITION 2.2. *The origin is said to be a* finite-time-stable equilibrium *of* (2.1) *if there exists an open neighborhood $\mathcal{N} \subseteq \mathcal{D}$ of the origin and a function $T\colon \mathcal{N}\backslash\{0\} \to (0, \infty)$, called the* settling-time function*, such that the following statements hold:*

(i) *Finite-time convergence: For every $x \in \mathcal{N}\backslash\{0\}$, ψ^x is defined on $[0, T(x))$, $\psi^x(t) \in \mathcal{N}\backslash\{0\}$ for all $t \in [0, T(x))$, and $\lim_{t\to T(x)} \psi^x(t) = 0$.*

(ii) *Lyapunov stability: For every open neighborhood $\mathcal{U}_\varepsilon$ of 0 there exists an open subset $\mathcal{U}_\delta$ of $\mathcal{N}$ containing 0 such that, for every $x \in \mathcal{U}_\delta\backslash\{0\}$, $\psi^x(t) \in \mathcal{U}_\varepsilon$ for all $t \in [0, T(x))$.*

The origin is said to be a globally finite-time-stable equilibrium *if it is a finite-time-stable equilibrium with $\mathcal{D} = \mathcal{N} = \mathbb{R}^n$.*

The following proposition shows that if the origin is a finite-time-stable equilibrium of (2.1), then (2.1) has a unique solution on $\mathbb{R}_+$ for every initial condition in an open neighborhood of 0, including 0 itself.

PROPOSITION 2.3. *Suppose the origin is a finite-time-stable equilibrium of* (2.1). *Let $\mathcal{N} \subseteq \mathcal{D}$ and let $T : \mathcal{N}\backslash\{0\} \to (0, \infty)$ be as in Definition* 2.2. *Then, ψ is defined on $\mathbb{R}_+ \times \mathcal{N}$ and $\psi(t, x) = 0$ for all $t \geq T(x)$, $x \in \mathcal{N}$, where $T(0) \triangleq 0$.*

Proof. It can be shown that Lyapunov stability of the origin implies that $y \equiv 0$ is the unique solution y of (2.1) satisfying $y(0) = 0$. This proves that $\mathbb{R}_+ \times \{0\}$ is contained in the domain of definition of ψ and $\psi^0 \equiv 0$.

Now, let $\mathcal{N} \subseteq \mathcal{D}$ and T be as in Definition 2.2 and let $x \in \mathcal{N}\backslash\{0\}$. Define

$$\begin{aligned} y(t) &= \psi(t, x), \quad 0 \leq t < T(x), \\ &= 0, \qquad T(x) \leq t. \end{aligned} \tag{2.2}$$

By construction, y is continuously differentiable on $\mathbb{R}_+\backslash\{T(x)\}$ and satisfies (2.1) on $\mathbb{R}_+\backslash\{T(x)\}$. Also, it follows from the continuity of f that

$$\lim_{t\to T(x)^-} \dot{y}(t) = \lim_{t\to T(x)^-} f(y(t)) = 0 = \lim_{t\to T(x)^+} \dot{y}(t),$$

so that y is continuously differentiable at $T(x)$ and satisfies (2.1). Thus y is a solution of (2.1) on $\mathbb{R}_+$. To prove uniqueness, suppose z is a solution of (2.1) on $\mathbb{R}_+$ satisfying $z(0) = x$. Then by the uniqueness assumption, y and z agree on $[0, T(x))$. By continuity, y and z must also agree on $[0, T(x)]$ so that $z(T(x)) = 0$. Lyapunov stability now implies that $z(t) = 0$ for $t > T(x)$. This proves uniqueness. By the definition of ψ, it follows that $\psi^x \equiv y$. Thus ψ^x is defined on $\mathbb{R}_+$ and satisfies $\psi^x(t) = 0$ on $[T(x), \infty)$ for every $x \in \mathcal{N}$. This proves the result. □

Proposition 2.3 implies that if the origin is a finite-time-stable equilibrium of (2.1), then the solutions of (2.1) define a continuous *global semiflow* [20] on $\mathcal{N}$; that is, $\psi : \mathbb{R}_+ \times \mathcal{N} \to \mathcal{N}$ is a (jointly) continuous function satisfying

$$\psi(0, x) = x, \tag{2.3}$$

$$\psi(t, \psi(h, x)) = \psi(t + h, x) \tag{2.4}$$

for every $x \in \mathcal{N}$ and $t, h \in \mathbb{R}_+$. In addition, ψ satisfies

$$\psi(T(x) + t, x) = 0 \tag{2.5}$$

for all $x \in \mathcal{N}$ and $t \in \mathbb{R}_+$.

Proposition 2.3 also indicates that it is reasonable to extend T to all of $\mathcal{N}$ by defining $T(0) = 0$. With a slight abuse of terminology, we will also call this extension the *settling-time function*. It is easy to see from Definition 2.2 that, for all $x \in \mathcal{N}$,

$$T(x) = \inf\{t \in \mathbb{R}_+ : \psi(t, x) = 0\}. \tag{2.6}$$

To illustrate finite-time stability, as well as for later use, we consider a scalar system with a finite-time-stable equilibrium.

Example 2.1. The right-hand side of the scalar system

$$\dot{y}(t) = -k\operatorname{sign}(y(t))|y(t)|^\alpha, \tag{2.7}$$

where $\operatorname{sign}(0) = 0$, $k > 0$, and $\alpha \in (0, 1)$, is continuous everywhere and locally Lipschitz everywhere except the origin. Hence every initial condition in $\mathbb{R}\backslash\{0\}$ has a unique solution in forward time on a sufficiently small time interval. The global semiflow for (2.7) is easily obtained by direct integration as

$$\begin{aligned} \mu(t, x) &= \operatorname{sign}(x)\left[|x|^{1-\alpha} - k(1-\alpha)t\right]^{\frac{1}{1-\alpha}}, && t < \tfrac{1}{k(1-\alpha)}|x|^{1-\alpha}, && x \neq 0, \\ &= 0, && t \geq \tfrac{1}{k(1-\alpha)}|x|^{1-\alpha}, && x \neq 0, \\ &= 0, && t \geq 0, && x = 0. \end{aligned} \tag{2.8}$$

It is clear from (2.8) that (i) in Definition 2.2 is satisfied with $\mathcal{D} = \mathcal{N} = \mathbb{R}$ and the settling-time function $T : \mathbb{R} \to \mathbb{R}_+$ given by

$$T(x) = \frac{1}{k(1-\alpha)}|x|^{1-\alpha}. \tag{2.9}$$

Lyapunov stability follows by considering, for instance, the Lyapunov function $V(x) = x^2$. Thus the origin is a globally finite-time-stable equilibrium for (2.7). Note that T is Hölder continuous but not Lipschitz continuous at the origin.

The following proposition investigates the properties of the settling-time function of a finite-time-stable system.

PROPOSITION 2.4. *Suppose the origin is a finite-time-stable equilibrium of* (2.1). *Let $\mathcal{N} \subseteq \mathcal{D}$ be as in Definition* 2.2 *and let $T : \mathcal{N} \to \mathbb{R}_+$ be the settling-time function. Then the following statements hold.*

(i) *If $x \in \mathcal{N}$ and $t \in \mathbb{R}_+$, then*

$$T(\psi(t,x)) = \max\{T(x) - t, 0\}. \tag{2.10}$$

(ii) *T is continuous on $\mathcal{N}$ if and only if T is continuous at 0.*

(iii) *For every $r > 0$, there exists an open neighborhood $\mathcal{U}_r \subset \mathcal{N}$ of 0 such that, for every $x \in \mathcal{U}_r \backslash \{0\}$,*

$$T(x) > r\|x\|. \tag{2.11}$$

Proof. (i) The result follows from (2.6), (2.4), and (2.5).

(ii) Necessity is immediate. To prove sufficiency, suppose that T is continuous at 0.

Let $z \in \mathcal{N}$ and consider a sequence $\{z_m\}$ in $\mathcal{N}$ that converges to z. Let $\tau^- = \liminf_{m\to\infty} T(z_m)$ and $\tau^+ = \limsup_{m\to\infty} T(z_m)$. Note that both τ^- and τ^+ are in $\overline{\mathbb{R}}_+$ and

$$\tau^- \le \tau^+. \tag{2.12}$$

Next, let $\{z_l^+\}$ be a subsequence of $\{z_m\}$ such that $T(z_l^+) \to \tau^+$ as $l \to \infty$. The sequence $\{(T(z), z_l^+)\}$ converges in $\mathbb{R}_+ \times \mathcal{N}$ to $(T(z), z)$. By continuity and equation (2.5), $\psi(T(z), z_l^+) \to \psi(T(z), z) = 0$ as $l \to \infty$. Since T is assumed to be continuous at 0, $T(\psi(T(z), z_l^+)) \to T(0) = 0$ as $l \to \infty$. Using (2.10) with $t = T(z)$ and $x = z_l^+$, we obtain $\max\{T(z_l^+) - T(z), 0\} \to 0$ as $l \to \infty$. Thus $\max\{\tau^+ - T(z), 0\} = 0$, that is,

$$\tau^+ \le T(z). \tag{2.13}$$

Now, let $\{z_l^-\}$ be a subsequence of $\{z_m\}$ such that $T(z_l^-) \to \tau^-$ as $l \to \infty$. It follows from (2.12) and (2.13) that $\tau^- \in \mathbb{R}_+$. Therefore, the sequence $\{(T(z_l^-), z_l^-)\}$ converges in $\mathbb{R}_+ \times \mathcal{N}$ to (τ^-, z). Since ψ is continuous, it follows that $\psi(T(z_l^-), z_l^-) \to \psi(\tau^-, z)$ as $l \to \infty$. Equation (2.5) implies that $\psi(T(z_l^-), z_l^-) = 0$ for each l. Hence $\psi(\tau^-, z) = 0$ and, by (2.6),

$$T(z) \le \tau^-. \tag{2.14}$$

From (2.12), (2.13), and (2.14) we conclude that $\tau^- = \tau^+ = T(z)$ and hence $T(z_m) \to T(z)$ as $m \to \infty$.

(iii) Let $r > 0$. The function $\|f(\cdot)\|$ is continuous on $\mathcal{D}$ and $f(0) = 0$ so that the set $\Omega_r = \left\{x \in \mathcal{N} : \|f(x)\| < \frac{1}{r}\right\}$ is open and contains 0. By Lyapunov stability, there exists an open set $\mathcal{U}_r$ such that $0 \in \mathcal{U}_r \subset \mathcal{N}$ and $\psi(t,x) \in \Omega_r$ for every $t \in \mathbb{R}_+$ and $x \in \mathcal{U}_r$. Letting $x \in \mathcal{U}_r \backslash \{0\}$, we have

$$0 = \psi(T(x), x) = x + \int_0^{T(x)} f(\psi(t,x))dt,$$

so that

$$\|x\| = \left\| - \int_0^{T(x)} f(\psi(t,x))dt \right\| \le \int_0^{T(x)} \|f(\psi(t,x))\|dt < \frac{T(x)}{r},$$

which proves the result. ∎

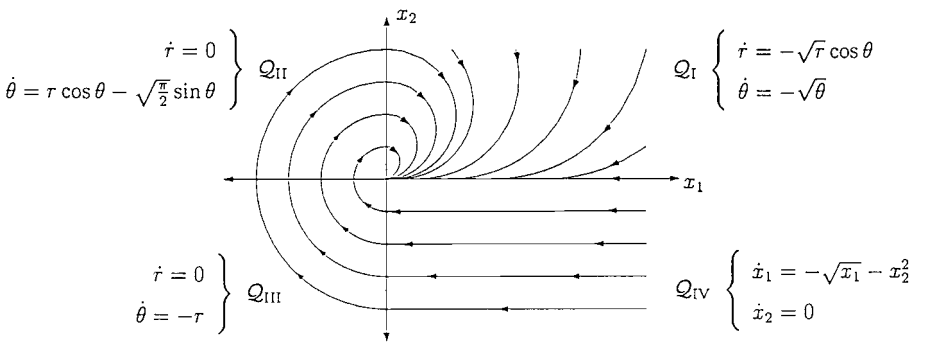


FIG. 2.1. *Finite-time stability with discontinuous settling-time function.*

Proposition 2.4 (ii) is significant because, in general, finite-time stability does not imply that the settling-time function T is continuous at the origin. Indeed, as the following example shows, the settling-time function can be unbounded in every neighborhood of the origin.

Example 2.2. Consider the system (2.1) where the vector field $f : \mathbb{R}^2 \to \mathbb{R}^2$ is defined on the quadrants

$$\begin{aligned} \mathcal{Q}_{\rm I} &= \{x \in \mathbb{R}^2 \backslash \{0\} : x_1 \geq 0, x_2 \geq 0\}, \quad \mathcal{Q}_{\rm II} = \{x \in \mathbb{R}^2 : x_1 < 0, x_2 \geq 0\}, \\ \mathcal{Q}_{\rm III} &= \{x \in \mathbb{R}^2 : x_1 \leq 0, x_2 < 0\}, \quad \mathcal{Q}_{\rm IV} = \{x \in \mathbb{R}^2 : x_1 > 0, x_2 < 0\}, \end{aligned}$$

as shown in Figure 2.1, with $f(0) = 0$, $r > 0$, $\theta \in [0, 2\pi)$, and $x = (x_1, x_2) = (r\cos\theta, r\sin\theta)$. It is easy to show that the vector field f is continuous on $\mathbb{R}^2$ and locally Lipschitz everywhere on $\mathbb{R}^2$ except on the positive x_1-axis, denoted by $\mathcal{X}_1^+$, the negative x_2-axis, denoted by $\mathcal{X}_2^-$, and the origin. Since the derivative of x_2^2 along the solutions of (2.1) is nonpositive in a sufficiently small neighborhood of every point $x \in \mathcal{X}_1^+$, every solution $y(\cdot)$ of (2.1) that satisfies $y(0) \in \mathcal{X}_1^+$ satisfies $y(t) \in \mathcal{X}_1^+$ for $t > 0$ sufficiently small, while on $\mathcal{X}_1^+$, f is simply given by $\dot{x}_1 = -\sqrt{x_1}$, $\dot{x}_2 = 0$ which is easily seen to have unique solutions for initial conditions in $\mathcal{X}_1^+$. In fact, by Example 2.1, solutions starting in $\mathcal{X}_1^+$ converge to the origin in finite time. The vector field f is also transversal to $\mathcal{X}_2^-$ at every point in $\mathcal{X}_2^-$. Hence it follows from [14, Prop. 2.2] or [9, Lem. 2, p. 107] that initial conditions in $\mathcal{X}_2^-$ possess unique solutions in forward time. Thus (2.1) possesses a unique solution in forward time for every initial condition in $\mathbb{R}^2 \backslash \{(0,0)\}$.

We show that the system given in Figure 2.1 has a globally finite-time-stable equilibrium at the origin and demonstrate a sequence $\{x_m\}$ in $\mathbb{R}^2$ such that $x_m \to 0$ and $T(x_m) \to \infty$, where T is the settling-time function.

Lyapunov stability of the origin is easily verified using the Lyapunov function $x_1^2 + x_2^2$. To show global finite-time convergence, we show that solutions starting in $\mathcal{Q}_{\rm IV}$ and $\mathcal{Q}_{\rm III} \cup \mathcal{Q}_{\rm II}$ enter $\mathcal{Q}_{\rm III}$ and $\mathcal{Q}_{\rm I}$, respectively, in a finite amount of time, while solutions starting in $\mathcal{Q}_{\rm I}$ converge to the origin in finite time.

On $\mathcal{Q}_{\rm IV}$, $\dot{x}_2 = 0$ and $\dot{x}_1 \leq -x_2^2 < 0$ so that after a finite amount of time (that depends on the initial condition) every solution starting in $\mathcal{Q}_{\rm IV}$ enters $\mathcal{Q}_{\rm III}$. Since $r\cos\theta - \sqrt{\frac{\pi}{2}}\sin\theta \leq \max\left\{-\sqrt{\frac{\pi}{2}}, -r\right\}$ for $r > 0$ and $\theta \in \left[\frac{\pi}{2}, \pi\right]$, it follows that $\dot{r} = 0$ and $\dot{\theta} \leq \max\left\{-\sqrt{\frac{\pi}{2}}, -r\right\} < 0$ on $\mathcal{Q}_{\rm III} \cup \mathcal{Q}_{\rm II}$ so that every solution starting in $\mathcal{Q}_{\rm III} \cup \mathcal{Q}_{\rm II}$ enters $\mathcal{Q}_{\rm I}$ after a finite amount of time. Now, $\mathcal{Q}_{\rm I}$ is positively invariant. Hence, if

a solution y starting in $\mathcal{Q}_{\mathrm{I}}$ does not converge to the origin for a sufficiently long time, then, since the scalar equation $\dot{\theta} = -\sqrt{\theta}$ has the origin as a finite-time-stable equilibrium by Example 2.1, y converges to $\mathcal{X}_1^+$ in finite time. We have already seen that solutions in $\mathcal{X}_1^+$ converge to the origin in finite-time. Thus the origin is a globally finite-time-stable equilibrium.

Now consider the sequence $\{x_m\}$, where $x_m = (x_{m1}, x_{m2}) = \left(0, -\frac{1}{m}\right)$, $m = 1, 2, \ldots$, in $\mathcal{X}_2^-$. Thus $\{x_m\}$ lies in $\mathcal{X}_2^-$ and $x_m \to 0$ as $m \to \infty$. Since $\dot{\theta} = -r$ on $\mathcal{Q}_{\mathrm{III}}$, for every m, the time taken by the solution y_m starting at x_m to enter $\mathcal{Q}_{\mathrm{II}}$ is equal to $\frac{\pi}{2\sqrt{x_{m1}^2 + x_{m2}^2}} = \frac{m\pi}{2}$. Since y_m must enter $\mathcal{Q}_{\mathrm{II}}$ before converging to the origin, it follows that $T(x_m) \geq \frac{m\pi}{2}$ for every m and hence $T(x_m) \to \infty$.

Proposition 2.4 (iii), which is equivalent to the statement that $\frac{\|x\|}{T(x)} \to 0$ as $x \to 0$, implies that the settling-time function is not Lipschitz continuous at the origin. This is consistent with Example 2.1 where the settling-time function is not Lipschitz continuous. However, as noted earlier, the settling-time function in Example 2.1 is Hölder continuous at the origin. In contrast, the following example shows that even if the settling-time function is continuous, it may not be Hölder continuous at the origin.

Example 2.3. Consider the system (2.1) with $\mathcal{D} = \{x \in \mathbb{R} : |x| < 1\}$ and $f : \mathcal{D} \to \mathbb{R}$ given by

$$\begin{aligned} f(x) &= -x(\ln|x|)^2, && x \in \mathcal{D}\backslash\{0\}, \\ &= 0, && x = 0. \end{aligned} \tag{2.15}$$

The system defined by (2.15) is continuous and has the global semiflow

$$\begin{aligned} \mu(t,x) &= \operatorname{sign}(x) e^{\frac{\ln|x|}{1+t\ln|x|}}, && t < -\tfrac{1}{\ln|x|}, && x \in \mathcal{D}\backslash\{0\}, \\ &= 0, && t \geq -\tfrac{1}{\ln|x|}, && x \in \mathcal{D}\backslash\{0\}, \\ &= 0, && t \geq 0, && x = 0. \end{aligned} \tag{2.16}$$

From the solution (2.16), it is clear that 0 is a finite-time-stable equilibrium in the neighborhood $\mathcal{N} = \mathcal{D}$ and the settling-time function, which is continuous, is given by

$$\begin{aligned} T(x) &= -\frac{1}{\ln|x|}, && x \in \mathcal{D}\backslash\{0\}, \\ &= 0, && x = 0. \end{aligned} \tag{2.17}$$

Since $\lim_{h\to 0^+} h^\gamma |\ln h| = 0$ for every $\gamma > 0$, it follows that for every $\gamma > 0$, $\frac{T(\cdot)}{|\cdot|^\gamma}$ is unbounded in every deleted neighborhood of 0. Thus T is not Hölder continuous at the origin.

3. Finite-time repellers. The results of this section do not depend upon the assumption of forward uniqueness.

If the origin is not Lyapunov stable, then there exists an open neighborhood $\mathcal{U}$ of the origin and solutions that start arbitrarily close to the origin and eventually leave $\mathcal{U}$. However, in the case of Lipschitzian dynamics, solutions are continuous in the initial condition over bounded time intervals so that solutions with initial conditions sufficiently close to the origin stay in $\mathcal{U}$ for arbitrarily large amounts of time. In the non-Lipschitzian case, where solutions need not be continuous in the initial condition even over a bounded time interval, it is natural to expect the existence of solutions that start arbitrarily close to the origin and yet leave a certain neighborhood in a fixed amount of time. We therefore have the following definition.

DEFINITION 3.1. *The origin is said to be a* finite-time repeller *if there exists a neighborhood* $\mathcal{U} \subset \mathcal{D}$ *of the origin and* $\tau > 0$ *such that, for every open neighborhood* $\mathcal{V} \subseteq \mathcal{U}$ *of the origin, there exists* $h \in (0, \tau]$ *and a solution* $y : [0, h] \to \mathcal{D}$ *of* (2.1) *such that* $y(0) \in \mathcal{V}$ *and* $y(h) \notin \mathcal{U}$. *The origin is said to be a* finite-time saddle *if the origin is a finite-time repeller in forward as well as reverse time.*

Definition 3.1 implies that solutions of (2.1) with initial conditions sufficiently close to a finite-time repeller do not depend continuously on the initial conditions over the bounded time interval $[0, \tau]$. In other words, a system is extremely sensitive to perturbations close to a finite-time repeller. As noted in section 2, under the assumption of uniqueness, solutions are continuous functions of the initial conditions and hence nonuniqueness is necessary for the existence of a finite-time repeller. The following proposition gives the precise connection between nonuniqueness and finite-time repellers.

PROPOSITION 3.2. *The origin is a finite-time repeller if and only if there exist more than one solution of* (2.1) *originating at the origin.*

Proof. Note that $z \equiv 0$ is a solution of (2.1) satisfying $z(0) = 0$. To prove sufficiency, suppose $y : [0, \tau] \to \mathcal{D}$, $\tau > 0$, is a solution of (2.1) such that $y(0) = 0$ and $y(\tau) \neq 0$. Then there exists an open set $\mathcal{U} \subset \mathcal{D}$ such that $0 \in \mathcal{U}$ and $y(\tau) \notin \mathcal{U}$. If $\mathcal{V} \subseteq \mathcal{U}$ is an open neighborhood of the origin, then $0 = y(0) \in \mathcal{V}$ and $y(h) \notin \mathcal{V}$ for $h = \tau$. Thus the origin is a finite-time repeller.

To prove necessity, suppose that the origin is a finite-time repeller and let $\mathcal{U}$ and τ be as in Definition 3.1. There exists a sequence $\{h_m\}$ of real numbers in $(0, \tau]$ and a sequence of solutions $y_m : [0, h_m] \to \mathcal{D}$ of (2.1) such that, $y_m(0) \to 0$ as $m \to \infty$ and $y_m(h_m) \notin \mathcal{U}$. Now suppose that $z \equiv 0$ is the unique solution of (2.1) satisfying $z(0) = 0$. Then there exists $N > 0$ such that for every $m > N$, y_m can be extended to a solution $\hat{y}_m$ of (2.1) defined on $[0, \tau]$ and $\hat{y}_m \to z$ uniformly on $[0, \tau]$ [12, Lem. I.3.1]. However, this contradicts the fact that, for every m, $h_m \in [0, \tau]$ and $\hat{y}_m(h_m) = y_m(h_m) \notin \mathcal{U}$. Hence we conclude that $z \equiv 0$ is not the unique solution of (2.1) satisfying $z(0) = 0$. □

Finite-time repellers are called terminal repellers in [7], [23] and some of the references therein. Reference [5] gives an example of a one-degree-of-freedom Lagrangian system having a finite-time saddle, while in [4] finite-time saddles arise in the controlled double integrator. Proposition 3.2 implies that a system exhibits spontaneous and unpredictable departure from an equilibrium state that is a finite-time repeller. This property of finite-time repellers was used in [5] as an example of indeterminacy in classical dynamics, while [23] and some of the references contained therein postulate finite-time repellers as models of irreversibility and unpredictability in complex systems. Finally, [7] proposed a fast global optimization algorithm which utilizes the tendency of solutions to rapidly escape from a neighborhood of a finite-time repeller.

Sections 3.25 and 3.26 in [1] contain sufficient conditions for (2.1) to possess multiple solutions with the initial value 0. In view of Proposition 3.2, these conditions can also be used to deduce whether the origin is a finite-time repeller. Therefore, sufficient Lyapunov conditions for the origin to be a finite-time repeller will not be considered in this paper.

4. Lyapunov theory. The upper right Dini derivative of a function $g : [a, b) \to \mathbb{R}$, $b > a$, is the function $D^+g : [a, b) \to \overline{\mathbb{R}}$ given by

$$(D^+g)(t) = \limsup_{h \to 0^+} \frac{1}{h}[g(t + h) - g(t)], \ t \in [a, b). \tag{4.1}$$

The function g is nonincreasing on $[a,b)$ if and only if $(D^+g)(t) \leq 0$ for all $t \in [a,b)$ [13, p. 84], [16, p. 347]. If g is differentiable at t, then $(D^+g)(t)$ is the ordinary derivative of g at t.

If the scalar differential equation $\dot{y}(t) = w(y(t))$ has the global semiflow $\mu : \mathbb{R}_+ \times \mathbb{R} \to \mathbb{R}$, where $w : \mathbb{R} \to \mathbb{R}$ is continuous, and $g : [a,b) \to \mathbb{R}$ is a continuous function such that $(D^+g)(t) \leq w(g(t))$ for all $t \in [a,b)$, then $g(t) \leq \mu(t, g(a))$ for all $t \in [a,b)$. Proofs and more general versions of this result, which is known as the comparison lemma, can be found in [13, sect. 5.2], [15, sect. 2.5], [16, Chap. IX], and [22, sect. 4]. The comparison lemma will be used along with the scalar system of Example 2.1 in the proofs of the main results of this section and the next.

The following lemma will prove useful in the rest of the development.

LEMMA 4.1. *Let $V : \mathcal{A} \to \mathbb{R}$ be a continuous function defined on the open set $\mathcal{A} \subseteq \mathbb{R}^n$. Let $\mathcal{B}$ be an open set such that $\overline{\mathcal{B}} \subset \mathcal{A}$, let $\Omega_\kappa = \{x \in \mathcal{B} : V(x) < \kappa\}$, where $\kappa < \inf_{z \in \text{bd } \mathcal{B}} V(z)$, and let $p : \mathbb{R} \to \mathbb{R}$ be a continuous function satisfying $p(\kappa) > 0$. If $y : [a,b) \to \mathcal{A}$ is a continuous function that satisfies $y(a) \in \overline{\Omega}_\kappa$ and satisfies*

$$(D^+(V \circ y))(t) \leq -p(V \circ y(t)) \tag{4.2}$$

for every $t \in [a,b)$ such that $y(t) \in \mathcal{B}$, then $y(t) \in \Omega_\kappa$ for all $t \in (a,b)$.

Proof. The assertion is vacuously true if Ω_κ is empty. Therefore, let $y : [a,b) \to \mathcal{A}$ be a continuous function satisfying the hypotheses in the statement of the lemma. Note that by the choice of κ and the continuity of V, bd $\Omega_\kappa \subseteq \{x \in \mathcal{B} : V(x) = \kappa\}$.

First suppose that $y(a) \in$ bd Ω_κ. Since p, V, and y are continuous and $p(V(y(a))) = p(\kappa) > 0$, it follows that there exists $s > 0$ such that $p(V(y(t))) > 0$ for all $t \in [a, a+s)$. Moreover, s may be chosen such that $y(t) \in \mathcal{B}$ for all $t \in [a, a+s)$. Equation (4.2) now implies that $V \circ y$ is strictly decreasing on $[a, a+s)$ so that $y(t) \in \Omega_\kappa$ for all $t \in (a, a+s)$.

Now suppose $y(h) \in \Omega_\kappa$ for some $h \in [a,b)$. If $y(t) \notin \Omega_\kappa$ for some $t \in [h,b)$, then, by continuity, there exists $\tau \in (h,b)$ such that $y(\tau) \in$ bd Ω_κ and $y(t) \in \Omega_\kappa$ for all $t \in [h,\tau)$. Therefore, y satisfies (4.2) on $[h,\tau)$. Since p, V, and y are continuous and $p(V(y(\tau))) = p(\kappa) > 0$, it follows that there exists $s > 0$ such that $p(V(y(t))) > 0$ for all $t \in [\tau - s, \tau)$. Equation (4.2) now implies that $V \circ y$ is nonincreasing on $[\tau - s, \tau)$ so that $\kappa = V(y(\tau)) \leq V(y(\tau - s)) < \kappa$, which is a contradiction. Hence we conclude that $y(t) \in \Omega_\kappa$ for all $t \in [h,b)$.

It follows from the above two facts that if $y(a) \in \overline{\Omega}_\kappa$, then $y(t) \in \Omega_\kappa$ for all $t \in (a,b)$. □

Given a continuous function $V : \mathcal{D} \to \mathbb{R}$, the upper-right Dini derivative of V along the solutions of (2.1) is a $\overline{\mathbb{R}}$-valued function $\dot{V}$ given by

$$\dot{V}(x) = (D^+(V \circ \psi^x))(0). \tag{4.3}$$

$\dot{V}(x)$ is defined for every $x \in \mathcal{D}$ for which ψ^x is defined. It is easy to see that $\dot{V}(0)$, if defined, is 0. Also, since ψ is a local semiflow, it can be shown that if $\psi^x(t)$ is defined, then

$$\dot{V}(\psi^x(t)) = (D^+(V \circ \psi^x))(t). \tag{4.4}$$

It can also be shown that if V is locally Lipschitz at $x \in \mathcal{D} \backslash \{0\}$, then [13, sect. 5.1], [16, p. 353], [22, p. 3]

$$\dot{V}(x) = \limsup_{h \to 0^+} \frac{1}{h}[V(x + hf(x)) - V(x)]. \tag{4.5}$$

If V is continuously differentiable on $\mathcal{D}\backslash\{0\}$, then (4.3) and (4.5) both yield the Lie derivative

$$\dot{V}(x) = \frac{d(V \circ \psi^x)}{dt}(0) = \frac{\partial V}{\partial x}(x) f(x), \ x \in \mathcal{D}\backslash\{0\}. \tag{4.6}$$

A function $V : \mathcal{D} \to \mathbb{R}$ is said to be *proper* if $V^{-1}(K)$ is compact for every compact set $K \subset \mathbb{R}$. Note that if $\mathcal{D} = \mathbb{R}^n$ and V is radially unbounded, then V is proper.

We are now ready to state the main result of this paper. Versions of this result have either appeared without proof or have been used implicitly in [3], [4], [11], [18], [19].

THEOREM 4.2. *Suppose there exists a continuous function $V\colon \mathcal{D} \to \mathbb{R}$ such that the following conditions hold:*

(i) *V is positive definite.*

(ii) *There exist real numbers $c > 0$ and $\alpha \in (0, 1)$ and an open neighborhood $\mathcal{V} \subseteq \mathcal{D}$ of the origin such that*

$$\dot{V}(x) + c(V(x))^\alpha \leq 0, \ x \in \mathcal{V}\backslash\{0\}. \tag{4.7}$$

Then the origin is a finite-time-stable equilibrium of (2.1). *Moreover, if $\mathcal{N}$ is as in Definition* 2.2 *and T is the settling-time function, then*

$$T(x) \leq \frac{1}{c(1-\alpha)} V(x)^{1-\alpha}, \ x \in \mathcal{N}, \tag{4.8}$$

and T is continuous on $\mathcal{N}$. If in addition $\mathcal{D} = \mathbb{R}^n$, V is proper, and $\dot{V}$ takes negative values on $\mathbb{R}^n\backslash\{0\}$, then the origin is a globally finite-time-stable equilibrium of (2.1).

Proof. Since V is positive definite and $\dot{V}$ takes negative values on $\mathcal{V}\backslash\{0\}$, it follows that $y \equiv 0$ is the unique solution of (2.1) on $\mathbb{R}_+$ satisfying $y(0) = 0$ [1, sect. 3.15] [22, Thm. 1.2, p. 5]. Thus every initial condition in $\mathcal{D}$ has a unique solution in forward time. Moreover, $\dot{V}(0) = 0$ and thus (4.7) holds on $\mathcal{V}$.

Let $\mathcal{U} \subseteq \mathcal{V}$ be a bounded open set such that $0 \in \mathcal{U}$ and $\overline{\mathcal{U}} \subset \mathcal{D}$. Then bd $\mathcal{U}$ is compact and $0 \notin$ bd $\mathcal{U}$. The continuous function V attains a minimum on bd $\mathcal{U}$ and by positive definiteness, $\min_{x \in \text{bd } \mathcal{U}} V(x) > 0$. Let $0 < \beta < \min_{x \in \text{bd } \mathcal{U}} V(x)$ and $\mathcal{N} = \{x \in \mathcal{U} : V(x) < \beta\}$. $\mathcal{N}$ is nonempty since $0 \in \mathcal{N}$, open since V is continuous, and bounded since $\mathcal{U}$ is bounded.

Now, consider $x \in \mathcal{N}$ and let c and α be as in the theorem statement above. By uniqueness, $\psi^x : [0, \tau_x) \to \mathcal{D}$ is the unique right maximally defined solution of (2.1) for the initial condition x. For every $t \in [0, \tau_x)$ such that $\psi^x(t) \in \mathcal{U}$, (4.4) and (4.7) yield

$$(D^+(V \circ \psi^x))(t) \leq -c(V \circ \psi^x(t))^\alpha. \tag{4.9}$$

Thus $y = \psi^x$ satisfies the hypotheses of Lemma 4.1 with $\mathcal{A} = \mathcal{D}$, $\mathcal{B} = \mathcal{U}$, $\kappa = \beta$, $\Omega_\kappa = \mathcal{N}$, and $p(h) = ch^\alpha$ for $h \in \mathbb{R}_+$. Therefore, by Lemma 4.1, $\psi^x(t) \in \mathcal{N}$ for all $t \in [0, \tau_x)$. Now ψ^x satisfies the hypotheses of Proposition 2.1 with $\mathcal{K} = \mathcal{N}$. Therefore, by Proposition 2.1, ψ^x is defined and satisfies (4.9) on $\mathbb{R}_+$. Thus $\psi : \mathbb{R}_+ \times \mathcal{N} \to \mathcal{N}$ is a continuous global semiflow satisfying (2.3) and (2.4).

Next, applying the comparison lemma to the differential inequality (4.9) and the scalar differential equation (2.7) yields

$$V(\psi(t, x)) \leq \mu(t, V(x)), \ t \in \mathbb{R}_+, \ x \in \mathcal{N}, \tag{4.10}$$

where μ is given by (2.8) with $k = c$. From (2.8), (4.10), and the positive-definiteness of V, we conclude that

$$\psi(t,x) = 0,\ t \geq \frac{1}{c(1-\alpha)}(V(x))^{1-\alpha},\ x \in \mathcal{N}. \tag{4.11}$$

Since $\psi(0,x) = x$ and ψ is continuous, $\inf\{t \in \mathbb{R}_+ : \psi(t,x) = 0\} > 0$ for $x \in \mathcal{N}\backslash\{0\}$. Also, it follows from (4.11) that $\inf\{t \in \mathbb{R}_+ : \psi(t,x) = 0\} < \infty$ for $x \in \mathcal{N}$. Define $T : \mathcal{N} \to \mathbb{R}_+$ by using (2.6). It is a simple matter to verify that T and $\mathcal{N}$ satisfy (i) of Definition 2.2 and thus T is the settling-time function on $\mathcal{N}$. Lyapunov stability follows by noting from (4.7) that $\dot{V}$ takes negative values on $\mathcal{V}\backslash\{0\}$. Equation (4.8) follows from (4.11) and (2.6). Equation (4.8) implies that T is continuous at the origin and hence, by Proposition 2.4, continuous on $\mathcal{N}$.

If $\mathcal{D} = \mathbb{R}^n$ and V is proper, then global finite-time-stability is proven in the same way that global asymptotic stability is proven using radially unbounded Lyapunov functions. See, for instance, [15, Thm. 3.2], [22, Thm. 11.5]. □

Remark 4.1. It is difficult to compute $\dot{V}$ by using (4.3) unless solutions to (2.1) are known. Thus, in practice, it will often be more convenient to apply Theorem 4.2 with a Lipschitz continuous or a continuously differentiable function V so that $\dot{V}$ is given by (4.5) or (4.6), respectively.

Theorem 4.2 implies that for a system with a finite-time-stable equilibrium and a discontinuous settling-time function, such as the system considered in Example 2.2, there does not exist a Lyapunov function satisfying the hypotheses of Theorem 4.2. In the case that the settling-time function is continuous, the following theorem provides a converse to the previous one.

THEOREM 4.3. *Suppose the origin is a finite-time-stable equilibrium of* (2.1) *and the settling-time function T is continuous at* 0. *Let $\mathcal{N}$ be as in Definition* 2.2 *and let $\alpha \in (0,1)$. Then there exists a continuous function $V : \mathcal{N} \to \mathbb{R}$ such that the following conditions are satisfied:*

(i) *V is positive definite.*

(ii) *$\dot{V}$ is real valued and continuous on $\mathcal{N}$ and there exists $c > 0$ such that*

$$\dot{V}(x) + c(V(x))^{\alpha} \leq 0,\ x \in \mathcal{N}.$$

Proof. By Proposition 2.4, the settling-time function $T\colon \mathcal{N} \to \mathbb{R}_+$ is continuous. Define $V : \mathcal{N} \to \mathbb{R}_+$ by $V(x) = (T(x))^{\frac{1}{1-\alpha}}$. Then V is continuous and positive definite and, by (2.5), $\dot{V}(0) = 0$. For $x \in \mathcal{N}\backslash\{0\}$, (2.10) implies that $V \circ \psi^x$ is continuously differentiable on $[0, T(x))$ so that (4.3) can be easily computed as $\dot{V}(x) = -\frac{1}{1-\alpha}(T(x))^{\frac{\alpha}{1-\alpha}} = -\frac{1}{1-\alpha}(V(x))^{\alpha}$. Thus $\dot{V}$ is real valued, continuous, and negative definite on $\mathcal{N}$ and satisfies $\dot{V}(x) + c(V(x))^{\alpha} = 0$ for all $x \in \mathcal{N}$ with $c = \frac{1}{1-\alpha}$. □

Equation (4.8) implies that if V in Theorem 4.2 is Hölder continuous at 0 then so is T. However, as shown by Example 2.3, the settling-time function need not be Hölder continuous at the origin. Thus the conclusion regarding the continuity of V in Theorem 4.3 cannot be strengthened to Hölder continuity. In particular, the scalar system considered in Example 2.3, where T is not Hölder continuous, does not admit a continuously differentiable or Lipschitz continuous Lyapunov function that satisfies the hypotheses of Theorem 4.2, since either Lipschitz continuity or differentiability implies Hölder continuity. As the next section shows, the existence of Lipschitz continuous Lyapunov functions is of importance in studying the behavior of finite-time-stable systems in the presence of perturbations.

5. Sensitivity to perturbations. In a realistic problem, (2.1) might represent a nominal model that is valid only under ideal conditions, while a more accurate description of the system might be provided by the perturbed model

$$\dot{y}(t) = f(y(t)) + g(t, y(t)), \tag{5.1}$$

where the perturbation term g results from disturbances, uncertainties, parameter variations, or modelling errors. In this section we investigate the sensitivity to perturbations of systems with a finite-time-stable equilibrium by studying the behavior of solutions of the perturbed system (5.1) in a neighborhood of the finite-time-stable equilibrium of the nominal system (2.1).

For simplicity, we consider only continuous perturbation terms $g : \mathbb{R}_+ \times \mathcal{D} \to \mathbb{R}^n$ so that the local existence of solutions of the perturbed system (5.1) is guaranteed. Right maximally defined solutions of (5.1) are defined as in section 2. We will need the following extension of Proposition 2.1 to time-varying systems. Proofs appear in [12, pp. 17–18], [22, Thm. 3.3, p. 12].

PROPOSITION 5.1. *If $y : [0, \tau) \to \mathcal{D}$ is a right maximally defined solution of* (5.1) *such that $y(t) \in \mathcal{K}$ for all $t \in [0, \tau)$, where $\overline{\mathcal{K}} \subset \mathcal{D}$ is compact, then $\tau = \infty$.*

If $y : [0, \tau) \to \mathcal{D}$ is a solution of (5.1) and $V : \mathcal{D} \to \mathbb{R}$ is Lipschitz continuous on $\mathcal{D}$ with Lipschitz constant M, then it can be shown that

$$(D^+(V \circ y))(t) \leq \dot{V}(y(t)) + M\|g(t, y(t))\|, \ t \in [0, \tau), \tag{5.2}$$

where $\dot{V}$ is computed along the solutions of the unperturbed system (2.1) using equation (4.3). See, for instance, the proof of Lemma X.5.1 in [12].

The following theorem concerns the behavior of finite-time-stable systems under bounded perturbations. Such perturbations, include, as a special case, bounded persistent disturbances of the form $g(t, y(t)) = v(t)$.

THEOREM 5.2. *Suppose there exists a function $V : \mathcal{D} \to \mathbb{R}$ such that V is positive definite and Lipschitz continuous on $\mathcal{D}$, and satisfies* (4.7), *where $\mathcal{V} \subseteq \mathcal{D}$ is an open neighborhood of the origin, $c > 0$ and $\alpha \in \left(0, \frac{1}{2}\right)$. Then there exist $\delta_0 > 0$, $l > 0$, $\Gamma > 0$, and an open neighborhood $\mathcal{U}$ of the origin such that, for every continuous function $g : \mathbb{R}_+ \times \mathcal{D} \to \mathbb{R}^n$ with*

$$\delta = \sup_{\mathbb{R}_+ \times \mathcal{D}} \|g(t, x)\| < \delta_0, \tag{5.3}$$

every right maximally defined solution y of (5.1) *with $y(0) \in \mathcal{U}$ is defined on $\mathbb{R}_+$ and satisfies $y(t) \in \mathcal{U}$ for all $t \in \mathbb{R}_+$ and*

$$\|y(t)\| \leq l\delta^\gamma, \ t \geq \Gamma, \tag{5.4}$$

where $\gamma = \frac{1-\alpha}{\alpha} > 1$.

Proof. By Theorem 4.2, the origin is a finite-time-stable equilibrium for (2.1). Let $\mathcal{N}$ be as in Definition 2.2 and let $T : \mathcal{N} \to \mathbb{R}^+$ be the settling-time function. By Proposition 2.4, there exists $r > 0$ and an open neighborhood $\mathcal{U}_r \subset \mathcal{N} \cap \mathcal{V}$ of 0 such that $T(x) \geq r\|x\|$ for $x \in \mathcal{U}_r$. Also, by Theorem 4.2, T satisfies (4.8). Without any loss of generality, we assume that $\overline{\mathcal{U}}_r$ is compact and $\overline{\mathcal{U}}_r \subset \mathcal{N} \cap \mathcal{V}$. Let $\mathcal{U} = \{x \in \mathcal{U}_r : V(x) < \beta\}$, where $0 < \beta < \min_{z \in \mathrm{bd}\ \mathcal{U}_r} V(z)$. Then $\mathcal{U}$ is nonempty, open, and bounded.

Let $M > 0$ be the Lipschitz constant of V and let $\delta_0 > 0$ satisfy $c\beta^\alpha - 2M\delta_0 > 0$. Suppose $g : \mathbb{R}_+ \times \mathcal{D} \to \mathbb{R}^n$ is a continuous function that satisfies (5.3) and consider a

right maximally defined solution $y : [0, \tau) \to \mathcal{D}$ of (5.1) with $x = y(0) \in \mathcal{U}$. Equations (4.7), (5.2), and (5.3) imply that for every $t \in [0, \tau)$ such that $y(t) \in \mathcal{U}_r$,

$$(D^+(V \circ y))(t) \leq -c(V(y(t)))^\alpha + M\delta. \tag{5.5}$$

Since $c\beta^\alpha - M\delta > c\beta^\alpha - 2M\delta > c\beta^\alpha - 2M\delta_0 > 0$, (5.5) implies that y satisfies the hypotheses of Lemma 4.1 with $\mathcal{A} = \mathcal{D}$, $\mathcal{B} = \mathcal{U}_r$, $\kappa = \beta$, $\Omega_\kappa = \mathcal{U}$, and $p(h) = ch^\alpha - M\delta$ for $h \in \mathbb{R}_+$. Therefore, Lemma 4.1 implies that $y(t) \in \mathcal{U}$ for $t \in [0, \tau)$. The right maximally defined solution y satisfies the hypotheses of Proposition 5.1 with $\mathcal{K} = \mathcal{U}$. Thus, by Proposition 5.1, y is defined on $\mathbb{R}_+$ and (5.5) holds on $\mathbb{R}_+$.

Now, let $\mathcal{W} = \{x \in \mathcal{U} : V(x) < (\frac{2M\delta}{c})^{\frac{1}{\alpha}}\}$. If $y(\tau) \in \overline{\mathcal{W}}$ for some $\tau > 0$, then (5.5) implies that y satisfies the hypotheses of Lemma 4.1 on $[\tau, \infty)$ with $\mathcal{A} = \mathcal{U}_r$, $\mathcal{B} = \mathcal{U}$, $\kappa = \left(\frac{2M\delta}{c}\right)^{\frac{1}{\alpha}}$, $\Omega_\kappa = \mathcal{W}$, and $p(h) = ch^\alpha - M\delta$ for $h \in \mathbb{R}_+$, and hence $y(t) \in \mathcal{W}$ for all $t > \tau$. Therefore, suppose $y(0) = x \notin \overline{\mathcal{W}}$ so that $y^{-1}(\mathcal{W})$, which is open by continuity, is of the form (t_x, ∞) with $t_x > 0$. Since $y(t) \notin \mathcal{W}$ for all $t \in [0, t_x]$, it follows that $V(y(t)) \geq \left(\frac{2M\delta}{c}\right)^{\frac{1}{\alpha}}$ for all $t \in [0, t_x]$. Equation (5.5) now implies that

$$(D^+(V \circ y))(t) \leq -\frac{1}{2}c(V(y(t)))^\alpha, \ t \in [0, t_x). \tag{5.6}$$

Applying the comparison principle to the differential inequality (5.6) and the scalar differential equation (2.7) we obtain

$$(V \circ y)(t) \leq \mu(t, V(x)), \ t \in [0, t_x), \tag{5.7}$$

where μ is given by (2.8) with $k = \frac{1}{2}c$. By continuity, the inequality (5.7) also holds for $t = t_x$. Since $V(y(t_x)) \geq \left(\frac{2M\delta}{c}\right)^{\frac{1}{\alpha}} > 0$, the comparison (5.7) yields $\mu(t_x, V(x)) > 0$. Equation (2.8) now gives $t_x < \frac{2}{c(1-\alpha)}(V(x))^{1-\alpha} < \frac{2}{c(1-\alpha)}\beta^{1-\alpha}$. Thus $V(y(t)) < \left(\frac{2M\delta}{c}\right)^{\frac{1}{\alpha}}$ for $t \geq \Gamma \triangleq \frac{2}{c(1-\alpha)}\beta^{1-\alpha}$. It now follows from (2.11) and (4.8) that for $t > \Gamma$,

$$\|y(t)\| \leq \frac{1}{rc(1-\alpha)}(V(y(t)))^{1-\alpha} \leq \frac{1}{rc(1-\alpha)}\left(\frac{2M\delta}{c}\right)^{\frac{1-\alpha}{\alpha}}.$$

Equation (5.4) now follows by choosing $l \triangleq \frac{1}{rc(1-\alpha)}\left(\frac{2M}{c}\right)^{\frac{1-\alpha}{\alpha}} > 0$. □

Note that in Theorem 4.3, α can be chosen to be arbitrarily small. Hence the requirement in Theorem 5.2 that α lie in $\left(0, \frac{1}{2}\right)$ is not restrictive. This choice of α leads to $\gamma > 1$ in (5.4) which implies that for δ in equation (5.3) sufficiently small, the ultimate bound (5.4) on the state is of higher order than the bound on the perturbation. In analogous theorems on exponential stability for Lipschitzian systems, α in equation (4.7) is at least 1 [15, Thm. 3.12], [22, Thm. 19.2] while γ in (5.4) is at most 1 [15, Lemma 5.2]. Thus for a Lipschitzian system with an exponentially stable equilibrium at the origin, the ultimate bound on the state can only be guaranteed to be of the same order of magnitude as the perturbation and not less. Consequently, finite-time stability of the origin leads to improved rejection of low-level persistent disturbances.

The following theorem deals with perturbations that are globally Lipschitz in the state variables uniformly in time. Such perturbations might represent model uncertainties.

THEOREM 5.3. *Suppose there exists a function $V : \mathcal{D} \to \mathbb{R}$ such that V is positive definite and Lipschitz continuous on $\mathcal{D}$, and satisfies* (4.7)*, where $\mathcal{V} \subseteq \mathcal{D}$ is an open neighborhood of the origin, $c > 0$ and $\alpha \in \left(0, \frac{1}{2}\right)$. Then, for every $L \geq 0$, there exists an open neighborhood $\mathcal{U}$ of the origin and $\Gamma > 0$ such that, for every continuous function $g : \mathbb{R}_+ \times \mathcal{D} \to \mathbb{R}^n$ satisfying*

$$\|g(t,x)\| \leq L\|x\|, \ (t,x) \in \mathbb{R}_+ \times \mathcal{D}, \tag{5.8}$$

every right maximally defined solution y of (5.1) *with $y(0) \in \mathcal{U}$ is defined on $\mathbb{R}_+$ and satisfies $y(t) \in \mathcal{U}$, for all $t \in \mathbb{R}_+$, and $y(t) = 0$ for all $t \geq \Gamma$.*

Proof. By Theorem 4.2, the origin is a finite-time-stable equilibrium for (2.1). Let $\mathcal{N}$ be as in Definition 2.2 and let $T : \mathcal{N} \to \mathbb{R}_+$ be the settling-time function. Fix $L \geq 0$ and let $r > 0$ be such that $c[r(1-\alpha)]^\alpha > (2ML)^{1-\alpha}$, where $M > 0$ is the Lipschitz constant of V. By Proposition 2.4, there exists an open set $\mathcal{U}_r \subset \mathcal{N} \cap \mathcal{V}$ such that $r\|x\| \leq T(x)$ for all $x \in \mathcal{U}_r$. Also, by Theorem 4.2, T satisfies (4.8). Without any loss of generality, we may assume that $\|x\| < 1$ for $x \in \mathcal{U}_r$ and $\overline{\mathcal{U}}_r \subset \mathcal{N} \cap \mathcal{V}$. Let $\mathcal{U} = \{x \in \mathcal{U}_r : V(x) < \beta\}$, where $0 < \beta < \min_{z \in \operatorname{bd} \mathcal{U}_r} V(z)$. Note that $\mathcal{U}$ is nonempty, open, and bounded. Also, $\frac{\alpha}{1-\alpha} < 1$ so that $\|x\| \leq \|x\|^{\frac{\alpha}{1-\alpha}}$ for $\|x\| < 1$. Therefore, (2.11) and (4.8) yield

$$2ML\|x\| \leq c[rc(1-\alpha)\|x\|]^{\frac{\alpha}{1-\alpha}} \leq c[c(1-\alpha)T(x)]^{\frac{\alpha}{1-\alpha}} \leq c(V(x))^\alpha, \ x \in \mathcal{U}_r. \tag{5.9}$$

Next, let $x \in \mathcal{U}$ and let $g : \mathbb{R}_+ \times \mathcal{D} \to \mathbb{R}^n$ be a continuous function satisfying (5.8). Consider a right maximally defined solution $y : [0, \tau) \to \mathcal{D}$ of (5.1) such that $y(0) = x$. For every $t \in [0, \tau)$ such that $y(t) \in \mathcal{U}_r$, (4.7), (5.2), and (5.8) yield

$$(D^+(V \circ y))(t) \leq -c(V \circ y(t))^\alpha + ML\|y(t)\|. \tag{5.10}$$

Using (5.9) in (5.10) we obtain

$$(D^+(V \circ y))(t) \leq -\frac{c}{2}(V \circ y(t))^\alpha, \ y(t) \in \mathcal{U}_r. \tag{5.11}$$

Lemma 4.1 now applies with $\mathcal{A} = \mathcal{D}$, $\mathcal{B} = \mathcal{U}_r$, $\kappa = \beta$, $\Omega_\kappa = \mathcal{U}$, and $p(h) = \frac{c}{2}h^\alpha$ for $h \in \mathbb{R}_+$ so that $y(t) \in \mathcal{U}$ for $t \in [0, \tau)$. The hypotheses of Proposition 5.1 are now satisfied by the right maximally defined solution y of (5.1) with $\mathcal{K} = \mathcal{U}$. Hence, by Proposition 5.1, $\tau = \infty$ and (5.11) holds on $\mathbb{R}_+$. Applying the comparison principle to the differential inequality (5.11) and the scalar differential equation (2.7) yields

$$(V \circ y)(t) \leq \mu(t, V(x)), \ t \in \mathbb{R}_+, \tag{5.12}$$

where μ is given by (2.8) with $k = \frac{1}{2}c$. Equation (2.8) and the inequality (5.12) imply that $y(t) = 0$ for $t \geq \Gamma \triangleq \frac{2\beta^{1-\alpha}}{c(1-\alpha)}$. ∎

The following theorem specializes Theorem 5.3 to time-invariant perturbations and shows that finite-time stability is preserved under Lipschitzian perturbations.

THEOREM 5.4. *Suppose there exists a function $V : \mathcal{D} \to \mathbb{R}$ such that V is positive definite and Lipschitz continuous on $\mathcal{D}$ and satisfies* (4.7)*, where $\mathcal{V} \subseteq \mathcal{D}$ is an open neighborhood of the origin, $c > 0$, and $\alpha \in \left(0, \frac{1}{2}\right)$. Let $g : \mathcal{D} \to \mathbb{R}^n$ be Lipschitz*

continuous on $\mathcal{D}$ and such that the differential equation

$$\dot{y}(t) = f(y(t)) + g(y(t)) \tag{5.13}$$

possesses unique solutions in forward time for initial conditions in $\mathcal{D}\backslash\{0\}$. Then the origin is a finite-time-stable equilibrium of (5.13).

Proof. Using steps similar to those used in deriving (5.11) above, it can be shown that $\dot{V}(x) \leq -\frac{c}{2}(V(x))^{\alpha}$ for all x in some open neighborhood of the origin, where $\dot{V}$ denotes the upper-right Dini derivative of V along the solutions of (5.13). Finite-time stability now follows from Theorem 4.2. □

The existence of a Lipschitz continuous function satisfying the hypotheses of Theorem 5.4 is sufficient but not necessary for the conclusions to hold. For instance, consider a scalar system of the form (5.13) where the nominal dynamics f are given by (2.15) in Example 2.3, and $g : \mathcal{D} \to \mathbb{R}$ is Lipschitz continuous on $\mathcal{D} = \{x \in \mathbb{R} : |x| < 1\}$ with Lipschitz constant L. As noted at the end of section 4, the nominal dynamics do not admit a Lipschitz continuous Lyapunov function satisfying the hypotheses of Theorem 5.4. However, the origin is still a finite-time-stable equilibrium for the perturbed system. This can be verified by considering the continuous Lyapunov function $V(x) = (\ln|x|)^{-2}$, $x \in \mathcal{D}\backslash\{0\}$, $V(0) = 0$. It is easy to compute $\dot{V}$ along the solutions of the perturbed system (5.13) for $x \neq 0$ and establish that $\dot{V}(x) < -\sqrt{V(x)}$ for $0 < |x| < e^{-\sqrt{2L}}$, thus proving finite-time stability by Theorem 4.2. This indicates that the main results of this section may be valid under the weaker assumption of finite-time stability with a continuous settling-time function. From the point of view of stability theory, proofs of these results under such weaker hypotheses are certainly of interest. However, as observed in Remark 4.1, the Lyapunov functions used to verify stability properties are often continuously differentiable in practice. In such a case, the results of this section are immediately applicable.

6. Conclusions. The notion of finite-time stability can be precisely formulated within the framework of continuous autonomous systems with forward uniqueness. These assumptions, however, do not imply any regularity properties for the settling-time function, which may be discontinuous or continuous yet Hölder discontinuous.

Lyapunov and converse Lyapunov results for finite-time stability naturally involve finite-time scalar differential inequalities. The regularity properties of a Lyapunov function satisfying such an inequality strongly depend on the regularity properties of the settling-time function.

Under the assumption of the existence of a Lipschitz continuous Lyapunov function, finite-time stability leads to better rejection of persistent as well as vanishing perturbations. Such an assumption, however, is not strictly necessary, as the discussion at the end of section 5 shows.

The paper thus raises certain questions that are important from the point of view of stability theory. In particular, conditions on the dynamics for the settling-time function to be Hölder continuous and conditions on the settling-time function that lead to a stronger converse result than Theorem 4.3 are of interest. Also of interest are results similar to those given in section 5 but with weaker hypotheses.

As mentioned earlier, a control system under the action of a time-optimal feedback controller yields a closed-loop system that exhibits finite-time convergence. Hence it would be interesting to explore the connections between finite-time-stability and time optimality and relate the results of this paper to results on the time-optimal control problem.

REFERENCES

[1] R. P. Agarwal and V. Lakshmikantham, *Uniqueness and Nonuniqueness Criteria for Ordinary Differential Equations*, Ser. Real Anal. 6, World Scientific, Singapore, 1993.

[2] M. Athans and P. L. Falb, *Optimal Control: An Introduction to the Theory and Its Applications*, McGraw-Hill, New York, 1966.

[3] S. P. Bhat and D. S. Bernstein, *Lyapunov Analysis of Finite-Time Differential Equations*, in Proceedings of the American Control Conference, Seattle, WA, 1995, pp. 1831–1832.

[4] S. P. Bhat and D. S. Bernstein, *Continuous, finite-time stabilization of the translational and rotational double integrators*, IEEE Trans. Automat. Control, 43 (1998), pp. 678–682.

[5] S. P. Bhat and D. S. Bernstein, *Example of indeterminacy in classical dynamics*, Internat. J. Theoret. Phys., 36 (1997), pp. 545–550.

[6] N. P. Bhatia and O. Hajek, *Local Semi-Dynamical Systems*, Lecture Notes in Math. 90, Springer-Verlag, Berlin, 1969.

[7] B. C. Cetin, J. Barhen, and J. W. Burdick, *Terminal repeller unconstrained subenergy tunneling (TRUST) for fast global optimization*, J. Optim. Theory Appl., 77 (1993), pp. 97–126.

[8] J.-M. Coron, *On the stabilization in finite time of locally controllable systems by means of continuous time-varying feedback law*, SIAM J. Control Optim., 33 (1995), pp. 804–833.

[9] A. F. Filippov, *Differential Equations with Discontinuous Righthand Sides*, Math. Appl., Kluwer Academic Publishers, Dordrecht, the Netherlands, 1988.

[10] I. Flügge-Lotz, *Discontinuous and Optimal Control*, McGraw–Hill, New York, 1968.

[11] V. T. Haimo, *Finite time controllers*, SIAM J. Control Optim., 24 (1986), pp. 760–770.

[12] J. K. Hale, *Ordinary Differential Equations*, 2nd ed., Pure and Applied Mathematics XXI, Krieger, Malabar, FL, 1980.

[13] A. G. Kartsatos, *Advanced Ordinary Differential Equations*, Mariner, Tampa, FL, 1980.

[14] M. Kawski, *Stabilization of nonlinear systems in the plane*, Systems Control Lett., 12 (1989), pp. 169–175.

[15] H. K. Khalil, *Nonlinear Systems*, 2nd ed., Tr. Prentice–Hall, Upper Saddle River, NJ, 1996.

[16] N. Rouche, P. Habets, and M. Laloy, *Stability Theory by Liapunov's Direct Method*, Appl. Math. Sciences, Springer-Verlag, New York, 1977.

[17] E. P. Ryan, *Optimal Relay and Saturating Control System Synthesis*, IEE Control Engrg. Ser. 14, Peter Peregrinus Ltd., London, 1982.

[18] E. P. Ryan, *Finite-time stabilization of uncertain nonlinear planar systems*, Dynam. Control, 1 (1991), pp. 83–94.

[19] S. V. Salehi and E. P. Ryan, *On optimal nonlinear feedback regulation of linear plants*, IEEE Trans. Automat. Control, 26 (1982), pp. 1260–1264.

[20] G. R. Sell, *Topological Dynamics and Ordinary Differential Equations*, Van Nostrand Reinhold, London, 1971.

[21] S. T. Venkataraman and S. Gulati, *Terminal slider control of nonlinear systems*, in Proceedings of the International Conference on Advanced Robotics, Pisa, Italy, 1990.

[22] T. Yoshizawa, *Stability Theory by Liapunov's Second Method*, the Mathematical Society of Japan, Tokyo, 1966.

[23] M. Zak, *Introduction to terminal dynamics*, Complex Systems, 7 (1993), pp. 59–87.

SIAM J. CONTROL OPTIM.
Vol. 38, No. 3, pp. 767–785

MESH-INDEPENDENCE FOR AN AUGMENTED LAGRANGIAN-SQP METHOD IN HILBERT SPACES*

S. VOLKWEIN†

Abstract. An augmented Lagrangian-SQP algorithm for optimal control of differential equations in Hilbert spaces is analyzed. This algorithm has second-order convergence rate provided that a second-order sufficient optimality condition is satisfied. The internal approximation of this method is investigated, and convergence results are presented. A mesh-independence principle for the augmented Lagrangian-SQP method is proved, which asserts that asymptotically the infinite dimensional algorithm and finite dimensional discretizations have the same convergence property. More precisely, for sufficiently small mesh-size there is at most a difference of one iteration step between the number of steps required by the infinite dimensional method and its discretization to converge within a given tolerance $\varepsilon > 0$. The theoretical results are demonstrated by two optimal control problems for the Burgers equation.

Key words. nonlinear programming, multiplier methods, internal approximation, mesh-independence, Burgers equation

AMS subject classifications. 49K20, 49M29, 49M37, 65Nxx

PII. S0363012998334468

1. Introduction.

The numerical treatment of optimal control problems for nonlinear partial differential equations arising in diverse areas of science has received an increasing amount of attention in the recent past. We mention optimal control problems in combustion (see [Lio85] for control of parabolic explosive systems), in phase field modeling [CH91, HS94], in superconductivity [GHS91, Tin75], and in fluid dynamics (see [Gun95] and references therein). There is great interest for the numerical treatment of such problems.

In this work we study a general class of optimal control problems for nonlinear differential equations of the following type

$$\text{(P)} \qquad \text{minimize } J(x) \text{ subject to } e(x) = 0,$$

where $J : X \to \mathbb{R}$ and $e : X \to Y$ are sufficiently smooth functions; X and Y are Hilbert spaces. We assume that (P) has a local solution x^* and denote by λ^* the associated Lagrange multiplier.

To solve (P) we use the augmented Lagrangian-SQP (sequential quadratic programming) technique as developed in [IK96b]. In this method the differential equation is treated as an equality constraint which is realized by a Lagrangian term together with a penalty functional. We present an algorithm which has second-order convergence rate and depends upon a second-order sufficient optimality condition. In comparison with SQP methods, the augmented Lagrangian-SQP method has the advantage of a more global behavior. For certain examples we found it to be less sensitive with respect to the starting values, and the region for second-order convergence rate was reached earlier. We shall point out that the penalty term of the augmented Lagrangian functional need not to be implemented but rather that it can be realized by a first-order Lagrangian update.

*Received by the editors February 24, 1998; accepted for publication (in revised form) September 21, 1999; published electronically February 29, 2000.
http://www.siam.org/journals/sicon/38-3/33446.html

†Institut für Mathematik, Karl-Franzens-Universität Graz, Heinrichstrasse 36, A-8010 Graz, Austria (stefan.volkwein@kfunigraz.ac.at).

Since the algorithm cannot be executed in infinite dimensional spaces, (P) is replaced by a family of internal approximations

$$(\mathrm{P}_h) \qquad \text{minimize } J_h(x_h) \text{ subject to } e_h(x_h) = 0,$$

indexed by some mesh-size parameter h, where now $J_h : X_h \to \mathbb{R}$, $e_h : X_h \to Y_h$, $\{X_h\}_{h>0}$, and $\{Y_h\}_{h>0}$ are given families of finite dimensional real Hilbert spaces. We propose the corresponding finite dimensional discretization of the augmented Lagrangian-SQP method. Using convergence properties in subspaces of X and Y, the existence of x^* solving (P) implies the existence of a solution to (P_h) for sufficiently small h. Convergence and rate of convergence results of the discretized algorithm are proved. Moreover, when the augmented Lagrangian-SQP algorithm and its finite dimensional discretization are stopped after the nth step, the difference between the iterates of the finite and the infinite dimensional methods can be estimated by the discretization error of the approximation scheme used. Since we have to solve a saddle-point problem at each level of the iteration process, these results depend on the discrete Babuška–Brezzi condition for the saddle-point problem.

We prove a mesh-independence principle for the augmented Lagrangian-SQP method. The proof is essentially based on the proof of the mesh-independence principle for the Newton method given in [ABPR86]. The mesh-independence principle asserts that, when the augmented Lagrangian-SQP method is applied to an optimal control problem in Hilbert spaces as well as to some finite dimensional discretization of that problem, then the behavior of the discretized process is asymptotically the same as that for the original problem. We shall give sufficient conditions for mesh-independence, provided the mesh-size is smaller than an explicitly given threshold parameter. The sufficient assumptions for mesh-independence are strongly related to the assumptions that are sufficient for the convergence of the discretized augmented Lagrangian-SQP algorithm.

Mesh-independence allows us to predict the convergence of the method applied to the discretized problem when the method has been analyzed for the infinite dimensional problem. Further, it can be used to improve the performance of the method. Since we are interested in the solution of an infinite dimensional problem, it is usually necessary to choose reasonably fine discretizations. This leads to a large number of variables in the discrete minimization problem and therefore to a large amount of work per iteration. If the method is fixed, the only possibility for reducing the total amount of work consists of a good choice in the starting value. For these problems it is obvious that we must use information from the coarse grids to obtain good starting values for the finer discretizations, which leads to mesh-refinement strategies. Mesh-independence is a theoretical justification for mesh-refinement strategies and, moreover, it can be used to design the refinement process and to predict the overall performance of the algorithm.

Let us put the present work into perspective with related research efforts. Polyak and Tret'yakov [PT73] give an elegant treatise of the augmented Lagrangian method. In the book of Fortin and Glowinski [FG83] augmented Lagrangian methods are developed systematically for equality constraints as a technique for solving nonlinear partial differential equations. In [IK90b] Ito and Kunisch apply a technique realized by an augmented Lagrangian formulation for parameter estimation in elliptic partial differential equations. An augmented Lagrangian method for the minimization of a nonlinear functional in the presence of an equality and an affine inequality constraint is considered in [IK90a]. Ito, Kroller, and Kunisch discuss their numerical experience with an augmented Lagrangian method for estimating the coefficient in an elliptic

equation from some given measurement [IKK91]. We refer to the paper [KP91], where Kunisch and Peichl develop an estimation procedure for the diffusion coefficient in a parabolic partial differential equation from the knowledge of the state. In [IK96b] Ito and Kunisch continue their efforts to develop stable and efficient techniques to solve parameter estimation problems formulated as nonlinear optimization problems. Such problems are ill-posed in the sense of a possible lack of continuous dependence of the minimizers with respect to perturbations of the problem. The authors concentrate on second-order methods and analyze the augmented Lagrangian-SQP technique with a second-order update of the Lagrange multiplier in a general Hilbert space setting. Quadratical convergence for smooth optimization problems satisfying the second-order sufficient optimality condition is proved. In [IK96a] the second-order sufficient optimality condition is analyzed and numerical test examples are given. We refer to the paper [KV97], where Kunisch and Volkwein study two augmented Lagrangian-SQP algorithms for optimal control of partial differential equations. The approximation of one of these methods is analyzed. Examples illustrate the theoretical investigations. Further, in [Vol99] the application of the augmented Lagrangian-SQP algorithm to optimal control problems for the stationary Burgers equation is studied. For some classes of nonlinear boundary value problems the mesh-independence principle for Newton's method is proved in [AM78] and [AMP79], and in [AM78], [ABM81], and [McC78] it is used for the construction of some mesh-refinement strategies. In [ABPR86] a proof of the mesh-independence principle for Newton's method is presented in its sharpest formulation for a general class of discretizations of nonlinear operator equations under fairly general and natural conditions on the operator and the discretizations. The paper combines the results of [AB87] with the procedure in [PR84]. Argyros extends the validity of the mesh-independence principle for nonlinear equations and their discretizations to include operators whose derivatives are only Hölder-continuous [Arg90]. In [Arg92] Argyros proved a mesh-independence principle for operator equations and the secant method. A mesh-independence result of Newton's method for generalized equations is achieved by Alt in [Alt95]. Kelley and Sachs prove a mesh-independence result for the Broyden–Fletcher–Goldfarb–Shanno (BFGS) method in Hilbert spaces and apply it to a class of unconstrained optimal control problems in [KS87]. In [KS91] Kelley and Sachs consider globally convergent modifications of Newton's method, such as the Armijo rule. It is shown that for proper discretization schemes the convergence behavior of the iteration is the same for the discrete problems as it is for the infinite dimensional problem. Optimal control problems and their discretizations are considered by Kelley and Sachs in [KS92]. To solve such discretized problems by the gradient projection method the finite identification of active constraints will be necessary. These authors prove a mesh-independence result using a proper condition on the convergence of the gradients. In [Hei93] Heinkenschloss extends the results of [ABPR86] to a norm constrained Gauss–Newton method using a somewhat different discretization scheme based on Galerkin approximations. Deuflhard and Potra present a theoretical characterization of the asymptotic mesh-independence of Newton's method [DP92]. The theory does not need any uniform Lipschitz assumptions. The results are obtained by means of a refined Newton–Mysovskii theorem in an affine invariant formulation. A theoretical refinement discussion is given in [Rhe80]. Further, refinement strategies are presented in [Axe93] for Newton-type methods, in [KS90] for quasi-Newton methods, and in [HLS91] for the Gauss–Newton method.

The paper is organized as follows. In section 2 the augmented Lagrangian-SQP method is proposed and a convergence result is presented, which states the second-order rate of convergence. In section 3 we treat the internal approximation of the

algorithm. We prove the converge and rate of convergence for the discretized version of the augmented Lagrangian-SQP method. The mesh-independence principle is developed in section 4. As an application we present numerical examples for optimal control problems for the Burgers equation in sections 5 and 6, where the cost functional is of tracking type.

It is appropriate to introduce some notations that will be used throughout the paper. Let $(V, \|\cdot\|_V)$ and $(W, \|\cdot\|_W)$ be a normed linear space. The set $B(v;r)$ denotes an open ball of radius $r > 0$ centered at the point $v \in V$. The open set $U \subset V$ is called a neighborhood of $v \in U$ if $B(v;r) \subset U$ for some $r > 0$. By $\mathcal{L}(V,W)$ we denote the normed linear space of all bounded linear operators from V into W and set $\mathcal{L}(V) = \mathcal{L}(V,V)$. For $A \in \mathcal{L}(V,W)$ the set $A(V) \subseteq W$ is called the range of A and is denoted by ran A. The set $\{v \in V : Av = 0\}$ is called the kernel of A and is denoted by ker A. For arbitrary $T > 0$ and B a Banach space $L^2(0,T;B)$ denotes the (equivalence class of) square integrable functions in the sense of Bochner.

2. The augmented Lagrangian-SQP method. We consider the constrained minimization problem

$$\text{(P)} \qquad \text{minimize } J(x) \text{ subject to } e(x) = 0,$$

where $J : X \to \mathbb{R}$, $e : X \to Y$ with X and Y real Hilbert spaces. For brevity we set $Z = X \times Y$. The space Z is endowed with the Hilbert space product topology. The Fréchet-derivatives with respect to the variable x will be denoted by primes. We do not distinguish by notation between a functional in the dual and its Riesz representation in the Hilbert space. Let us start with two examples that motivate our work.

Example 2.1. Let $\Omega = (0,1)$, $\Omega_\circ$ a nonempty subset of Ω with positive measure, and $f \in L^2(\Omega)$. For a control $u \in L^2(\Omega_\circ)$ the state $y \in H^1(\Omega)$ is given by the solution of the stationary Burgers equation

$$\begin{cases} -\nu y'' + yy' = f + Bu & \text{in } \Omega, \\ y(0) = y_l,\ y(1) = y_r, \end{cases}$$

where $B \in \mathcal{L}(L^2(\Omega_\circ), L^2(\Omega))$ is an extension operator defined by

$$Bq = \begin{cases} q & \text{in } \Omega_\circ, \\ 0 & \text{in } \Omega \setminus \Omega_\circ. \end{cases}$$

For controls $u \in L^2(\Omega_\circ)$ we associate the cost of tracking type

$$J(y,u) = \frac{1}{2} \int_\Omega |y - z|^2 \, dx + \frac{\alpha}{2} \int_{\Omega_\circ} |u|^2 \, dx,$$

where $z \in L^2(\Omega)$ and $\alpha > 0$ is fixed. We introduce the operator $e = (e_1, e_2, e_3) : H^1(\Omega) \times L^2(\Omega_\circ) \to H_0^1(\Omega) \times \mathbb{R}^2$ by

$$e(y,u) = \left((-\Delta)^{-1}(-\nu y'' + yy' - f - Bu), y(0) - y_l, y(1) - y_r\right),$$

where Δ denotes the Laplace operator from $H_0^1(\Omega)$ to $H^{-1}(\Omega)$. The resulting optimal control problem is in the form (**P**) with $X = H^1(\Omega) \times L^2(\Omega_\circ)$ and $Y = H^{-1}(\Omega) \times \mathbb{R}^2$.

Example 2.2. Let Ω be the interval $(0,1)$ and $Q = (0,T] \times \Omega$ for given $T > 0$. Moreover, let $f \in L^2(0,T;L^2(\Omega))$ and $\phi \in L^2(\Omega)$. We define the space $W(0,T)$ by

$$W(0,T) = \left\{\varphi : \varphi \in L^2(0,T;H_0^1(\Omega)),\ \varphi_t \in L^2(0,T;H^{-1}(\Omega))\right\},$$

which is a Hilbert space endowed with the common inner product. For a control $u \in L^2(0,T;L^2(\Omega))$ the state $y \in W(0,T)$ is given by the weak solution of the unsteady Burgers equation

$$\begin{aligned} y_t - \nu y_{xx} + y y_x &= f + u \quad \text{in } Q, \\ y(t,0) = y(t,1) &= 0 \quad \text{for } t \in (0,T) \text{ a.e.}, \\ y(0,x) &= \phi(x) \quad \text{for all } x \in \Omega, \end{aligned} \tag{2.1}$$

where $\nu > 0$ denotes the viscosity parameter. This equation was extensively developed by Burgers as a simplified fluid flow model which nevertheless exhibits some of the important aspects of turbulence [Bur40]. Later it was derived by Lighthill as a second-order approximation to the one-dimensional unsteady Navier–Stokes equations [Lig56].

With controls $u \in L^2(0,T;L^2(\Omega))$ we associate the cost of tracking type

$$J(y,u) = \frac{1}{2}\|y - z\|^2_{L^2(0,T;L^2(\Omega))} + \frac{\alpha}{2}\|u\|^2_{L^2(0,T;L^2(\Omega))},$$

where $z \in W(0,T)$ and $\alpha > 0$ is fixed. Setting $X = W(0,T) \times L^2(0,T;L^2(\Omega))$ and $Y = L^2(0,T;H_0^1(\Omega)) \times L^2(\Omega)$ and defining $e = (e_1, e_2)$ by

$$e(y,u) = \left((-\Delta)^{-1}\left(y_t - y_{xx} + y y_x - f - u\right), y(0,\cdot) - \phi\right),$$

where the Laplace operator Δ was introduced in Example 2.2, the optimal control problem can be written in the form (P).

ASSUMPTION 1. *Problem* (P) *has a local solution x^*; i.e., there exists an $\varepsilon > 0$ such that $J(x) \geq J(x^*)$ for all $x \in B(x^*;\varepsilon)$ satisfying $e(x) = 0$. Further, J and e are twice continuously Fréchet-differentiable, and the mappings J'' and e'' are Lipschitz-continuous in the neighborhood $U_1 \subset X$ of x^*.*

Remark 2.3. It is proved in [Vol97] that Assumption 1 holds for Example 2.2 and Example 2.1.

The Lagrangian functional $L : Z \to \mathbb{R}$ associated with (P) is given by

$$L(x,\lambda) = J(x) + \langle e(x), \lambda \rangle_Y,$$

where $\langle \cdot, \cdot \rangle_Y$ denotes the inner product in Y.

ASSUMPTION 2. *The linearization $e'(x^*)$ of e at x^* is surjective.*

Remark 2.4. Assumption 2 is satisfied for Examples 2.2 and 2.1; see [Vol97].

With Assumption 2 holding there exists a Lagrange multiplier $\lambda^* \in Y$ such that the following first-order necessary optimality condition is satisfied [Lue69]:

$$L'(x^*,\lambda^*) = 0\,, \quad e(x^*) = 0. \tag{2.2}$$

We also make use of the second-order sufficient optimality condition.

ASSUMPTION 3. *The operator $L''(x^*,\lambda^*)$ is coercive on the kernel of $e'(x^*)$; i.e., there exists a constant $\kappa > 0$ such that*

$$\langle L''(x^*,\lambda^*)v, v\rangle_X \geq \kappa \|v\|^2_X \text{ for all } v \in \ker e'(x^*),$$

where $\langle \cdot, \cdot \rangle_X$ denotes the inner product in X.

Remark 2.5. For Example 2.2 and Example 2.1 it is shown in [Vol97] that the second-order sufficient optimality condition is satisfied if the residues $\|y^* - z\|_{L^2(Q)}$ and $\|y^* - z\|_{L^2(\Omega)}$, respectively, are sufficiently small.

Let J and e be twice continuously Fréchet-differentiable in U_1. With Assumption 3 holding, the solution x^* of (2.2) is a locally unique solution to (P) [MZ79]. Due to Assumption 2 we can further assume that there exists a neighborhood $U_2 \subseteq U_1$ of x^* such that $e'(x)$ is surjective for all $x \in U_2$. Assumption 3 implies that there is a neighborhood $U_3 \subset Z$ of (x^*, λ^*) such that

$$\langle L''(x,\lambda)v, v\rangle_X \geq \kappa_\circ \|v\|_X^2 \text{ for all } v \in \ker e'(x)$$

for $(x,\lambda) \in U_3$ and for a constant $\kappa_\circ > 0$. Henceforth we denote by $U = U_X \times U_Y$ with $U_X \subset X$ and $U_Y \subset Y$ a neighborhood of (x^*, λ^*) such that

(a) J and e are twice Fréchet-differentiable on $\overline{U_X}$ and their second Fréchet-derivatives are Lipschitz-continuous,
(b) $e'(x)$ is surjective, and
(c) $L''(x,\lambda)$ is coercive on the kernel of $e'(x)$ for all $(x,\lambda) \in U$.

For any $c \geq 0$ the augmented Lagrangian functional is defined by

$$L_c(x,\lambda) = L(x,\lambda) + \frac{c}{2}\|e(x)\|_Y^2.$$

It will be convenient to introduce the matrix of operators

$$M(x,\lambda) = \begin{pmatrix} L''(x,\lambda) & e'(x)^* \\ e'(x) & 0 \end{pmatrix} \text{ for all } (x,\lambda) \in \overline{U}, \tag{2.3}$$

where $e'(x)^* : Y \to X$ denotes the adjoint of $e'(x)$ in $\overline{U_X}$. To find x^* numerically we observe that the first-order necessary optimality condition (2.2) implies that (x^*, λ^*) is a solution to

$$L_c'(x,\lambda) = L'(x, \lambda + ce(x)) = 0, \quad e(x) = 0 \quad \text{for all } c \geq 0. \tag{OS}$$

Remark 2.6. With Assumptions 1–2 holding, (x^*, λ^*) is a locally unique solution to (OS).

We solve problem (OS) by the Newton method, where we avoid the explicit calculation of the augmented Lagrangian functional and its derivatives [IK96b].

Algorithm 1.

(a) *Choose* $(x^0, \lambda^0) \in U$, $c \geq 0$ *and put* $n = 0$.
(b) *Set* $\tilde{\lambda}^n = \lambda^n + ce(x^n)$.
(c) *Solve for* $(\bar{x}, \bar{\lambda})$

$$M(x^n, \tilde{\lambda}^n)\begin{pmatrix} \bar{x} - x^n \\ \bar{\lambda} - \tilde{\lambda}^n \end{pmatrix} = -\begin{pmatrix} L'(x^n, \tilde{\lambda}^n) \\ e(x^n) \end{pmatrix}. \tag{2.4}$$

(d) *Set* $(x^{n+1}, \lambda^{n+1}) = (\bar{x}, \bar{\lambda})$, $n = n+1$ *and go back to* (b).

Remark 2.7.

(a) Since X and Y are Hilbert spaces, it can be shown that the update $(\bar{x}, \bar{\lambda})$ of Algorithm 1 can equivalently be obtained from

$$\begin{pmatrix} L_c''(x^n, \lambda^n) & e'(x^n)^* \\ e'(x^n) & 0 \end{pmatrix}\begin{pmatrix} \bar{x} - x^n \\ \bar{\lambda} - \lambda^n \end{pmatrix} = -\begin{pmatrix} L_c'(x^n, \lambda^n) \\ e(x^n) \end{pmatrix},$$

which corresponds to a Newton step applied to (OS). This form of the iteration requires the implementation of $e(x^n)^* e'(x^n)$, whereas steps (b) and (c) of Algorithm 1 do not. In the case of Examples 2.1 and 2.2 this implies an additional solve of the Poisson equation.

(b) If we set $c = 0$, Algorithm 1 becomes the SQP-method.

In each iteration of Algorithm 1 the saddle-point problem (2.4) needs to be solved. With Assumptions 2 and 3 holding, problem (2.4) has a unique solution $(\bar{x}, \bar{\lambda})$ for all $(x^n, \tilde{\lambda}^n) \in U$ (see [BF91]) and there exists a constant $\eta > 0$ satisfying

$$\|M^{-1}(x, \lambda)\|_{\mathcal{L}(Z)} \leq \eta \text{ for all } (x, \lambda) \in U.$$

To formulate the convergence result for Algorithm 1 we introduce the following constants. Let $r_{\max} > 0$ denote the largest radius of a ball centered at (x^*, λ^*) and contained in the neighborhood U, γ_e stands for the uniform Lipschitz-constant of e in U_X, and γ_M denotes the uniform Lipschitz-constant of M in U. Moreover, we set

$$r^* = \min\left(\frac{r_{\max}}{\sqrt{\max(2, 1 + 2c^2\gamma_e^2)}}, \frac{2}{\gamma_M \eta \max(2, 1 + 2c^2\gamma_e^2)}\right). \tag{2.5}$$

For the proof of the next theorem we refer the reader to [IK96b].

THEOREM 2.8. *Let Assumptions* 1, 2, *and* 3 *hold and let the starting value* (x^0, λ^0) *belong to the open ball* $B((x^*, \lambda^*); r^*)$. *Then Algorithm* 1 *is well-defined and*

$$\|(x^{n+1}, \lambda^{n+1}) - (x^*, \lambda^*)\|_Z \leq \widetilde{C} \, \|(x^n, \lambda^n) - (x^*, \lambda^*)\|_Z^2, \tag{2.6}$$

where $\widetilde{C} = \frac{1}{2} \gamma_M \eta \max(2, 1 + 2c^2\gamma_e^2)$.

Remark 2.9.

(a) With Assumptions 1–3 holding, Algorithm 1 has second-order rate of convergence, where the convergence rate factor $\widetilde{C}$ depends quadratically on c.

(b) We shall need the radius r^* in the formulation of Corollary 3.19.

(c) From (2.5) and (2.6) we derive that $(x^n, \lambda^n) \in U$ holds for all $n = 0, 1, \ldots$ and that the sequence $\{\|(x^n, \lambda^n) - (x^*, \lambda^*)\|_Z\}_{n \in \mathbb{N}}$ is strictly decreasing.

3. Internal approximation and convergence theorems. This section is devoted to the internal approximation of Algorithm 1. In many applications X and Y are infinite dimensional Hilbert spaces, which have to be discretized for numerical realization of (P); see Examples 2.2 and 2.1. For this purpose we suppose that we are given a family $\{h\}$, $h > 0$, with accumulation point zero and families $\{X_h\}_h$ and $\{Y_h\}_h$ of Hilbert spaces with finite dimension. In practice, the parameter h is called the mesh-size of the finite dimensional spaces and h varies over a sequence. To shorten notation we set $Z_h = X_h \times Y_h$ for every h.

Let $p_h^X \in \mathcal{L}(X_h, X)$ and $p_h^Y \in \mathcal{L}(Y_h, Y)$ be given injective prolongations. In addition, we introduce surjective restrictions $r_h^X \in \mathcal{L}(X, X_h)$ and $r_h^Y \in \mathcal{L}(Y, Y_h)$. For brevity we set $p_h = (p_h^X, p_h^Y)$ and $r_h = (r_h^X, r_h^Y)$. For every h the data (X_h, p_h^X, r_h^X), (Y_h, p_h^Y, r_h^Y), and (Z_h, p_h, r_h) are called approximations of the spaces X, Y, and Z, respectively.

Remark 3.1. Let us turn to Example 2.1. For $m \in \mathbb{N}$ the mesh-size $h = h(m)$ and the associated grid-points $x_i \in [0, 1]$ are given by $h = \frac{1}{m}$ and $x_i = ih$ for $i = 0, \ldots, m$, respectively. Let $\varphi_0, \ldots, \varphi_m \in H^1(0, 1)$ denote the well-known piecewise linear finite elements satisfying $\varphi_i(x_j) = \delta_{ij}$ for $0 \leq i, j \leq m$. We will restrict our discussions to the case that the set $\Omega_\circ = (a, b) \subseteq \Omega$ is an open interval. The numbers $i_a, i_b \in \{1, \ldots, m-1\}$ are defined by

$$a < x_{i_a} < x_{i_a+1} < \cdots < x_{i_b} < b.$$

For the approximation of the Hilbert spaces we set

$$\begin{aligned} X_h &= \operatorname{Span}\{\varphi_0, \dots, \varphi_m\} \times \operatorname{Span}\{\varphi_{i_a}, \dots, \varphi_{i_b}\}, \\ Y_h &= \operatorname{Span}\{\varphi_1, \dots, \varphi_{m-1}\} \times \mathbb{R}^2, \\ Z_h &= X_h \times Y_h. \end{aligned}$$

Now we are going to define the prolongations and restrictions. We set

$$p_h(y_h, u_h, \lambda_h, \mu, \xi) = (y_h, u_h, \lambda_h, \mu, \xi) \quad \text{for } (y_h, u_h) \in X_h \text{ and } (\lambda_h, \mu, \xi) \in Y_h.$$

Hence, the finite dimensional Hilbert spaces X_h and Y_h are endowed with the inner products and norms introduced in X and Y, respectively. We introduce the restriction

$$r_h(y, u, \lambda, \mu, \xi) = \left(\sum_{i=0}^{m} y(x_i)\varphi_i, \frac{1}{h} \sum_{i=i_a}^{i_b} \int_{x_i - \frac{h}{2}}^{x_i + \frac{h}{2}} u \, dx \, \varphi_i, \sum_{i=1}^{m-1} \lambda(x_i)\varphi_i, \mu, \xi \right)$$

for $(y, u) \in X$ and $(\lambda, \mu, \xi) \in Y$. Note that $u \in L^2(\Omega_\circ)$ holds so that $u(x_i)$ need not be meaningful.

With the prolongations we define discrete inner products and discrete norms in the following way.

Definition 3.2. *We introduce the* discrete inner product *in X_h by*

$$\langle x_h, \bar{x}_h \rangle_{X_h} = \langle p_h^X x_h, p_h^X \bar{x}_h \rangle_X$$

and the discrete norm *in X_h by*

$$\|x_h\|_{X_h} = \|p_h^X x_h\|_X$$

for all $x_h, \bar{x}_h \in X_h$. Analogously, for all $\lambda_h, \bar{\lambda}_h \in Y_h$ we define by

$$\langle \lambda_h, \bar{\lambda}_h \rangle_{Y_h} = \langle p_h^Y \lambda_h, p_h^Y \bar{\lambda}_h \rangle_Y, \quad \|\lambda_h\|_{Y_h} = \|p_h^Y \lambda_h\|_Y$$

the discrete inner product and discrete norm in Y_h. In Z_h the corresponding inner product is the sum of the inner products $\langle \cdot, \cdot \rangle_{X_h}$ and $\langle \cdot, \cdot \rangle_{Y_h}$ and the discrete norm is

$$\|(x_h, \lambda_h)\|_{Z_h} = \|p_h(x_h, \lambda_h)\|_Z \text{ for all } (x_h, \lambda_h) \in Z_h.$$

Remark 3.3. For these discrete norms p_h^X, p_h^Y, and p_h have norm 1.

Let us introduce the restriction $\hat{r}_h^X = (p_h^X)^* \in \mathcal{L}(X, X_h)$, i.e.,

$$\langle \hat{r}_h^X x, x_h \rangle_{X_h} = \langle x, p_h^X x_h \rangle_X \text{ for all } x \in X \text{ and } x_h \in X_h.$$

Analogously, $\hat{r}_h^Y$ is defined, and we put $\hat{r}_h = (\hat{r}_h^X, \hat{r}_h^Y)$. As p_h is injective and $\operatorname{ran} \hat{r}_h$ has finite dimension, we infer that $\hat{r}_h$ is surjective.

Lemma 3.4. *For every h the restriction $\hat{r}_h$ has the following properties:*

(a) *$\hat{r}_h p_h(x_h, \lambda_h) = (x_h, \lambda_h)$ for all $(x_h, \lambda_h) \in Z_h$ and*

(b) *$p_h \hat{r}_h$ is the orthogonal projector of Z onto the subspace* $\operatorname{ran} p_h$*, i.e.*

$$\|(x, \lambda) - p_h \hat{r}_h(x, \lambda)\|_Z = \inf_{(x_h, \lambda_h) \in Z_h} \|(x, \lambda) - p_h(x_h, \lambda_h)\|_Z \text{ for all } (x, \lambda) \in Z.$$

Proof.
(a) We derive from $\hat{r}_h = (p_h)^*$ and Definition 3.2 that

$$\langle \hat{r}_h p_h(x_h, \lambda_h), (\bar{x}_h, \bar{\lambda}_h) \rangle_{Z_h} = \langle (x_h, \lambda_h), (\bar{x}_h, \bar{\lambda}_h) \rangle_{Z_h}$$

for all $(x_h, \lambda_h), (\bar{x}_h, \bar{\lambda}_h) \in Z_h$. Hence, part (a) is shown.
(b) Any $(x, \lambda) \in Z$ can be expressed as

$$(x, \lambda) = [(x, \lambda) - p_h \hat{r}_h(x, \lambda)] + p_h \hat{r}_h(x, \lambda).$$

From $\hat{r}_h = (p_h)^*$ and part (a) we infer that

$$\langle (x, \lambda) - p_h \hat{r}_h(x, \lambda), p_h \hat{r}_h(x, \lambda) \rangle_Z = 0$$

such that $p_h \hat{r}_h$ is the orthogonal projector of Z onto the subspace ran p_h. □

Remark 3.5. The restrictions $\hat{r}_h^X$, $\hat{r}_h^Y$, and $\hat{r}_h$ are called the optimal restrictions associated with the prolongations p_h^X, p_h^Y, and p_h, respectively [Aub72]. Remark 3.3 implies that the norms of the optimal restrictions are equal to 1 [DL88].

DEFINITION 3.6. *The approximations (X_h, p_h^X, r_h^X) of X are called* convergent *if*

$$\lim_{h \to 0} \| x - p_h^X r_h^X x \|_X = 0 \text{ for all } x \in X.$$

Convergence of (Y_h, p_h^Y, r_h^Y) and (Z_h, p_h, r_h) is defined analogously.

ASSUMPTION 4. *The approximations (Z_h, p_h, r_h) of the Hilbert space Z are convergent, and $p_h r_h(x, \lambda) \in U$ holds for $(x, \lambda) \in U$ and all h sufficiently small.*

Remark 3.7. In Remark 3.1 we introduced approximations for Example 2.1, which are convergent [Aub72]. Hence, Assumption 4 is satisfied.

Remark 3.8. With Assumption 4 holding, $\| p_h r_h(x, \lambda) \|_Z$ is bounded for all $(x, \lambda) \in Z$. According to the principle of uniform boundedness [Wou79] there exists a constant $c_\circ > 0$ such that $\| p_h r_h \|_{\mathcal{L}(Z)} \leq c_\circ$ for all h.

Definition 3.6 leads directly to the following corollary.

COROLLARY 3.9. *The approximations (Z_h, p_h, r_h) of Z are convergent if and only if the approximations (X_h, p_h^X, r_h^X) of X and (Y_h, p_h^Y, r_h^Y) of Y are convergent.*

Now we define the following internal approximations of the cost functional J and the constraint e [Aub72]:

$$J_h = J\, p_h^X : X_h \to \mathbb{R}, \quad e_h = \hat{r}_h^Y\, e\, p_h^X : X_h \to Y_h. \tag{3.1}$$

Remark 3.10. Using (3.1) and $\hat{r}_h^Y = (p_h^Y)^*$ we have

$$\langle e_h(x_h), \lambda_h \rangle_{Y_h} = \langle e(p_h^X x_h), p_h^Y \lambda_h \rangle_Y \text{ for all } (x_h, \lambda_h) \in X_h \times Y_h. \tag{3.2}$$

From (3.2) we observe that the restriction $\hat{r}_h$ is more a theoretical tool, whereas the operator r_h is used in the numerical implementation.

Together with (P) we investigate the finite dimensional discrete problems

$$\text{minimize } J_h(x_h) \text{ subject to } e_h(x_h) = 0. \tag{P_h}$$

ASSUMPTION 5. *For every h there exist neighborhoods $V_h \subset X_h$ of $r_h^X x^*$ and $W_h \subset Y_h$ of $r_h^Y \lambda^*$ and a constant $r_{\min} > 0$ independent of h such that $p_h(U_h) \subset U$ and $B(r_h(x^*, \lambda^*); r_{\min})$ is a subset of $U_h = V_h \times W_h$ for all h.*

The Lagrangian for (P_h) is given by

$$L_h(x_h, \lambda_h) = J_h(x_h) + \langle e_h(x_h), \lambda_h \rangle_{Y_h}.$$

From $\hat{r}_h^Y = (p_h^Y)^*$ and (3.1) we derive that $L_h = L\, p_h$. Since all restrictions and prolongations are linear, we are able to differentiate the Lagrangian L_h with respect to the variable x_h for all $x_h \in V_h$ and $\lambda_h \in Y_h$:

$$L_h'(x_h, \lambda_h) = \hat{r}_h^X L'(p_h(x_h, \lambda_h)).$$

Both $L_h'(x_h, \lambda_h)$ and $e_h(x_h)$ are also continuously Fréchet-differentiable with respect to the variable x_h and their Fréchet-derivatives are Lipschitz-continuous. Explicitly, we get

$$L_h''(x_h, \lambda_h) = \hat{r}_h^X L''(p_h(x_h, \lambda_h)) p_h^X, \quad e_h'(x_h) = \hat{r}_h^Y e'(p_h^X x_h) p_h^X \tag{3.3}$$

for all $x_h \in V_h$ and $\lambda_h \in Y_h$. Further, we derive that

$$e_h'(x_h)^* = \hat{r}_h^X e'(p_h^X x_h)^* p_h^Y \text{ for all } x_h \in V_h.$$

For every h the matrix of operators M given in (2.3) is approximated by

$$M_h(x_h, \lambda_h) = \hat{r}_h\, M(p_h(x_h, \lambda_h))\, p_h = \begin{pmatrix} L_h''(x_h, \lambda_h) & e_h'(x_h)^* \\ e_h'(x_h) & 0 \end{pmatrix} \tag{3.4}$$

for all $(x_h, \lambda_h) \in U_h$. Since L'' is a self-adjoint operator in U we conclude that L_h'' is self-adjoint in U_h. Hence M_h is self-adjoint in U_h.

The approximation of (OS) is formulated as

$$L_h'(x_h, \lambda + c e_h(x_h)) = 0, \quad e_h(x_h) = 0 \quad \text{for all } c \geq 0, \tag{OS_h}$$

which is the first-order necessary optimality condition for (P_h).

We shall require a uniform Babuška–Brezzi condition in V_h.

Assumption 6. *There exists a constant $\beta^* > 0$ independent of h with*

$$\inf_{\bar{\lambda}_h \in Y_h} \sup_{\bar{x}_h \in X_h} \frac{\langle e_h'(x_h)^* \bar{\lambda}_h, \bar{x}_h \rangle_{X_h}}{\|\bar{x}_h\|_{X_h} \|\bar{\lambda}_h\|_{Y_h}} \geq \beta^* \text{ for all } x_h \in V_h \text{ and for all } h. \tag{3.5}$$

Remark 3.11. By closed range theory [Bre87] the uniform Babuška–Brezzi condition implies that the operator $e_h'(y_h, u_h)$ is surjective on V_h. Note that (3.5) leads to

$$\|e_h'(x_h)^* \lambda_h\|_{X_h} \geq \beta^* \|\lambda_h\|_{Y_h} \text{ for all } \lambda_h \in Y_h.$$

We refer the reader to [Vol99] for the proof of the uniform Babuška–Brezzi condition in the case of Example 2.1 and the piecewise linear approximations introduced in Remark 3.1.

We shall also make use of a uniform second-order sufficient optimality condition in U_h.

Assumption 7. *There exists $\kappa^* > 0$ independent of h such that*

$$\langle L_h''(x_h, \lambda_h) \bar{x}_h, \bar{x}_h \rangle_{X_h} \geq \kappa^* \|\bar{x}_h\|_{X_h}^2 \text{ for all } \bar{x}_h \in \ker e_h'(x_h)$$

for all $(x_h, \lambda_h) \in U_h$ and for all h.

Remark 3.12. Using (3.3) we have

$$\langle L_h''(r_h(x^*,\lambda^*))x_h, x_h\rangle_{X_h} = \langle L''(p_h r_h(x^*,\lambda^*))p_h^X x_h, p_h^X x_h\rangle_X \text{ for all } x_h \in X_h.$$

If the operator $L''(x^*,\lambda^*)$ is coercive on the whole space X and Assumption 4 holds, then Assumption 7 is satisfied in a neighborhood of $r_h(x^*,\lambda^*)$ if h is sufficiently small.

Now we consider the internal approximation of Algorithm 1.

ALGORITHM 2.

(a) *Choose* $(x_h^0,\lambda_h^0) \in U_h$, $c \geq 0$ *and put* $n = 0$.

(b) *Set* $\tilde{\lambda}_h^n = \lambda_h^n + ce_h(x_h^n)$.

(c) *Solve for* $(\bar{x}_h, \bar{\lambda}_h)$

$$M_h(x_h^n, \tilde{\lambda}_h^n)\begin{pmatrix} \bar{x}_h - x_h^n \\ \bar{\lambda}_h - \tilde{\lambda}_h^n \end{pmatrix} = -\begin{pmatrix} L_h'(x_h^n, \tilde{\lambda}_h^n) \\ e_h(x_h^n) \end{pmatrix}. \tag{3.6}$$

(d) *Set* $(x_h^{n+1}, \lambda_h^{n+1}) = (\bar{x}_h, \bar{\lambda}_h)$, $n = n+1$, *and go back to* (b).

Remark 3.13. Since the matrix M_h has the same structure as M we have to solve a saddle-point problem in (3.6). But now the linear system (3.6) has finite dimension.

LEMMA 3.14. *With Assumption* 6 *holding, the saddle-point problem*

$$M_h(x_h,\lambda_h)\begin{pmatrix} \bar{x}_h - x_h \\ \bar{\lambda}_h - \lambda_h \end{pmatrix} = -\begin{pmatrix} L_h'(x_h,\lambda_h) \\ e_h(x_h) \end{pmatrix}$$

has a unique solution $(\bar{x}_h, \bar{\lambda}_h)$ *for every* $(x_h,\lambda_h) \in U_h$.

For the proof we refer the reader to [GR92].

Remark 3.15. With Assumption 6 holding there exists a bound $\eta^* > 0$, which may depend on β^* or κ^*, but is independent of h satisfying

$$\|M_h^{-1}(r_h(x^*,\lambda^*))\|_{\mathcal{L}(Z_h)} \leq \eta^* \text{ for all } h. \tag{3.7}$$

Using Remark 3.5 and (3.1) we derive

$$\|e_h(x_h) - e_h(\bar{x}_h)\|_{Y_h} \leq \gamma_e \|x_h - \bar{x}_h\|_{X_h},$$

and from (3.4) we infer that

$$\|M_h(x_h,\lambda_h) - M_h(\bar{x}_h,\bar{\lambda}_h)\|_{\mathcal{L}(Z_h)} \leq \gamma_M \|(x_h,\lambda_h) - (\bar{x}_h,\bar{\lambda}_h)\|_{Z_h}$$

for all $(x_h,\lambda_h),(\bar{x}_h,\bar{\lambda}_h) \in U_h$. As a consequence, uniform Lipschitz-constants of e_h in V_h and M_h in U_h can be chosen of the uniform Lipschitz-properties of e and M in U_X and U, respectively.

In many applications it turns out that the solution (x^*,λ^*) of (OS) as well as the iterates (x^n,λ^n) of Algorithm 1 have "better smoothness" properties than the elements of Z. This is a motivation for the following assumption.

ASSUMPTION 8. *There are bounded subsets* $V \subset X$ *and* $W \subset Y$ *such that*

$$x^*,\ x^n,\ x^n - x^*,\ x^{n+1} - x^n \in V$$

and

$$\lambda^*,\ \lambda^n,\ \lambda^n - \lambda^*,\ \lambda^{n+1} - \tilde{\lambda}^n \in W$$

holds for all $n = 0, 1, \ldots$.

Moreover, there exists a bounded function $\rho : [0,1] \to [0,\infty)$ *which is right-continuous at* $h = 0$ *with* $\rho(0) = 0$ *satisfying*

$$\|(x,\lambda) - p_h r_h(x,\lambda)\|_Z \le \rho(h)$$

for all $(x,\lambda) \in V \times W$ *and for all* $h > 0$.

Let

$$F(x,\lambda) = \begin{pmatrix} L'(x,\lambda) \\ e(x) \end{pmatrix} \text{ and } F_h = \hat{r}_h F p_h. \tag{3.8}$$

The next lemma shows that the internal approximations of F and M given by F_h and M_h, respectively, are consistent of order $\rho(h)$.

LEMMA 3.16. *With Assumptions* 4 *and* 8 *holding, there exists a constant* $\tilde{C} > 0$ *independent of* h *such that*

$$\|\hat{r}_h F(x,\lambda) - F_h(r_h(x,\lambda))\|_{Z_h} \le \tilde{C}\rho(h)$$

and

$$\|\hat{r}_h M(x,\lambda)(\bar{x},\bar{\lambda})^T - M_h(r_h(x,\lambda))[r_h(\bar{x},\bar{\lambda})]^T\|_{Z_h} \le \tilde{C}\rho(h)$$

for all $(x,\lambda) \in (V \times W) \cap U$, $(\bar{x},\bar{\lambda}) \in V \times W$ *and for all* $h > 0$.

Proof. Let $x, \bar{x} \in V$ and $\lambda, \bar{\lambda} \in W$. By Assumption 8 we conclude

$$\|(x,\lambda) - p_h r_h(x,\lambda)\|_Z \le \rho(h) \text{ and } \|(\bar{x},\bar{\lambda}) - p_h r_h(\bar{x},\bar{\lambda})\|_Z \le \rho(h)$$

and $\|(\bar{x},\bar{\lambda})\|_Z \le \bar{c}$ for a constant $\bar{c} > 0$. Using Remark 3.5 and setting $\tilde{C} = \max(\gamma_F, \bar{c}\gamma_M + \eta_M)$, where $\gamma_F > 0$ denotes the Lipschitz-constant of F in U and η_M is a uniform upper bound of M in $\overline{U}$, we derive from (3.8) that

$$\|\hat{r}_h F(x,\lambda) - F_h(r_h(x,\lambda))\|_{Z_h} \le \gamma_F \|(x,\lambda) - p_h r_h(x,\lambda)\|_Z \le \tilde{C}\rho(h)$$

and by triangle inequality and (3.4) that

$$\begin{aligned}&\|\hat{r}_h M(x,\lambda)(\bar{x},\bar{\lambda})^T - M_h(r_h(x,\lambda))[r_h(\bar{x},\bar{\lambda})]^T\|_{Z_h} \\ &\quad \le \gamma_M \|(x,\lambda) - p_h r_h(x,\lambda)\|_Z \|(\bar{x},\bar{\lambda})\|_Z + \eta_M \|(\bar{x},\bar{\lambda}) - p_h r_h(\bar{x},\bar{\lambda})\|_Z \le \tilde{C}\rho(h). \quad \square\end{aligned}$$

Now we are going to formulate the convergence results.

THEOREM 3.17. *Let Assumptions* 1–8 *hold. Then there is a constant* $\bar{h} \in (0,1]$ *so that* (OS$_h$) *has a locally unique solution satisfying*

$$(x_h^*, \lambda_h^*) = r_h(x^*,\lambda^*) + O(\rho(h)) \text{ for all } h \in (0,\bar{h}]. \tag{3.9}$$

Proof. Since Algorithms 1 and 2 are equivalent to the Newton method applied to the operator equations (OS$_h$) and (OS), respectively, the theorem follows directly from Theorem 2 in [ABPR86]. $\square$

Remark 3.18. Since ρ is right-continuous at $h = 0$ with $\rho(0) = 0$, we derive from (3.9) that

$$\lim_{h\to 0} \|(x_h^*, \lambda_h^*) - r_h(x^*,\lambda^*)\|_{Z_h} = 0$$

holds.

From existence of (x_h^*, λ_h^*) the second-order convergence of Algorithm 2 is proved for starting values $(x_h^0, \lambda_h^0) = r_h(x^0, \lambda^0)$, where (x^0, λ^0) is the starting value of Algorithm 1 chosen in an appropriate neighborhood of (x^*, λ^*). This is formulated precisely in the next corollary. Note that we introduced the radius r^* in (2.5).

COROLLARY 3.19. *Let the assumptions of Theorem* 3.17 *hold. Then there exist* $\tilde{h} \in (0, \bar{h}]$ *and* $r_1 \in (0, r^*]$ *such that Algorithm* 2 *converges to* (x_h^*, λ_h^*) *for starting values* $(x_h^0, \lambda_h^0) = r_h(x^0, \lambda^0)$ *with* $(x^0, \lambda^0) \in (V \times W) \cap B((x^*, \lambda^*); r_1)$*, and the iterates satisfy*

$$\|(x_h^{n+1}, \lambda_h^{n+1}) - (x_h^*, \lambda_h^*)\|_{Z_h} \le C \, \|(x_h^n, \lambda_h^n) - (x_h^*, \lambda_h^*)\|_{Z_h}^2$$

for a constant $C > 0$ *independent of* $h \in (0, \tilde{h}]$ *and* n.

Proof. We shall apply Theorem 2.8 to the discrete Algorithm 2 in order to prove the claim of the corollary. From Theorem 3.17 we conclude the existence of a locally unique (x_h^*, λ_h^*) solving (OS_h), of a constant $\tilde{C} > 0$, and of $\bar{h} > 0$ such that

$$\|(x_h^*, \lambda_h^*) - r_h(x^*, \lambda^*)\|_{Z_h} \le \tilde{C} \, \rho(h) \text{ for every } h \in (0, \bar{h}]. \tag{3.10}$$

Applying the perturbation lemma (see [OR70]) we obtain that

$$\|M_h^{-1}(x_h, \lambda_h)\|_{\mathcal{L}(Z_h)} \le \frac{\eta^*}{1 - \eta^* \gamma_M \, \|(x_h, \lambda_h) - r_h(x^*, \lambda^*)\|_{Z_h}} < \frac{\eta^*}{2}$$

for all $(x_h, \lambda_h) \in B(r_h(x^*, \lambda^*), \frac{1}{2\eta^* \gamma_M})$. Let $(x^0, \lambda^0) \in U$ denote a starting value of Algorithm 1 satisfying

$$\|(x^0, \lambda^0) - (x^*, \lambda^*)\|_Z < \min\left(r^*, \frac{3r_{\min}}{8c_\circ \sqrt{\max(2, 1 + c^2\gamma_e^2)}}, \frac{2}{c_\circ \gamma_M \eta^* \max(2, 1 + c^2\gamma_e^2)} \right), \tag{3.11}$$

where the constant $c_\circ$ is specified in Remark 3.8. Since ρ is right-continuous at $h = 0$ there exists $\tilde{h} \in (0, \bar{h}]$ such that

$$\rho(h) \le \min\left(\frac{r_{\min}}{4\tilde{C}}, \frac{3r_{\min}}{8\tilde{C}\sqrt{\max(2, 1 + c^2\gamma_e^2)}}, \frac{2}{\gamma_M \eta^* \tilde{C} \max(2, 1 + c^2\gamma_e^2)} \right) \tag{3.12}$$

for all $h \in (0, \tilde{h}]$. Applying (3.11) yields

$$\|(x^0, \lambda^0) - (x^*, \lambda^*)\|_Z < \frac{3r_{\min}}{8c_\circ \sqrt{\max(2, 1 + c^2\gamma_e^2)}} < \frac{r_{\min}}{2c_\circ}. \tag{3.13}$$

From (3.10), (3.12), and (3.13) we infer that

$$\begin{aligned} &\|(x_h^0, \lambda_h^0) - (x_h^*, \lambda_h^*)\|_{Z_h} + \|(x_h^*, \lambda_h^*) - r_h(x^*, \lambda^*)\|_{Z_h} \\ &\qquad \le c_\circ \, \|(x^0, \lambda^0) - (x^*, \lambda^*)\|_Z + 2 \, \|(x_h^*, \lambda_h^*) - r_h(x^*, \lambda^*)\|_{Z_h} \le r_{\min}. \end{aligned}$$

By Assumption 5 it follows that

$$B((x_h^*, \lambda_h^*); \|(x_h^0, \lambda_h^0) - (x_h^*, \lambda_h^*)\|_{Z_h}) \subseteq U_h.$$

To apply Theorem 2.8 we set

$$\bar{r}^* = \min\left(\frac{3r_{\min}}{4\sqrt{\max(2, 1 + c^2\gamma_e^2)}}, \frac{4}{\gamma_M \eta^* \max(2, 1 + c^2\gamma_e^2)} \right)$$

(compare with (2.5)). From (3.10) and (3.12) we conclude that

$$\|(x_h^*, \lambda_h^*) - r_h(x^*, \lambda^*)\|_{Z_h} \leq \frac{r_{\min}}{4}$$

for every $h \in (0, \bar{h}]$. Moreover, it follows that $\bar{r}^* < \frac{3}{4} r_{\min}$. Consequently, we have

$$B((x_h^*, \lambda_h^*); \bar{r}^*) \subset U_h.$$

Then Theorem 2.8 yields

$$\lim_{n\to\infty} \|(x_h^n, \lambda_h^n) - (x_h^*, \lambda_h^*)\|_{Z_h} = 0$$

and

$$\begin{aligned} &\|(x_h^{n+1}, \lambda_h^{n+1}) - (x_h^*, \lambda_h^*)\|_{Z_h} \\ &\qquad \leq \frac{1}{\gamma_M \eta^*} \max(2, 1 + c^2\gamma_e^2) \, \|(x_h^n, \lambda_h^n) - (x_h^*, \lambda_h^*)\|_{Z_h}^2 \end{aligned} \tag{3.14}$$

for the sequence $\{(x_h^n, \lambda_h^n)\}_{n\in\mathbb{N}}$, which is generated by Algorithm 2 with initial value $r_h(x^0, \lambda^0)$, if $\|r_h(x^0, \lambda^0) - (x_h^*, \lambda_h^*)\|_{Z_h} < \bar{r}^*$; i.e.,

$$\|r_h(x^0, \lambda^0) - (x_h^*, \lambda_h^*)\|_{Z_h} < \frac{3r_{\min}}{4\sqrt{\max(2, 1 + c^2\gamma_e^2)}} \tag{3.15}$$

and

$$\|r_h(x^0, \lambda^0) - (x_h^*, \lambda_h^*)\|_{Z_h} < \frac{4}{\gamma_M \eta^* \max(2, 1 + c^2\gamma_e^2)} \, . \tag{3.16}$$

Using Remark 3.8, (3.11), (3.10), and (3.12) we derive that

$$\begin{aligned} \|r_h(x^0, \lambda^0) - (x_h^*, \lambda_h^*)\|_{Z_h} &\leq \|r_h[(x^0, \lambda^0) - (x^*, \lambda^*)]\|_{Z_h} \\ &\quad + \|r_h(x^*, \lambda^*) - (x_h^*, \lambda_h^*)\|_{Z_h} \\ &\leq c_\circ \, \|(x^0, \lambda^0) - (x^*, \lambda^*)\|_Z + \tilde{C}\rho(h) \\ &< \frac{3r_{\min}}{4\sqrt{\max(2, 1 + c^2\gamma_e^2)}} \, . \end{aligned} \tag{3.17}$$

Analogously, estimate (3.16) follows. □

Remark 3.20. We infer from the proof of Corollary 3.19 that the choice of r_1 guarantees that the sequence $\{\|(x_h^n, \lambda_h^n) - (x_h^*, \lambda_h^*)\|_{Z_h}\}_{n\in\mathbb{N}}$ is strictly decreasing.

When Algorithm 2 and Algorithm 1 are stopped after the nth step, the difference between the iterates of the finite dimensional and the infinite dimensional methods can be estimated by the discretization error of the approximation scheme used. This is the assertion of the following corollary following from Theorem 3.17 (see [ABPR86]).

COROLLARY 3.21. *Let the assumptions of Corollary* 3.19 *hold. By* $\{(x^n, \lambda^n)\}_{n\in\mathbb{N}}$ *we denote the sequence generated by Algorithm* 1. *Then there exist* $\hat{h} \in (0, \tilde{h}]$ *and a constant* $r_2 \in (0, r_1]$ *such that*

$$(x_h^n, \lambda_h^n) = r_h(x^n, \lambda^n) + O\left(\rho(h)\right), \tag{3.18}$$

$$F_h(x_h^n, \tilde{\lambda}_h^n) = \hat{r}_h F(x^n, \tilde{\lambda}^n) + O\left(\rho(h)\right), \tag{3.19}$$

$$(x_h^n - x_h^*, \lambda_h^n - \lambda_h^*) = r_h(x^n - x^*, \lambda^n - \lambda^*) + O\left(\rho(h)\right) \tag{3.20}$$

for $n = 0, 1, \ldots$, *for starting values* $(x^0, \lambda^0) \in (V \times W) \cap B((x^*, \lambda^*); r_2)$, *for the sequence* $\{(x_h^n, \lambda_h^n)\}_{n\in\mathbb{N}}$ *generated by Algorithm* 2 *using the starting value* $r_h(x^0, \lambda^0)$ *with* $h \in (0, \hat{h}]$.

Remark 3.22. We emphasize that the results presented also hold for $c = 0$. Hence the convergence is proved for internal approximations of SQP-methods in Hilbert spaces.

4. Mesh-independence. Now we present the mesh-independence principle in its sharpest formulation; i.e., we give sufficient conditions for mesh-independence provided the mesh-size is smaller than an explicitly given threshold parameter. The proof is based on Corollary 2 in [ABPR86]. Since $c = 0$ is allowed, the mesh-independence principle also applies to internal approximations of SQP methods.

Let $\varepsilon > 0$. We define

$$\ell(\varepsilon) = \min\{\ell_0 \in \mathbb{N} |\text{ for } \ell \geq \ell_0 : \|(x^\ell, \lambda^\ell) - (x^*, \lambda^*)\|_Z < \varepsilon\},$$
$$\ell_h(\varepsilon) = \min\{\ell_0 \in \mathbb{N} |\text{ for } \ell \geq \ell_0 : \|(x_h^\ell, \lambda_h^\ell) - (x_h^*, \lambda_h^*)\|_{Z_h} < \varepsilon\}.$$

Let us point out that $\ell(\varepsilon)$ and $\ell_h(\varepsilon)$ depend on the starting values (x^0, λ^0) and $(x_h^0, \lambda_h^0) = r_h(x^0, \lambda^0)$ of the infinite and finite dimensional method.

THEOREM 4.1. *Suppose that the hypotheses of Corollary* 3.21 *hold. Let* $\varepsilon > 0$ *be chosen. Then there exist* $h^* \in (0, \hat{h}]$ *and* $r_3 \in (0, r_2]$, *so that*

$$|\ell(\varepsilon) - \ell_h(\varepsilon)| \leq 1 \tag{4.1}$$

for all $h \in (0, h^*]$ *and* $(x^0, \lambda^0) \in (V \times W) \cap B((x^*, \lambda^*); r_3)$.

Remark 4.2. The constant h^* depends of course on the given tolerance ε.

5. Numerical example for Example 2.1. We present a numerical example which illustrates that mesh-independence can be observed numerically choosing the control interval $\Omega_\circ = (0, 0.1)$, the viscosity parameter $\nu = \frac{1}{25}$, the regularity parameter $\alpha = 1$, the right-hand side $f(x) = x$, and the desired state $z(x) = x$. The boundary values are given by $y_l = 0$ and $y_r = 1$. Then the optimal solution of (P) is $(y^*, u^*) = (z, 0)$. It follows from (2.2) that $(\lambda^*, \mu^*, \xi^*) = (0, 0, 0)$ hold. Note that $(y_h^*, u_h^*) = (z, 0)$ is also the optimal solution of the finite dimensional problem (P_h).

Assumptions 1–5 are satisfied; see sections 2 and 3. For the proof of the uniform Babuška–Brezzi condition we refer the reader to [Vol99]. Now we are going to verify the uniform second-order sufficient optimality condition (Assumption 7).

LEMMA 5.1. *Let* $(v_h, q_h) \in \ker e_h'(z, 0)$. *Then we have*

$$\|v_h\|_{H^1}^2 \leq c_{\ker}\ \|q_h\|_{L^2(\Omega_\circ)}^2$$

for a constant $c_{\ker} > 0$.

Proof. Using $z(x) = x$ the property $(v_h, q_h) \in \ker e_h'(z, 0)$ implies that $v_h \in H_0^1(\Omega)$ and

$$\begin{aligned}
0 &= \langle e_h'(z, 0)(v_h, q_h), (v_h, 0, 0)\rangle_{Y_h} = \langle e'(z, 0)(v_h, q_h), (v_h, 0, 0)\rangle_Y \\
&= \nu\ \|v_h'\|_{L^2}^2 + \int_0^1 (zv_h)' v_h\, dx - \int_{\Omega_\circ} q_h v_h\, dx \\
&= \nu\ \|v_h'\|_{L^2}^2 + \|v_h\|_{L^2}^2 + \frac{1}{2}\ xv_h^2\Big|_{x=0}^{x=1} - \frac{1}{2}\ \|v_h\|_{L^2}^2 - \int_{\Omega_\circ} q_h v_h\, dx \\
&\geq \min\left(\nu, \frac{1}{2}\right) \|v_h\|_{H^1}^2 - \|q_h\|_{L^2(\Omega_\circ)}\ \|v_h\|_{H^1}.
\end{aligned}$$

Setting $c_{\ker} = [\min(\nu, \frac{1}{2})]^{-2}$ the claim follows. ∎

TABLE 5.1
$\ell_h(\varepsilon)$ for different tolerance ε and mesh-size h.

m	200	250	300	350	400	450	500	1000	1800
$\varepsilon = 10^{0}$	13	14	13	13	13	13	13	13	13
$\varepsilon = 10^{-2}$	14	15	14	14	14	14	14	14	14
$\varepsilon = 10^{-4}$	15	15	15	15	15	15	15	15	15
$\varepsilon = 10^{-6}$	15	16	15	15	15	15	15	15	15
$\varepsilon = 10^{-8}$	16	16	16	16	16	15	15	15	15
$\varepsilon = 10^{-10}$	16	16	16	16	16	16	16	16	16

PROPOSITION 5.2. *The uniform second-order sufficient optimality condition is satisfied.*

Proof. We have $p_h r_h(y^*, u^*, \lambda^*, \mu^*, \xi^*) = (z, 0, 0, 0, 0)$. Hence,

$$\begin{aligned}\langle L_h''(r_h(y^*, u^*, \lambda^*, \mu^*, \xi^*))(v_h, q_h), (v_h, q_h)\rangle_{X_h} &= \langle L''(z, 0, 0, 0, 0)(v_h, q_h), (v_h, q_h)\rangle_X \\ &= \|v_h\|_{L^2}^2 + \|q_h\|_{L^2(\Omega_\circ)}^2 \\ &\geq \min\left(\frac{1}{2}, \frac{1}{2c_{\ker}}\right) \|(v_h, q_h)\|_{X_h}\end{aligned}$$

by Lemma 5.1. Thus, the operator $L_h''(r_h(y^*, u^*, \lambda^*, \mu^*, \xi^*))$ is coercive on the kernel of $e_h'(r_h(y^*, u^*))$. □

PROPOSITION 5.3. *The optimal solution of* (OS) *satisfies* $(y^*, u^*, \lambda^*) \in H^2(\Omega) \times H^2(\Omega_\circ) \times H^2(\Omega)$. *If* $(y^0, u^0, \lambda^0) \in H^2(\Omega) \times H^2(\Omega_\circ) \times H^2(\Omega)$ *holds, then we have* $y^n \in H^2(\Omega)$, $u^n \in H^2(\Omega_\circ)$, *and* $\lambda^n \in H^2(\Omega)$. *Furthermore, Assumption* 8 *is satisfied.*

Proof. Since the optimal solution solves (OS), the regularity result for (y^*, u^*, λ^*) follows. The second claim is proved by induction. Let $(y^n, u^n, \lambda^n) \in H^2(\Omega) \times H^2(\Omega_\circ) \times H^2(\Omega)$ for $n \geq 0$. Then we have $e(y^n, u^n) \in H^2(\Omega) \times \mathbb{R}^2$, which implies that $\tilde{\lambda}^n \in H^2(\Omega)$. The saddle-point problem (2.4) implies that

$$\begin{aligned} -\nu(\lambda^{n+1})'' &= -y^{n+1} + (\tilde{\lambda}^n)' y^{n+1} + y^n(\lambda^{n+1})' + x - y^n(\tilde{\lambda}^n)' && \text{in } \Omega, \\ u^{n+1} &= \lambda^{n+1} && \text{in } \Omega_\circ, \\ -\nu(y^{n+1})'' &= -(y^n y^{n+1})' + \overline{u^{n+1}} + x + y^n(y^n)' && \text{in } \Omega. \end{aligned} \tag{5.1}$$

The first and third equations imply that $y^{n+1}, \lambda^{n+1} \in H^2(\Omega)$ hold. Using the second equation of (5.1) we see that $u^{n+1} \in H^2(\Omega_\circ)$. As we use piecewise linear finite elements, we obtain

$$\|(y, u, \lambda, \mu, \xi) - r_h(y, u, \lambda, \mu, \xi)\|_Z \leq \check{C} h \, \|(y'', u', \lambda'')\|_{L^2 \times L^2(\Omega_\circ) \times L^2}$$

for all $(y, u, \lambda, \mu, \xi) \in H^2(\Omega) \times H^2(\Omega_\circ) \times H^2(\Omega) \times \mathbb{R}^2$, where the constant $\check{C} > 0$ does not depend on (y, u, λ, μ) and h. If the starting value y^0 is close to y^*, then $\|y^n\|_{H^1} \leq C^* + \|y^*\|_{H^1}$ for all n. Thus, we infer from (5.1) that $\|(y^n)''\|_{L^2}$ is bounded for all n. Analogously, we derive that $\|(\lambda^n)''\|_{L^2}$ is bounded for all n. Since $u^n = \lambda^n$ in $\Omega_\circ$, the sequence $\|(u^n)''\|_{L^2(\Omega_\circ)}$ is also bounded for all n. We conclude that Assumption 8 holds. □

Remark 5.4. Note that the previous proposition also holds for arbitrary $f \in L^2(\Omega)$.

The program was written in MATLAB version 5.1, and we used an IBM RS/6000 590-workstation. We chose $c = 1$, $y^0(x) = -x^2$, $u^0(x) = 0$, $\lambda^0(x) = -1.5x$, $\mu^0 = 0$, and $\xi^0 = 0$. For $c = 0$ divergence was observed numerically. Table 5.1 yields that

TABLE 6.1
$\ell_h(\varepsilon)$ for different tolerance ε and mesh-size h.

$1/h$	30	40	60	80	100	160	200	250
$\varepsilon = 10^{0}$	3	3	3	3	3	4	4	4
$\varepsilon = 10^{-2}$	5	5	6	6	6	6	5	5
$\varepsilon = 10^{-4}$	6	6	7	7	7	7	6	6
$\varepsilon = 10^{-6}$	6	7	7	7	7	7	7	6
$\varepsilon = 10^{-8}$	7	7	8	8	8	7	7	7
$\varepsilon = 10^{-10}$	7	7	8	8	8	8	7	7

there is at most a difference of one between the number of steps required to converge within a given tolerance $\varepsilon > 0$. We emphasize that the mesh-independence principle is satisfied for $h \leq 0.005$.

6. Numerical example for Example 2.2. Let $l \in \mathbb{N}$ be given. We introduce a discretization of the time interval $[0, T]$ by setting $k = \frac{T}{l}$ and $t_j = jk$ for $j = 0, \ldots, l$. We used the implicit Euler method for the time integration and piecewise linear finite elements introduced in Remark 3.1.

Let $T = 1$, $\Omega = (0, 1)$, $\alpha = 0.1$, and $\nu = 0.01$. Further, we put

$$z(t, x) = \left(\left(x - \frac{1}{2}\right)^2 - \frac{1}{4}\right)\left(\left(t - \frac{1}{2}\right)^2 - \frac{1}{4}\right)$$

and $f = z_t - \nu z_{xx} + z z_x$. Clearly, $(y^*, u^*) = (z, 0)$ is the optimal solution of the optimal control problem. For the augmented Lagrangian-SQP method we chose $c = 50$. In the computations we used a uniform mesh h for the time as well as for the spatial discretization. To solve the linear system (3.6) we applied the MATLAB function `gmres` with preconditioning. The starting values of Algorithm 2 were chosen to be $y_h^0 = -z$, $u_h^0 = \lambda_h^0 = 0$. Note that the iteration y_h^n has to "pass through" 0. In Table 6.1 we specify the smallest iteration number $\ell = \ell_h(\varepsilon)$ such that

$$\|(y_h^\ell, u_h^\ell, \lambda_h^\ell) - (y_h^*, u_h^*, \lambda_h^*)\|_{Z_h} < \varepsilon.$$

We proved in Theorem 4.1 that a strong mesh-independence principle is valid for $\ell_h(\varepsilon)$ and the corresponding iteration count $\ell(\varepsilon)$ of the infinite dimensional method. But the approximation scheme used does not generally satisfy the assumptions of Theorem 4.1. However, Table 6.1 illustrates that the asymptotic mesh-independence can be observed numerically for $h \leq \frac{1}{30}$.

REFERENCES

[AB87] E. L. ALLGOWER AND K. BÖHMER, *Application of the mesh-independence principle to mesh refinement strategies*, SIAM J. Numer. Anal., 24 (1987), pp. 1335–1351.

[ABM81] E. L. ALLGOWER, K. BÖHMER, AND S. F. MCCORMICK, *Discrete correction methods for operator equations*, in Numerical Solution of Nonlinear Equations: Simplicial and Classical Methods, Bremen 1980, Lecture Notes in Math. 878, Springer-Verlag, Berlin, 1981, pp. 30–97.

[ABPR86] E. L. ALLGOWER, K. BÖHMER, F. A. POTRA, AND W. C. RHEINBOLDT, *A mesh-independence principle for operator equations and their discretizations*, SIAM J. Numer. Anal., 23 (1986), pp. 160–169.

[Alt95] W. ALT, *Discretization and mesh-independence of Newton's method for generalized equations*, in Mathematical Programming with Data Pertubations, Lecture Notes in Pure and Appl. Math. 195, A. V. Fiacco, ed., Dekker, New York, 1998, pp. 1–30.

[AM78] E. L. Allgower and S. F. McCormick, *Newton's method with mesh refinements for numerical solution of nonlinear two-point boundary value problems*, Numer. Math., 29 (1978), pp. 237–260.

[AMP79] E. L. Allgower, S. F. McCormick, and D. V. Pryor, *A general mesh independence principle for Newton's method applied to second order boundary value problems*, Computing, 23 (1979), pp. 233–246.

[Arg90] I. K. Argyros, *A mesh-independence principle for nonlinear operator equations and their discretizations under mild differentiability conditions*, Computing, 45 (1990), pp. 265–268.

[Arg92] I. K. Argyros, *On a mesh-independence principle for operator equations and the secant method*, Acta Math. Hungar., 60 (1992), pp. 7–19.

[Aub72] J.-P. Aubin, *Approximation of Elliptic Boundary-Value Problems*, Pure Appl. Math. 26, Wiley-Interscience, New York, 1972.

[Axe93] O. Axelsson, *On mesh-independence and Newton-type methods*, Appl. Math., 38 (1993), pp. 249–265.

[BF91] F. Brezzi and M. Fortin, *Mixed and Hybrid Finite Element Methods*, Springer Ser. Comput. Math. 15, Springer-Verlag, New York, 1991.

[Bre87] H. Brezis, *Analyse fonctionnelle. Théorie et applications*, Masson, Paris, 1987.

[Bur40] J. M. Burgers, *Application of a model system to illustrate some points of the statistical theory of free turbulence*, Nederl. Akad. Wefensh. Proc., 43 (1940), pp. 2–12.

[CH91] Z. Chen and K.-H. Hoffmann, *Numerical solutions of the optimal control problem governed by a phase field model*, in Optimal Control of Partial Differential Equations, Internat. Ser. Numer. Math. 100, Birkhäuser, Basel, 1991, pp. 79–97.

[DL88] R. Dautray and J.-L. Lions, *Mathematical Analysis and Numerical Methods for Science and Technology*. Vol. 2: *Functional and Variational Methods*, Springer-Verlag, New York, 1988.

[DP92] P. Deuflhard and F. A. Potra, *Asymptotic mesh independence of Newton-Galerkin methods via a refined Mysovskii theorem*, SIAM J. Numer. Anal., 29 (1992), pp. 1395–1412.

[FG83] M. Fortin and R. Glowinski, *Augmented Lagrangian Methods: Application to Numerical Solutions of Boundary Value Problems*, North–Holland, Amsterdam, 1983.

[GHS91] M. D. Gunzburger, L. Hou, and T. P. Svobodny, *Finite element approximations of an optimal control problem associated with the scalar Ginzburg-Landau equation*, Computers Math. Appl., 21 (1991), pp. 123–131.

[GR92] C. Grossmann and H.-G. Roos, *Numerik partieller Differentialgleichungen*, 2nd ed., Teubner Studienbüch. Math., Teubner, Stuttgart, 1992.

[Gun95] M. D. Gunzburger, *Flow Control* I, MA Vol. Math. Appl. 68, Springer-Verlag, New York, 1995.

[Hei93] M. Heinkenschloss, *Mesh independence for nonlinear least squares problems with norm constraints*, SIAM J. Optim., 3 (1993), pp. 81–117.

[HLS91] M. Heinkenschloss, M. Laumen, and E. W. Sachs, *Gauss-Newton methods with grid refinement*, in Estimation and Control of Distributed Parameter Systems, Internat. Ser. Numer. Math. 100, Birkhäuser–Verlag, Basel, 1991, pp. 161–174.

[HS94] M. Heinkenschloss and E. W. Sachs, *Numerical solution of a constrained control problem for a phase field model*, in Control and Estimation of Distributed Parameter Systems: Nonlinear Phenomena, Birkhäuser–Verlag, Basel, 1994, pp. 171–187.

[IK90a] K. Ito and K. Kunisch, *The augmented Lagrangian method for equality and inequality constraints in Hilbert spaces*, Math. Programming, 46 (1990), pp. 341–360.

[IK90b] K. Ito and K. Kunisch, *The augmented Lagrangian method for parameter estimation in elliptic systems*, SIAM J. Control Optim., 28 (1990), pp. 113–136.

[IK91] K. Ito and S. Kang, *Estimation of a temporally and spatially varying diffusion coefficient in a parabolic system by an augmented Lagrangian technique*, Numer. Math., 59 (1991), pp. 473–509.

[IK96a] K. Ito and K. Kunisch, *Augmented Lagrangian-SQP methods for nonlinear optimal control problems of tracking type*, SIAM J. Control Optim., 34 (1996), pp. 874–891.

[IK96b] K. Ito and K. Kunisch, *Augmented Lagrangian-SQP methods in Hilbert spaces and application to control in the coefficients problems*, SIAM J. Optim., 6 (1996), pp. 96–125.

[IKK91] K. Ito, M. Kroller, and K. Kunisch, *A numerical study of an augmented Lagrangian method for the estimation of parameters in elliptic systems*, SIAM J. Sci. Stat. Comput., 12 (1991), pp. 884–910.

[KS87] C. T. Kelley and E. W. Sachs, *Quasi-Newton methods and unconstrained optimal control problems*, SIAM J. Control Optim., 25 (1987), pp. 1503–1516.

[KS90] C. T. Kelley and E. W. Sachs, *Approximate quasi-Newton methods*, Math. Programming, 48 (1990), pp. 41–70.

[KS91] C. T. Kelley and E. W. Sachs, *Mesh independence of Newton-like methods for infinite dimensional problems*, J. Integral Equations Appl., 3 (1991), pp. 549–573.

[KS92] C. T. Kelley and E. W. Sachs, *Mesh independence of the gradient projection method for optimal control problems*, SIAM J. Control Optim., 30 (1992), pp. 477–493.

[KP91] K. Kunisch and G. Peichl, *Estimation of a temporally and spatially varying coefficient in a parabolic system by an augemented Lagrangian technique*, Numer. Math., 59 (1991), pp. 473–509.

[KV97] K. Kunisch and S. Volkwein, *Augmented Lagrangian-SQP techniques and their approximations*, in Optimization Methods in Partial Differential Equations, Contemp. Math. 209, S. Cox and I. Lasiecka, eds., AMS, Providence, RI, 1997, pp. 147–159.

[Lig56] M. J. Lighthill, *Viscosity effects in sound waves of finite amplitude*, Surveys in Mechanics, Cambridge University Press, Cambridge, UK, 1956, pp. 250–351.

[Lio85] J.-L. Lions, *Control of Distributed Singular Systems*, Gauthier-Villars, Paris, 1985.

[Lue69] D. G. Luenberger, *Optimization by Vector Space Methods*, John Wiley & Sons, Inc., New York, 1969.

[McC78] S. F. McCormick, *A revised mesh refinement strategy for Newton's method applied to nonlinear two-point boundary value problems*, in Numerical Treatment of Differential Equations in Applications, Lecture Notes in Math. 679, Springer-Verlag, New York, 1978, pp. 15–23.

[MZ79] H. Maurer and J. Zowe, *First and second order necessary and sufficient optimality conditions for infinite-dimensional programming problems*, Math. Programming, 16 (1979), pp. 98–110.

[OR70] J. M. Ortega and W. C. Rheinboldt, *Iterative Solutions of Nonlinear Equations in Several Variables*, Academic Press, New York, 1970.

[PR84] F. A. Potra and W. C. Rheinboldt, *On the Mesh-Independence Principle for Discretizations of Nonlinear Operator Equations*, Technical Report ICMA-84-74, Institute for Comp. Math. and Appl., Univ. Pittsburgh, Pittsburgh, PA, 1984.

[PT73] V. T. Polyak and N. Y. Tret'yakov, *The method of penalty estimates for conditional extremum problems*, Zh. Vychisl. Mat. Mat. Fiz., 13 (1973), pp. 34–46.

[Rhe80] W. Rheinboldt, *On a theory of mesh-refinement processes*, SIAM J. Numer. Anal., 17 (1980), pp. 766–778.

[Tin75] M. Tinkham, *Introduction to Superconductivity*, McGraw–Hill, New York, 1975.

[Vol97] S. Volkwein, *Mesh-Independence of an Augmented Lagrangian-SQP Method in Hilbert Spaces and Control Problems for the Burgers Equation*, Ph.D. thesis, Department of Mathematics, Technical University of Berlin, 1997.

[Vol99] S. Volkwein, *Augmented Lagrangian-SQP Techniques and Optimal Control Problems for the Stationary Burgers Equation*, Comput. Optim. Appl., to appear.

[Wou79] A. Wouk, *A Course of Applied Functional Analysis*, Wiley-Interscience, New York, 1979.

SIAM J. CONTROL OPTIM.
Vol. 38, No. 3, pp. 786–802

EXISTENCE OF RIGHT AND LEFT REPRESENTATIONS OF THE GRAPH FOR LINEAR PERIODICALLY TIME-VARYING SYSTEMS*

MICHAEL CANTONI† AND KEITH GLOVER†

Abstract. The graph representation of a system (the set of all input-output pairs) has gained considerable attention in the control literature in view of its usefulness for the analysis of feedback systems. In this paper it is shown that the graph of any *stabilizable*, linear, periodically time-varying (LPTV), continuous-time system can be expressed as the range and kernel of bounded, causal, LPTV systems that are, respectively, left and right invertible by bounded, causal, LPTV systems. These so-called strong-right and strong-left representations are closely related to the perhaps more common notion of coprime factor representations. As an example of their usefulness, a neat characterization of closed-loop stability is obtained in terms of strong-right and strong-left representations of the plant and controller graphs. This in turn leads to a Youla-style parametrization of stabilizing controllers. All of the results obtained accommodate possibly infinite-dimensional input and output spaces and apply, as a special case, to sampled-data control-systems. Furthermore, they are particularly useful for robustness analysis in terms of the gap metric.

Key words. graphs, periodic time-variation, system representations, stabilizing controllers

AMS subject classifications. 93C05, 93C25, 9C350, 93C60, 93D15, 93D25

PII. S0363012998346608

1. Introduction.

Fractional system representations are useful for the analysis and design of feedback systems [7, 24, 6]. For example, it is well known that the fundamental objective of closed-loop stability can be neatly characterized in terms of the coprime factors (taken over a set of stable systems) of a closed-loop's component systems [23]. Furthermore, the coprime factorization framework is closely related to robustness analysis in the graph topology [22, 11, 19, 25], and it also lies at the heart of the H^∞-loopshaping design paradigm for linear, time-invariant systems [14].

The existence of coprime factors has been linked to stabilizability for several classes of systems. For linear, time-invariant systems that evolve in either continuous or discrete time, this result is often expressed in terms of the corresponding frequency-domain transfer-function and a factorization of this over the Hardy space H^∞ (see [20, 23] and references therein). Existence results are also available for linear, time-varying systems described by finite-dimensional, stabilizable, and detectable state-space realizations [18, 16].

In [5], an input-output perspective is taken to study linear, time-varying, discrete-time systems. More specifically, it is shown in [5] that a causal, linear, time-varying, discrete-time system is *stabilizable* if and only if its graph (the set of all input-output pairs) can be expressed as the range (respectively, kernel) of a bounded, causal, linear system that is left (respectively, right) invertible by another bounded, causal, linear system. This so-called strong-right (respectively, strong-left) representation of the graph is notionally equivalent to the representation of the system as the ratio of right-coprime (respectively, left-coprime) factors, because a strong-right (respectively,

*Received by the editors October 29, 1998; accepted for publication (in revised form) March 22, 1999; published electronically March 8, 2000. This research was supported in part by the Gledden Studentship of the University of Western Australia and the Engineering and Physical Sciences Research Council (EPSRC), UK.

http://www.siam.org/journals/sicon/38-3/34660.html

†Department of Engineering, University of Cambridge, Cambridge CB2 1PZ, UK (mwc@eng.cam.ac.uk, kg@eng.cam.ac.uk).

strong-left) representation of the graph can be constructed straightforwardly from any right-coprime (respectively, left-coprime) factorization. Unfortunately, the results as developed in [5] do not extend directly to accommodate continuous-time systems.

In this paper, the input-output perspective described above is employed to study linear, periodically time-varying (LPTV), continuous-time systems. Such systems arise in sampled-data control [3] and the analysis of nonlinear systems along specified periodic trajectories, for example. The main results presented in this paper imply that the graph of a stabilizable, LPTV, continuous-time system can be expressed as the range (respectively, kernel) of a bounded, causal, LPTV system that is left (respectively, right) invertible by a bounded, causal, LPTV system. These so-called strong-right and strong-left representations of the graph facilitate a neat characterization of closed-loop stability, a parametrization of stabilizing controllers and, correspondingly, gap-metric robustness analysis of LPTV feedback-systems (see [4]). All of the results obtained accommodate possibly infinite-dimensional input and output spaces, and as such, apply to systems that are spatially distributed.

The paper has the following structure. Preliminary results from functional analysis and general systems theory are presented in the next section. Underpinning many of the results developed in this paper is the equivalence of LPTV, continuous-time systems to linear, shift-invariant (LSI), discrete-time systems. This equivalence implies that the graph of an LPTV system is isomorphic to a shift-invariant subspace of the Hardy space H^2. As such, section 3 is devoted to the characterization of shift-invariant subspaces. There, it is shown that if a shift-invariant subspace is shift-invariantly coordinatizable, then it can be expressed as the range and kernel of multiplication operators with symbols in H^∞. It is also shown that these range and kernel representations are, respectively, left and right invertible by multiplication operators with symbols in H^∞. Once certain causality issues have been addressed (as described in section 4), this extension of the Beurling–Lax–Halmos theorem is used to prove the existence of strong-right and strong-left representations of the graph for stabilizable, LPTV systems. Finally, in section 5 a neat characterization of closed-loop stability and a Youla-style parametrization of stabilizing, LPTV controllers for LPTV plants are given in terms of strong-right and strong-left representations of the plant and controller graphs.

2. Preliminaries. In this section, general notation and definitions used throughout the paper are introduced. Let $\mathbb{R}$, $\mathbb{Z}$, $\mathbb{R}^+$, $\mathbb{Z}^+$, $\mathbb{D}$, and $\mathbb{T}$ denote the reals, integers, nonnegative reals, nonnegative integers, open unit-disc, and unit-circle, respectively. For convenience, given a real number $h > 0$, the interval $[0, h) \subset \mathbb{R}$ is denoted by $\mathbb{H}$. In any Hilbert space $\mathcal{H}$, the inner-product is denoted by $\langle \cdot, \cdot \rangle_{\mathcal{H}}$ and the norm by $\| \cdot \|_{\mathcal{H}}$. For a subspace $\mathcal{U} \subset \mathcal{H}$, the orthogonal complement of $\mathcal{U}$ in $\mathcal{H}$ is denoted by $\mathcal{H} \ominus \mathcal{U}$ (or $\mathcal{U}^\perp$ when the ambient space $\mathcal{H}$ is clear from context) and the orthogonal projection onto $\mathcal{U}$ is denoted by $\mathbf{\Pi}_{\mathcal{U}}$.

Consider two closed subspaces $\mathcal{G}$ and $\mathcal{F}$ of a Hilbert space $\mathcal{H}$. If $\mathcal{G} \cap \mathcal{F} = \{0\}$ and $\mathcal{G} + \mathcal{F} = \mathcal{H}$, then $\mathcal{G}$ and $\mathcal{F}$ are said to induce a *coordinatization* of $\mathcal{H}$. In this case, any $h \in \mathcal{H}$ can be uniquely decomposed as the sum $h = g + f$, where $g \in \mathcal{G}$ and $f \in \mathcal{F}$ (see [8], for example). The bounded, linear operator $\mathbf{\Pi}_{\mathcal{G}\|\mathcal{F}} : h \mapsto g$ is called the parallel projection onto $\mathcal{G}$ along $\mathcal{F}$. Similarly, $\mathbf{\Pi}_{\mathcal{F}\|\mathcal{G}} : h \mapsto f$ is called the *parallel projection* onto $\mathcal{F}$ along $\mathcal{G}$.

Let $\mathcal{U}$ and $\mathcal{Y}$ be Hilbert spaces and consider a linear operator $P : \mathcal{D}_P \subset \mathcal{U} \to \mathcal{Y}$, where $\mathcal{D}_P := \{u \in \mathcal{U} : Pu \in \mathcal{Y}\}$ is called the domain of P. The range of P is defined to be $\mathcal{R}_P := \{Pu : u \in \mathcal{D}_P\}$, and $\mathcal{K}_P := \{u \in \mathcal{D}_P : Pu = 0\}$ is called the kernel of P.

The *graph* of P is defined to be the set of all input-output pairs

$$\mathcal{G}_P := \begin{bmatrix} I \\ P \end{bmatrix} \mathcal{D}_P \subset \mathcal{U} \oplus \mathcal{Y},$$

and for notational convenience the *inverse graph* is denoted by $\mathcal{G}_P^\sharp := [\begin{smallmatrix} 0 & I \\ I & 0 \end{smallmatrix}]\mathcal{G}_P$. Note that a (linear) subspace $\mathcal{G} \subset \mathcal{U} \oplus \mathcal{Y}$ corresponds to the graph of a linear operator if and only if $\left[\begin{smallmatrix} 0 \\ y \end{smallmatrix}\right] \in \mathcal{G}$ implies that $y = 0$. The symbol $\mathcal{B}_{\mathcal{U},\mathcal{Y}}$ is used to denote the Banach space of all bounded, linear operators $P : \mathcal{U} \to \mathcal{Y}$, that is, all such operators with $\mathcal{D}_P = \mathcal{U}$ and finite induced-norm

$$\|P\| := \sup_{\substack{u \in \mathcal{D}_P \\ u \neq 0}} \frac{\|Pu\|_{\mathcal{Y}}}{\|u\|_{\mathcal{U}}} < \infty.$$

Given an operator $P \in \mathcal{B}_{\mathcal{U},\mathcal{Y}}$, there exists a unique operator $P^* \in \mathcal{B}_{\mathcal{Y},\mathcal{U}}$ such that for all $u \in \mathcal{U}$ and $y \in \mathcal{Y}$,

$$\langle y, Pu \rangle_{\mathcal{Y}} = \langle P^* y, u \rangle_{\mathcal{U}}.$$

The operator P^* is called the (Hilbert space) adjoint. Note that for any $P \in \mathcal{B}_{\mathcal{U},\mathcal{Y}}$, $P = (P^*)^*$ and $\mathcal{R}_P$ is orthogonal to $\mathcal{K}_{P^*}$. An operator $P \in \mathcal{B}_{\mathcal{U},\mathcal{Y}}$ is called an isometry if $\langle Pu, Pu \rangle_{\mathcal{Y}} = \langle u, u \rangle_{\mathcal{U}}$ for all $u \in \mathcal{U}$ (or equivalently, $P^*P = I$).

2.1. Signals and systems. In this paper, a system is simply considered to be an operator mapping between signal-spaces. Primarily, attention is focused on systems mapping between continuous-time spaces of signals with finite-energy. Mathematically such signals can be thought of as functions in $\mathrm{L}^2_{\mathbb{R}^+}(\mathcal{H})$, the Hilbert space of $\mathcal{H}$-valued, (Lebesgue) square-integrable functions $f : \mathbb{R}^+ \to \mathcal{H}$. By virtue of the class of systems considered in the sequel and the analysis technique employed to study these systems, the following signal-spaces also play an important role: the discrete-time signal-space $\ell^2_{\mathbb{Z}^+}(\mathcal{H})$ of square-summable sequences $\mathsf{f} : \mathbb{Z}^+ \to \mathcal{H}$; and the frequency-domain signal-space $\mathrm{H}^2_{\mathbb{D}}(\mathcal{H})$ of functions $\varphi : \mathbb{D} \to \mathcal{H}$ that are analytic in $\mathbb{D}$ and satisfy

$$\int_0^{2\pi} \left\langle \varphi(re^{j\omega}), \varphi(re^{j\omega}) \right\rangle_{\mathcal{H}} d\omega < M$$

for some $M < \infty$ and all $0 \leq r < 1$. Note that $\mathrm{H}^2_{\mathbb{D}}(\mathcal{H})$ is isomorphic to $\ell^2_{\mathbb{Z}^+}(\mathcal{H})$ via the $\mathbf{Z}$-transform isomorphism [21, pp. 184–185], defined for all $\mathsf{f} \in \ell^2_{\mathbb{Z}^+}(\mathcal{H})$ by

$$(\mathbf{Z}\mathsf{f})(\lambda) := \sum_{i=0}^{\infty} \mathsf{f}_k \lambda^k, \quad \lambda \in \mathbb{D}.$$

Let $\mathcal{U}$ and $\mathcal{Y}$ be Hilbert spaces. For a linear, continuous-time system $P : \mathcal{D}_P \subset \mathrm{L}^2_{\mathbb{R}^+}(\mathcal{U}) \to \mathrm{L}^2_{\mathbb{R}^+}(\mathcal{Y})$, the standard notion of causality can be expressed as follows: P is *causal* if for all $\tau \in \mathbb{R}^+$, $\mathbf{T}_\tau \mathcal{G}_P$ corresponds to the graph of a linear operator, where $\mathbf{T}_\tau$ is the projection that truncates a signal to zero after time τ. Given a real number $h > 0$, such a system P is called *periodically time-varying* (with period h) if $\mathbf{U}_{kh}\mathcal{G}_P \subset \mathcal{G}_P$ for all $k \in \mathbb{Z}^+$, where $\mathbf{U}_\tau$ denotes the unilateral (forward) shift on $\mathrm{L}^2_{\mathbb{R}^+}(\cdot)$.[1]

[1] According to this definition a linear, time-invariant system is also LPTV, since such a system must, by definition, satisfy $\mathbf{U}_\tau \mathcal{G}_P \subset \mathcal{G}_P$ for all $\tau \in \mathbb{R}^+$.

The following technical definitions are necessary to facilitate a precise definition of the class of systems considered in that which follows. A causal, linear, continuous-time system $P : \mathcal{D}_P \subset \mathrm{L}^2_{\mathbb{R}^+}(\mathcal{U}) \to \mathrm{L}^2_{\mathbb{R}^+}(\mathcal{U})$ is said to be *causally extendible* if $\mathbf{T}_\tau \mathcal{D}_P = \mathbf{T}_\tau \mathrm{L}^2_{\mathbb{R}^+}(\mathcal{U})$ for all $\tau < \infty$. In this way, the input to P can be chosen arbitrarily over any finite interval $[0, \tau)$ and then continued into $\mathcal{D}_P$. Since P is causal, the corresponding output over $[0, \tau)$ is defined *uniquely* by the input up to time τ. Accordingly, when P is causally extendible there is a one-to-one correspondence between P and a system $P_e : \mathrm{L}^{2,\mathrm{e}}_{\mathbb{R}^+}(\mathcal{U}) \to \mathrm{L}^{2,\mathrm{e}}_{\mathbb{R}^+}(\mathcal{Y})$ such that for all $u_e \in \mathrm{L}^{2,\mathrm{e}}_{\mathbb{R}^+}(\mathcal{U})$ and $\tau < \infty$, $\mathbf{T}_\tau P_e u_e = \mathbf{T}_\tau P_e \mathbf{T}_\tau u_e := \mathbf{T}_\tau P u$ for any $u \in \mathcal{D}_P$ that satisfies $\mathbf{T}_\tau u_e = \mathbf{T}_\tau u$, where $\mathrm{L}^{2,\mathrm{e}}_{\mathbb{R}^+}(\mathcal{H}) := \{f : \mathbb{R}^+ \to \mathcal{H} \,:\, \mathbf{T}_\tau f \in \mathrm{L}^2_{\mathbb{R}^+}(\mathcal{H}) \;\forall\, \tau < \infty\}$ denotes the extended space associated with $\mathrm{L}^2_{\mathbb{R}^+}(\mathcal{H})$ for any Hilbert space $\mathcal{H}$. If P and P_e are related in this way, P_e is called the (causal) extension of P and it is said that P is induced by P_e. When P is causally extendible, its causal extension P_e is said to be *locally Lipschitz-continuous* if for all $\tau \in \mathbb{R}^+$,

$$\sup_{\substack{u_1, u_2 \in \mathrm{L}^{2,\mathrm{e}}_{\mathbb{R}^+}(\mathcal{U}) \\ \mathbf{T}_\tau u_1 \neq \mathbf{T}_\tau u_2}} \frac{\|\mathbf{T}_\tau (P u_1 - P u_2)\|_{\mathrm{L}^2_{\mathbb{R}^+}(\mathcal{Y})}}{\|\mathbf{T}_\tau (u_1 - u_2)\|_{\mathrm{L}^2_{\mathbb{R}^+}(\mathcal{U})}} < \infty.$$

DEFINITION 2.1. *Given two Hilbert spaces $\mathcal{U}$ and $\mathcal{Y}$, let $\mathscr{P}_{\mathcal{U},\mathcal{Y}}$ denote the set of causal, LPTV (with period h), continuous-time systems*

$$P : \mathcal{D}_P \subset \mathrm{L}^2_{\mathbb{R}^+}(\mathcal{U}) \to \mathrm{L}^2_{\mathbb{R}^+}(\mathcal{Y}),$$

with closed graphs.[2] *Furthermore, let $\mathscr{P}^e_{\mathcal{U},\mathcal{Y}} \subset \mathscr{P}_{\mathcal{U},\mathcal{Y}}$ denote the subset of causally extendible systems with locally Lipschitz-continuous extensions.*

A continuous-time system $P : \mathcal{D}_P \subset \mathrm{L}^2_{\mathbb{R}^+}(\mathcal{U}) \to \mathrm{L}^2_{\mathbb{R}^+}(\mathcal{Y})$ is said to be stable if it is causal and $\mathcal{D}_P = \mathrm{L}^2_{\mathbb{R}^+}(\mathcal{U})$. If the system also has a closed graph, then $\|P\| < \infty$ by the closed graph theorem [2, p. 80]. Note that any stable system $P \in \mathscr{P}_{\mathcal{U},\mathcal{Y}}$ is an element of $\mathscr{P}^e_{\mathcal{U},\mathcal{Y}}$.

Importantly, each system in $\mathscr{P}_{\mathcal{U},\mathcal{Y}}$ is equivalent (via the time-lifting isomorphism defined next) to a discrete-time system that is shift invariant [1]. Let $\mathbf{W} : \mathrm{L}^2_{\mathbb{R}^+}(\mathcal{H}) \to \ell^2_{\mathbb{Z}^+}(\mathrm{L}^2_{\mathbb{H}}(\mathcal{H}))$ denote the *time-lifting isomorphism* defined for each $f \in \mathrm{L}^2_{\mathbb{R}^+}(\mathcal{H})$ by

$$\vec{\mathsf{f}}^{\circ}_k(\theta) := (\mathbf{W} f)_k = f(kh + \theta), \quad \theta \in \mathbb{H},$$

where $\mathrm{L}^2_{\mathbb{H}}(\mathcal{H}) := \mathbf{T}_h \mathrm{L}^2_{\mathbb{R}^+}(\mathcal{H})$. Then given any system $P \in \mathscr{P}_{\mathcal{U},\mathcal{Y}}$, the time-lifted equivalent, discrete-time system

$$\mathsf{P} := \mathbf{W} P \mathbf{W}^{-1} : \mathcal{D}_{\mathsf{P}} \subset \ell^2_{\mathbb{Z}^+}(\mathrm{L}^2_{\mathbb{H}}(\mathcal{U})) \to \ell^2_{\mathbb{Z}^+}(\mathrm{L}^2_{\mathbb{H}}(\mathcal{Y}))$$

is causal and LSI in the sense that its graph is a shift-invariant (linear) subspace, meaning $\mathbf{S}\mathcal{G}_{\mathsf{P}} \subset \mathcal{G}_{\mathsf{P}} = \mathbf{W}\mathcal{G}_P$, and $\mathbf{T}_k \mathcal{G}_{\mathsf{P}}$ corresponds to the graph of a linear operator for all $k \in \mathbb{Z}^+$ (implying causality), where $\mathbf{S}$ denotes the unilateral (forward) shift on $\ell^2_{\mathbb{Z}^+}(\cdot)$, and $\mathbf{T}_k$ denotes the truncation to zero after time k.[3] The key here is that $\mathbf{S}^k \mathbf{W} = \mathbf{W} \mathbf{U}_{kh}$ for all $k \in \mathbb{Z}^+$, so that a shift by h in the continuous-time signal

[2] That a linear system has a closed graph is necessary for it to be stabilizable [10]. Hence, this is assumed.

[3] Throughout, the sans serif font (e.g., P) is used to distinguish objects associated with discrete-time signals/systems from continuous-time objects, for which Roman italics are used (e.g., P).

space $\mathrm{L}^2_{\mathbb{R}^+}(\cdot)$ corresponds to a shift by a single time-step in the isomorphic, discrete-time signal-space $\ell^2_{\mathbb{Z}^+}(\mathrm{L}^2_{\mathbb{H}}(\cdot))$. Note that by representing each signal in $\ell^2_{\mathbb{Z}^+}(\mathrm{L}^2_{\mathbb{H}}(\cdot))$ as a column vector with kth entry corresponding to the value of the signal at time k, the time-lifted equivalent system $\mathsf{P}(:=\mathbf{W}P\mathbf{W}^{-1})$ has lower-triangular, block-Toeplitz structure

$$\begin{bmatrix} \mathsf{P}_{[0]} & 0 & \cdots & \cdots & \cdots \\ \mathsf{P}_{[1]} & \mathsf{P}_{[0]} & 0 & \cdots & \cdots \\ \mathsf{P}_{[2]} & \mathsf{P}_{[1]} & \mathsf{P}_{[0]} & 0 & \cdots \\ \vdots & \ddots & \ddots & \ddots & \ddots \end{bmatrix},$$

which can be uniquely identified with the sequence $\{\mathsf{P}_{[i]}\}_{i=0}^{\infty}$. This representation plays a significant role in that which follows. If the original LPTV system P is causally extendible to a locally Lipschitz-continuous system, then each $\mathsf{P}_{[i]} \in \mathcal{B}_{\mathrm{L}^2_{\mathbb{H}}(\mathcal{U}),\mathrm{L}^2_{\mathbb{H}}(\mathcal{Y})}$.

REMARK 2.2. *Given an LSI system* $\mathsf{P} : \mathcal{D}_{\mathsf{P}} \subset \ell^2_{\mathbb{Z}^+}(\mathrm{L}^2_{\mathbb{H}}(\mathcal{U})) \to \ell^2_{\mathbb{Z}^+}(\mathrm{L}^2_{\mathbb{H}}(\mathcal{Y}))$,[4] *the equivalent continuous-time system* $P(:=\mathbf{W}^{-1}\mathsf{P}\mathbf{W}$ *may not be locally causal in each* $[kh,(k+1)h)$ *interval of time. The equivalent system* P *is causal if and only if* $\mathsf{P}_{[0]}$ *(the first element of the sequence uniquely identifiable with the block-Toeplitz representation of* P*) is a causal mapping from* $\mathrm{L}^2_{\mathbb{H}}(\mathcal{U})$ *to* $\mathrm{L}^2_{\mathbb{H}}(\mathcal{Y})$*, in the sense that* $\mathbf{T}_\tau \mathcal{G}_{\mathsf{P}_{[0]}}$ *corresponds to the graph of a linear operator for all* $\tau \in \mathbb{H}$.

An LSI system $\mathsf{P} : \mathcal{D}_{\mathsf{P}} \subset \ell^2_{\mathbb{Z}^+}(\mathcal{U}) \to \ell^2_{\mathbb{Z}^+}(\mathcal{Y})$ is said to be stable if $\mathcal{D}_{\mathsf{P}} = \ell^2_{\mathbb{Z}^+}(\mathcal{U})$. If in addition $\mathcal{G}_{\mathsf{P}}$ is closed, then $\|\mathsf{P}\| < \infty$ by the closed graph theorem. Related to such operators is the Hardy space $\mathrm{H}^\infty_{\mathbb{D}}(\mathcal{B}_{\mathcal{U},\mathcal{Y}})$ of functions $\Phi : \mathbb{D} \to \mathcal{B}_{\mathcal{U},\mathcal{Y}}$ that are bounded and analytic in the open unit-disc. Given a function $\Phi \in \mathrm{H}^\infty_{\mathbb{D}}(\mathcal{B}_{\mathcal{U},\mathcal{Y}})$, boundary values can be defined almost everywhere on the unit-circle $\mathbb{T}$, and the resulting boundary values $\Phi(e^{j\omega})$ are essentially bounded. A function $\Phi \in \mathrm{H}^\infty_{\mathbb{D}}(\mathcal{B}_{\mathcal{U},\mathcal{Y}})$ is said to be inner if $\Phi(e^{j\omega})$ is an isometry almost everywhere on $\mathbb{T}$. Corresponding to each $\Phi \in \mathrm{H}^\infty_{\mathbb{D}}(\mathcal{B}_{\mathcal{U},\mathcal{Y}})$ is a multiplication operator $\mathbf{M}_\Phi : \mathrm{H}^2_{\mathbb{D}}(\mathcal{U}) \to \mathrm{H}^2_{\mathbb{D}}(\mathcal{Y})$, defined by $(\mathbf{M}_\Phi\varphi)(\lambda)(:=\Phi(\lambda)\varphi(\lambda)$ for all $\varphi \in \mathrm{H}^2_{\mathbb{D}}(\mathcal{U})$ and $\lambda \in \mathbb{D}$.

PROPOSITION 2.3 (see [9, p. 235]). *Given a stable, LSI system* $\mathsf{P} : \ell^2_{\mathbb{Z}^+}(\mathcal{U}) \to \ell^2_{\mathbb{Z}^+}(\mathcal{Y})$ *with a closed graph, there exists a function* $\hat{\mathsf{P}} \in \mathrm{H}^\infty_{\mathbb{D}}(\mathcal{B}_{\mathcal{U},\mathcal{Y}})$ *such that* $\mathsf{P} = \mathbf{Z}^{-1}\mathbf{M}_{\hat{\mathsf{P}}}\mathbf{Z}$. *Furthermore,* P *is an isometry if and only if the corresponding symbol* $\hat{\mathsf{P}}$ *is inner.*

2.2. Feedback systems. Consider the feedback configuration shown in Figure 2.1, and suppose that $P \in \mathscr{P}^e_{\mathcal{U},\mathcal{Y}}$ and $C \in \mathscr{P}^e_{\mathcal{Y},\mathcal{U}}$. Denote by P_e and C_e the respective locally Lipschitz-continuous causal extensions (which exist by assumption), and for notational convenience let $\mathcal{V}(:=\mathcal{U}\oplus\mathcal{Y}$. The closed-loop, denoted by $[P,C]$, is said to be well posed if the following three conditions hold:

(i) $\left[\begin{smallmatrix} I & C_e \\ P_e & I \end{smallmatrix}\right] : \mathrm{L}^{2,\mathrm{e}}_{\mathbb{R}^+}(\mathcal{V}) \to \mathrm{L}^{2,\mathrm{e}}_{\mathbb{R}^+}(\mathcal{V})$, the system mapping $\left[\begin{smallmatrix} u_1 \\ u_2 \end{smallmatrix}\right]$ to $\left[\begin{smallmatrix} d_1 \\ d_2 \end{smallmatrix}\right]$, is bijective, implying the existence of a *unique* solution to the functional equations that describe the closed-loop;

(ii) $H_e(P,C)(:=\left[\begin{smallmatrix} I & C_e \\ P_e & I \end{smallmatrix}\right]^{-1}$ is causal and locally Lipschitz-continuous;

(iii) The solution to the functional equations that describe the closed-loop is insensitive to very high frequency modeling errors, such as small transmission delays.

For further details and a discussion of the physical significance of well-posedness, the reader is referred to the seminal work of Willems [26, Chap. 4]. In addition to well-

[4] Since this system is only defined for positive time, shift-invariance implies causality.

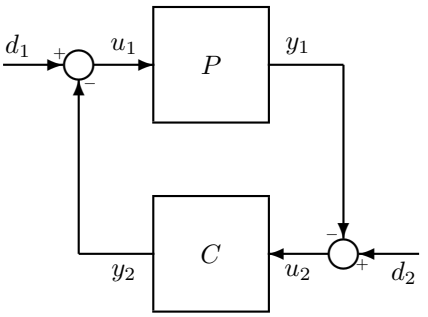


FIG. 2.1. *Standard feedback configuration.*

posedness, it is desirable for the system $H(P,C) : \mathcal{D}_{H(P,C)} \subset \mathrm{L}^2_{\mathbb{R}^+}(\mathcal{V}) \to \mathrm{L}^2_{\mathbb{R}^+}(\mathcal{V})$ induced by $H_e(P,C)$, to satisfy $\mathcal{D}_{H(P,C)} = \mathrm{L}^2_{\mathbb{R}^+}(\mathcal{V})$. In this case, the closed-loop is said to be stable and P is said to be stabilized by C. Since P and C are both linear, it can be shown that this also implies closed-loop stability with finite-gain [26, p. 117]; that is, $\|H(P,C)\| < \infty$.

Given a plant $P \in \mathscr{P}^e_{\mathcal{U},\mathcal{Y}}$ and a controller $C \in \mathscr{P}^e_{\mathcal{Y},\mathcal{U}}$, note that

$$\begin{bmatrix} I & C \\ P & I \end{bmatrix} : \mathcal{D}_P \oplus \mathcal{D}_C \to \mathcal{G}_P + \mathcal{G}^\sharp_C.$$

Thus, if $[P,C]$ is well posed, then $\mathcal{G}_P \cap \mathcal{G}^\sharp_C = \{0\}$. Furthermore, observe that $\mathcal{D}_{H(P,C)} = \mathcal{G}_P + \mathcal{G}^\sharp_C$ and hence that the closed-loop is stable if and only if $\mathcal{G}_P + \mathcal{G}^\sharp_C = \mathrm{L}^2_{\mathbb{R}^+}(\mathcal{V})$ (with $\mathcal{G}_P \cap \mathcal{G}^\sharp_C = \{0\}$). This useful geometric characterization of closed-loop stability is summarized in the following proposition, which has also appeared in [10, 12, 15] for other classes of systems.

PROPOSITION 2.4. *Given a well-posed plant/controller pair $P \in \mathscr{P}^e_{\mathcal{U},\mathcal{Y}}$ and $C \in \mathscr{P}^e_{\mathcal{Y},\mathcal{U}}$, the closed-loop $[P,C]$ is stable if and only if $\mathcal{G}_P$ and $\mathcal{G}^\sharp_C$ induce a coordinatization of $\mathrm{L}^2_{\mathbb{R}^+}(\mathcal{V})$. In this case, the parallel projections $\mathbf{\Pi}_{\mathcal{G}_P \| \mathcal{G}^\sharp_C}$ and $\mathbf{\Pi}_{\mathcal{G}^\sharp_C \| \mathcal{G}_P}$ are stable systems in $\mathscr{P}^e_{\mathcal{V},\mathcal{V}}$.*

REMARK 2.5. *Let the systems $P \in \mathscr{P}^e_{\mathcal{U},\mathcal{Y}}$ and $C \in \mathscr{P}^e_{\mathcal{Y},\mathcal{U}}$ constitute a well-posed plant/controller pair, and recall that $\mathcal{G}_P$ and $\mathcal{G}^\sharp_C$ are isomorphic to shift-invariant subspaces $\hat{\mathcal{G}}_{\mathsf{P}} := \mathbf{ZW}\mathcal{G}_P \subset \mathrm{H}^2_{\mathbb{D}}(\mathrm{L}^2_{\mathbb{H}}(\mathcal{V}))$ and $\hat{\mathcal{G}}^\sharp_{\mathsf{C}} := \mathbf{ZW}\mathcal{G}^\sharp_C \subset \mathrm{H}^2_{\mathbb{D}}(\mathrm{L}^2_{\mathbb{H}}(\mathcal{V}))$, respectively. It follows directly by Proposition 2.4 that the corresponding closed-loop $[P,C]$ is stable if and only if $\hat{\mathcal{G}}_{\mathsf{P}}$ and $\hat{\mathcal{G}}^\sharp_{\mathsf{C}}$ induce a coordinatization of $\mathrm{H}^2_{\mathbb{D}}(\mathrm{L}^2_{\mathbb{H}}(\mathcal{V}))$.*

3. Shift-invariant subspaces. In view of Remark 2.5, shift-invariant subspaces arise somewhat naturally in the study of LPTV, continuous-time systems. Accordingly, this section is concerned with the characterization of shift-invariant subspaces.[5] To this end, the Beurling–Lax–Halmos theorem [13, 17, 9] is used, which states that any shift-invariant subspace of $\mathrm{H}^2_{\mathbb{D}}$ can be expressed as the range of a multiplication operator with inner symbol. This result is extended here, to show that if the shift-invariant subspace is also shift-invariantly coordinatizable, then it can be expressed as the kernel of a multiplication operator with symbol in $\mathrm{H}^\infty_{\mathbb{D}}$. Furthermore, it is shown

[5] A subspace of a Hilbert space is linear by definition. Furthermore, a shift-invariant subspace $\mathcal{G}$ satisfies $\mathbf{S}\mathcal{G} \subset \mathcal{G}$, by definition.

that these range and kernel representations are, respectively, right and left invertible by multiplication operators with symbols in $\mathrm{H}^\infty_{\mathbb{D}}$. Using this result, strong-right and strong-left representations of the graph of an LPTV, continuous-time system may be constructed as shown in section 4, provided the system is stabilizable.

Let $\mathcal{G} \subset \mathrm{H}^2_{\mathbb{D}}(\mathcal{H})$ be a closed, shift-invariant subspace, where $\mathcal{H}$ is a Hilbert space. Then $\mathcal{G}$ is said to be *shift-invariantly coordinatizable* if there exists a closed, shift-invariant subspace $\mathcal{F} \subset \mathrm{H}^2_{\mathbb{D}}(\mathcal{H})$ such that $\mathcal{G}$ and $\mathcal{F}$ induce a coordinatization of $\mathrm{H}^2_{\mathbb{D}}(\mathcal{H})$; that is, such that $\mathcal{G} \cap \mathcal{F} = \{0\}$ and $\mathcal{G} + \mathcal{F} = \mathrm{H}^2_{\mathbb{D}}(\mathcal{H})$. Recall that when two (closed) subspaces, $\mathcal{G}$ and $\mathcal{F}$ say, induce a coordinatization of $\mathrm{H}^2_{\mathbb{D}}(\mathcal{H})$, any $\eta \in \mathrm{H}^2_{\mathbb{D}}(\mathcal{H})$ can be uniquely decomposed as the sum $\eta = \gamma + \varphi$, where $\gamma \in \mathcal{G}$ and $\varphi \in \mathcal{F}$. Furthermore, recall that the operator $\mathbf{\Pi}_{\mathcal{G}\|\mathcal{F}} := \eta \mapsto \gamma$ is called the parallel projection onto $\mathcal{G}$ along $\mathcal{F}$ and similarly, that $\mathbf{\Pi}_{\mathcal{F}\|\mathcal{G}} := \eta \mapsto \varphi$ is called the parallel projection onto $\mathcal{F}$ along $\mathcal{G}$. Since $\mathcal{G}$ and $\mathcal{F}$ are closed subspaces it follows that the graphs of the parallel-projection operators are closed and thus, that the parallel-projection operators are bounded, by the closed graph theorem [2, p. 80]. If $\mathcal{G}$ and $\mathcal{F}$ are shift-invariant subspaces, then it follows immediately that the parallel-projection operators are LSI.

PROPOSITION 3.1 (Beurling–Lax–Halmos theorem [13, 17, 9]). *Given a shift-invariant subspace $\mathcal{L} \subset \mathrm{H}^2_{\mathbb{D}}(\mathcal{H})$, there exists an inner function $\Lambda_r \in \mathrm{H}^\infty_{\mathbb{D}}(\mathcal{B}_{\mathcal{H}_1,\mathcal{H}})$ such that $\mathcal{L} = \Lambda_r \mathrm{H}^2_{\mathbb{D}}(\mathcal{H}_1)$, where $\mathcal{H}_1$ can be any Hilbert space isomorphic to $\mathcal{L} \ominus \mathbf{S}\mathcal{L}$.*

Let $\mathcal{G}$ and $\mathcal{F}$ be two shift-invariant subspaces that induce a coordinatization of $\mathrm{H}^2_{\mathbb{D}}(\mathcal{H})$. Then in view of Proposition 3.1, there exist *inner* functions $\Gamma_r \in \mathrm{H}^\infty_{\mathbb{D}}(\mathcal{B}_{\mathcal{H}_1,\mathcal{H}})$ and $\Phi_r \in \mathrm{H}^\infty_{\mathbb{D}}(\mathcal{B}_{\mathcal{H}_2,\mathcal{H}})$ such that $\mathcal{G} = \Gamma_r \mathrm{H}^2_{\mathbb{D}}(\mathcal{H}_1)$ and $\mathcal{F} = \Phi_r \mathrm{H}^2_{\mathbb{D}}(\mathcal{H}_2)$, where $\mathcal{H}_1$ is any Hilbert space isomorphic to $\mathcal{G} \ominus \mathbf{S}\mathcal{G}$ and $\mathcal{H}_2$ is any Hilbert space isomorphic to $\mathcal{F} \ominus \mathbf{S}\mathcal{F}$. Note that since $\mathbf{M}_{\Gamma_r}$ and $\mathbf{M}_{\Phi_r}$ are isometries (because their symbols are inner), the orthogonal projections $\mathbf{\Pi}_{\mathcal{G}}$ onto $\mathcal{G}$ and $\mathbf{\Pi}_{\mathcal{F}}$ onto $\mathcal{F}$ can be expressed as $\mathbf{M}_{\Gamma_r}(\mathbf{M}_{\Gamma_r})^*$ and $\mathbf{M}_{\Phi_r}(\mathbf{M}_{\Phi_r})^*$, respectively. This fact and the following technical result are used in the proof of the main theorem of this section.

LEMMA 3.2. *Given two closed, shift-invariant subspaces $\mathcal{G}$ and $\mathcal{F}$ that induce a coordinatization of $\mathrm{H}^2_{\mathbb{D}}(\mathcal{H})$, let $\Gamma_r \in \mathrm{H}^\infty_{\mathbb{D}}(\mathcal{B}_{\mathcal{H}_1,\mathcal{H}})$ and $\Phi_r \in \mathrm{H}^\infty_{\mathbb{D}}(\mathcal{B}_{\mathcal{H}_2,\mathcal{H}})$ be inner functions that satisfy $\mathcal{G} = \Gamma_r \mathrm{H}^2_{\mathbb{D}}(\mathcal{H}_1)$ and $\mathcal{F} = \Phi_r \mathrm{H}^2_{\mathbb{D}}(\mathcal{H}_2)$. Then the operators $(\mathbf{M}_{\Gamma_r})^* \mathbf{\Pi}_{\mathcal{G}\|\mathcal{F}}$ and $(\mathbf{M}_{\Phi_r})^* \mathbf{\Pi}_{\mathcal{F}\|\mathcal{G}}$ are bounded and LSI.*

Proof. It is only shown that $(\mathbf{M}_{\Gamma_r})^* \mathbf{\Pi}_{\mathcal{G}\|\mathcal{F}}$ is bounded and LSI. The proof that $(\mathbf{M}_{\Phi_r})^* \mathbf{\Pi}_{\mathcal{F}\|\mathcal{G}}$ is also bounded and LSI follows similarly. $(\mathbf{M}_{\Gamma_r})^* \mathbf{\Pi}_{\mathcal{G}\|\mathcal{F}}$ is clearly linear and bounded on $\mathrm{H}^2_{\mathbb{D}}(\mathcal{H})$, since it is the product of two bounded, linear operators. To see that it is also shift invariant, note that any $\eta \in \mathrm{H}^2_{\mathbb{D}}(\mathcal{H})$ is uniquely expressible as $\eta = \gamma + \varphi$, where $\gamma = \mathbf{\Pi}_{\mathcal{G}\|\mathcal{F}} \eta \in \mathcal{G}$ and $\varphi = \mathbf{\Pi}_{\mathcal{F}\|\mathcal{G}} \eta \in \mathcal{F}$, and since $\mathbf{M}_{\Gamma_r}$ is an isometry, that there exists a unique $\zeta \in \mathrm{H}^2_{\mathbb{D}}(\mathcal{H}_1)$ such that $\gamma = \mathbf{M}_{\Gamma_r} \zeta$. In fact, $\zeta = (\mathbf{M}_{\Gamma_r})^* \mathbf{\Pi}_{\mathcal{G}\|\mathcal{F}} \eta$. Now using the relationships $\mathbf{S}\eta = \mathbf{S}\gamma + \mathbf{S}\varphi$, $\mathbf{S}\gamma = \mathbf{M}_{\Gamma_r} \mathbf{S}\zeta \in \mathcal{G}$ and $\mathbf{S}\varphi \in \mathcal{F}$, it follows that

$$(\mathbf{M}_{\Gamma_r})^* \mathbf{\Pi}_{\mathcal{G}\|\mathcal{F}} \mathbf{S}\eta = (\mathbf{M}_{\Gamma_r})^* \mathbf{S}\gamma = \mathbf{S}\zeta = \mathbf{S}(\mathbf{M}_{\Gamma_r})^* \mathbf{\Pi}_{\mathcal{G}\|\mathcal{F}} \eta.$$

That is, $(\mathbf{M}_{\Gamma_r})^* \mathbf{\Pi}_{\mathcal{G}\|\mathcal{F}}$ is also shift invariant. ∎

The following theorem is the main result of this section. It is essentially an extension of the Beurling–Lax–Halmos theorem, in that it guarantees the existence of a kernel representation of a shift-invariant subspace if it is also shift-invariantly coordinatizable.

THEOREM 3.3. *Given a closed, shift-invariant subspace $\mathcal{G} \subset \mathrm{H}^2_{\mathbb{D}}(\mathcal{H})$ that is shift-invariantly coordinatizable, there exist functions $\Gamma_r \in \mathrm{H}^\infty_{\mathbb{D}}(\mathcal{B}_{\mathcal{H}_1,\mathcal{H}})$, $\Gamma_l \in \mathrm{H}^\infty_{\mathbb{D}}(\mathcal{B}_{\mathcal{H},\mathcal{H}_2})$,*

$\Phi_r \in \mathrm{H}^\infty_{\mathbb{D}}(\mathcal{B}_{\mathcal{H}_2,\mathcal{H}})$, *and* $\Phi_l \in \mathrm{H}^\infty_{\mathbb{D}}(\mathcal{B}_{\mathcal{H},\mathcal{H}_1})$ *such that*

$$\mathcal{G} = \mathcal{R}_{\mathbf{M}_{\Gamma_r}} = \mathcal{K}_{\mathbf{M}_{\Gamma_l}},$$

$$\begin{bmatrix}\mathbf{M}_{\Gamma_r} & \mathbf{M}_{\Phi_r}\end{bmatrix}\begin{bmatrix}\mathbf{M}_{\Phi_l}\\ \mathbf{M}_{\Gamma_l}\end{bmatrix} = \mathrm{I} \quad \textit{and} \quad \begin{bmatrix}\mathbf{M}_{\Phi_l}\\ \mathbf{M}_{\Gamma_l}\end{bmatrix}\begin{bmatrix}\mathbf{M}_{\Gamma_r} & \mathbf{M}_{\Phi_r}\end{bmatrix} = \begin{bmatrix}\mathrm{I} & 0\\ 0 & \mathrm{I}\end{bmatrix},$$

where $\mathcal{H}_1$ *can be any Hilbert space isomorphic to* $\mathcal{G} \ominus \mathbf{S}\mathcal{G}$, $\mathcal{H}_2$ *any Hilbert space isomorphic to* $\mathcal{F} \ominus \mathbf{S}\mathcal{F}$ *and* $\mathcal{F}$ *any closed, shift-invariant subspace that in conjunction with* $\mathcal{G}$ *induces a coordinatization of* $\mathrm{H}^2_{\mathbb{D}}(\mathcal{H})$.

Proof. By assumption there exists a shift-invariant subspace $\mathcal{F} \subset \mathrm{H}^2_{\mathbb{D}}(\mathcal{H})$ such that $\mathcal{G}$ and $\mathcal{F}$ induce a coordinatization of $\mathrm{H}^2_{\mathbb{D}}(\mathcal{H})$. Furthermore, by the Beurling–Lax–Halmos theorem, there exist inner functions $\Gamma_r \in \mathrm{H}^\infty_{\mathbb{D}}(\mathcal{B}_{\mathcal{H}_1,\mathcal{H}})$ and $\Phi_r \in \mathrm{H}^\infty_{\mathbb{D}}(\mathcal{B}_{\mathcal{H}_2,\mathcal{H}})$, with $\mathcal{H}_1 \sim \mathcal{G} \ominus \mathbf{S}\mathcal{G}$ and $\mathcal{H}_2 \sim \mathcal{F} \ominus \mathbf{S}\mathcal{F}$, such that $\mathcal{G} = \mathcal{R}_{\mathbf{M}_{\Gamma_r}}$ and $\mathcal{F} = \mathcal{R}_{\mathbf{M}_{\Phi_r}}$.

Now by Lemma 3.2, $(\mathbf{M}_{\Gamma_r})^*\mathbf{\Pi}_{\mathcal{G}\|\mathcal{F}}$ and $(\mathbf{M}_{\Phi_r})^*\mathbf{\Pi}_{\mathcal{F}\|\mathcal{G}}$ are bounded, LSI operators on $\mathrm{H}^2_{\mathbb{D}}(\mathcal{H})$. As such, by Proposition 2.3, there exist functions $\Phi_l \in \mathrm{H}^\infty_{\mathbb{D}}(\mathcal{B}_{\mathcal{H},\mathcal{H}_1})$ and $\Gamma_l \in \mathrm{H}^\infty_{\mathbb{D}}(\mathcal{B}_{\mathcal{H},\mathcal{H}_2})$, so that

$$\mathbf{M}_{\Phi_l} = (\mathbf{M}_{\Gamma_r})^*\mathbf{\Pi}_{\mathcal{G}\|\mathcal{F}} \quad \text{and} \quad \mathbf{M}_{\Gamma_l} = (\mathbf{M}_{\Phi_r})^*\mathbf{\Pi}_{\mathcal{F}\|\mathcal{G}}. \tag{3.1}$$

Since $\mathcal{R}_{\mathbf{\Pi}_{\mathcal{G}\|\mathcal{F}}} (= \mathcal{R}_{\mathbf{M}_{\Gamma_r}})$ is orthogonal to $\mathcal{K}_{(\mathbf{M}_{\Gamma_r})^*}$, it follows from (3.1) that $\mathcal{K}_{\mathbf{M}_{\Phi_l}} = \mathcal{K}_{\mathbf{\Pi}_{\mathcal{G}\|\mathcal{F}}} = \mathcal{F}$. Similarly, $\mathcal{K}_{\mathbf{M}_{\Gamma_l}} = \mathcal{G}$. Consequently it remains to show that

$$\begin{bmatrix}\mathbf{M}_{\Gamma_r} & \mathbf{M}_{\Phi_r}\end{bmatrix}\begin{bmatrix}\mathbf{M}_{\Phi_l}\\ \mathbf{M}_{\Gamma_l}\end{bmatrix} = \begin{bmatrix}\mathbf{M}_{\Phi_l}\\ \mathbf{M}_{\Gamma_l}\end{bmatrix}\begin{bmatrix}\mathbf{M}_{\Gamma_r} & \mathbf{M}_{\Phi_r}\end{bmatrix} = \begin{bmatrix}\mathrm{I} & 0\\ 0 & \mathrm{I}\end{bmatrix}.$$

To see this, first note that

$$\mathbf{\Pi}_{\mathcal{G}\|\mathcal{F}} = \mathbf{\Pi}_{\mathcal{G}}\mathbf{\Pi}_{\mathcal{G}\|\mathcal{F}} = \mathbf{M}_{\Gamma_r}(\mathbf{M}_{\Gamma_r})^*\mathbf{\Pi}_{\mathcal{G}\|\mathcal{F}} = \mathbf{M}_{\Gamma_r}\mathbf{M}_{\Phi_l}$$

and

$$\mathbf{\Pi}_{\mathcal{F}\|\mathcal{G}} = \mathbf{\Pi}_{\mathcal{F}}\mathbf{\Pi}_{\mathcal{F}\|\mathcal{G}} = \mathbf{M}_{\Phi_r}(\mathbf{M}_{\Phi_r})^*\mathbf{\Pi}_{\mathcal{F}\|\mathcal{G}} = \mathbf{M}_{\Phi_r}\mathbf{M}_{\Gamma_l}.$$

Using the relationship $\mathbf{\Pi}_{\mathcal{G}\|\mathcal{F}} + \mathbf{\Pi}_{\mathcal{F}\|\mathcal{G}} = \mathrm{I}$, it is then immediate that

$$\begin{bmatrix}\mathbf{M}_{\Gamma_r} & \mathbf{M}_{\Phi_r}\end{bmatrix}\begin{bmatrix}\mathbf{M}_{\Phi_l}\\ \mathbf{M}_{\Gamma_l}\end{bmatrix} = \mathrm{I}. \tag{3.2}$$

Now since $\mathcal{G} = \mathcal{R}_{\mathbf{M}_{\Gamma_r}} = \mathcal{K}_{\mathbf{M}_{\Gamma_l}}$, $\mathcal{F} = \mathcal{R}_{\mathbf{M}_{\Phi_r}} = \mathcal{K}_{\mathbf{M}_{\Phi_l}}$ and the subspaces $\mathcal{G}$ and $\mathcal{F}$ induce a coordinatization of $\mathrm{H}^2_{\mathbb{D}}(\mathcal{H})$, both $\begin{bmatrix}\mathbf{M}_{\Gamma_r} & \mathbf{M}_{\Phi_r}\end{bmatrix}$ and $\begin{bmatrix}\mathrm{M}_{\Phi_l}\\ \mathrm{M}_{\Gamma_1}\end{bmatrix}$ have zero kernel. Thus, they are the inverse of each other (see (3.2)) and

$$\begin{bmatrix}\mathbf{M}_{\Phi_l}\\ \mathbf{M}_{\Gamma_l}\end{bmatrix}\begin{bmatrix}\mathbf{M}_{\Gamma_r} & \mathbf{M}_{\Phi_r}\end{bmatrix} = \begin{bmatrix}\mathrm{I} & 0\\ 0 & \mathrm{I}\end{bmatrix}.$$

This completes the proof. ∎

REMARK 3.4. *The multiplication operator symbols in the last theorem are not unique. In fact, given any functions* $\Theta_0 \in \mathrm{H}^\infty_{\mathbb{D}}(\mathcal{B}_{\mathcal{H}_2,\mathcal{H}_1})$, $\Theta_1(\Theta_1^{-1}) \in \mathrm{H}^\infty_{\mathbb{D}}(\mathcal{B}_{\mathcal{H}_1,\mathcal{H}_1})$ *and* $\Theta_2(\Theta_2^{-1}) \in \mathrm{H}^\infty_{\mathbb{D}}(\mathcal{B}_{\mathcal{H}_2,\mathcal{H}_2})$, *the functions*

$$\Gamma_{r\star} := \Gamma_r\Theta_1, \quad \Gamma_{l\star} := \Theta_2^{-1}\Gamma_l, \quad \Phi_{r\star} := (\Phi_r - \Gamma_r\Theta_0)\Theta_2,$$

and

$$\Phi_{l\star}{:=}\Theta_1^{-1}(\Phi_l+\Theta_0\Gamma_l)$$

satisfy $\mathcal{G}=\mathcal{R}_{\mathbf{M}_{\Gamma_{r\star}}}=\mathcal{K}_{\mathbf{M}_{\Gamma_{l\star}}}$,

$$\begin{bmatrix}\mathbf{M}_{\Gamma_{r\star}} & \mathbf{M}_{\Phi_{r\star}}\end{bmatrix}\begin{bmatrix}\mathbf{M}_{\Phi_{l\star}}\\ \mathbf{M}_{\Gamma_{l\star}}\end{bmatrix}=\mathrm{I},\quad \textit{and}\quad \begin{bmatrix}\mathbf{M}_{\Phi_{l\star}}\\ \mathbf{M}_{\Gamma_{l\star}}\end{bmatrix}\begin{bmatrix}\mathbf{M}_{\Gamma_{r\star}} & \mathbf{M}_{\Phi_{r\star}}\end{bmatrix}=\begin{bmatrix}\mathrm{I} & 0\\ 0 & \mathrm{I}\end{bmatrix}.$$

4. Stabilizability and existence of strong-right/left representations. Recall the standard closed-loop configuration shown in Figure 2.1, which is denoted by $[P,C]$. Also recall that a system $P\in\mathscr{P}^e_{\mathcal{U},\mathcal{Y}}$ is said to be stabilizable if there exists a system $C\in\mathscr{P}^e_{\mathcal{Y},\mathcal{U}}$ such that the closed-loop $[P,C]$ is stable. In this section, it is shown that provided an LPTV system is stabilizable, there exist strong-right and strong-left representations of its graph. That is, it is shown that the graph of any stabilizable, LPTV system can be expressed as the range (respectively, kernel) of a stable, LPTV system that is left (respectively, right) invertible by a stable, LPTV system. In view of Proposition 2.4 and Remark 2.5, this result follows *almost* immediately by Theorem 3.3. The difficulty, however, is that the relevant multiplication operators shown to exist in Theorem 3.3 may correspond (via the **Z**-transform and time-lifting isomorphisms) to systems that are *not* locally causal in each $[kh,(k+1)h)$ interval of continuous-time (see Remark 2.2). Fortunately though, it is possible to exploit the nonuniqueness of these range and kernel representations (see Remark 3.4) to ensure causality, as detailed in the proof of the following theorem.

THEOREM 4.1. *Let* $\mathcal{U}$ *and* $\mathcal{Y}$ *be Hilbert spaces and for notational convenience define* $\mathcal{V}{:=}\mathcal{U}\oplus\mathcal{Y}$. *If* $P\in\mathscr{P}^e_{\mathcal{U},\mathcal{Y}}$ *can be stabilized by some* $C\in\mathscr{P}^e_{\mathcal{Y},\mathcal{U}}$, *then there exist stable systems* $G_r\in\mathscr{P}^e_{\mathcal{U},\mathcal{V}}$, $G_l\in\mathscr{P}^e_{\mathcal{V},\mathcal{Y}}$, $K_l\in\mathscr{P}^e_{\mathcal{V},\mathcal{U}}$, *and* $K_r\in\mathscr{P}^e_{\mathcal{Y},\mathcal{V}}$ *such that*

$$\mathcal{G}_P=\mathcal{R}_{G_r}=\mathcal{K}_{G_l}$$

and

$$\begin{bmatrix}G_r & K_r\end{bmatrix}\begin{bmatrix}K_l\\ G_l\end{bmatrix}=\begin{bmatrix}K_l\\ G_l\end{bmatrix}\begin{bmatrix}G_r & K_r\end{bmatrix}=\begin{bmatrix}I & 0\\ 0 & I\end{bmatrix}.$$

Proof. Let $C\in\mathscr{P}^e_{\mathcal{Y},\mathcal{Y}}$ be a stabilizing controller for the plant P. Then by Remark 2.5 it follows that the shift-invariant subspaces $\hat{\mathcal{G}}_{\mathsf{P}}{:=}\mathbf{ZW}\mathcal{G}_P\subset\mathrm{H}^2_{\mathbb{D}}(\mathrm{L}^2_{\mathbb{H}}(\mathcal{V}))$ and $\hat{\mathcal{G}}^{\sharp}_{\mathsf{C}}{:=}\mathbf{ZW}\mathcal{G}^{\sharp}_C\subset\mathrm{H}^2_{\mathbb{D}}(\mathrm{L}^2_{\mathbb{H}}(\mathcal{V}))$ induce a coordinatization of $\mathrm{H}^2_{\mathbb{D}}(\mathrm{L}^2_{\mathbb{H}}(\mathcal{V}))$. Correspondingly, there exist, by Theorem 3.3, functions

$$\hat{\mathsf{G}}_r\in\mathrm{H}^\infty_{\mathbb{D}}(\mathcal{B}_{\mathcal{H}_1,\mathrm{L}^2_{\mathbb{H}}(\mathcal{V})}),\quad \hat{\mathsf{G}}_l\in\mathrm{H}^\infty_{\mathbb{D}}(\mathcal{B}_{\mathrm{L}^2_{\mathbb{H}}(\mathcal{V}),\mathcal{H}_2}),\quad \hat{\mathsf{K}}_l\in\mathrm{H}^\infty_{\mathbb{D}}(\mathcal{B}_{\mathrm{L}^2_{\mathbb{H}}(\mathcal{V}),\mathcal{H}_1}),$$

and $\hat{\mathsf{K}}_r\in\mathrm{H}^\infty_{\mathbb{D}}(\mathcal{B}_{\mathcal{H}_2,\mathrm{L}^2_{\mathbb{H}}(\mathcal{V})})$ such that

$$\hat{\mathcal{G}}_{\mathsf{P}}=\mathcal{R}_{\mathbf{M}_{\hat{\mathsf{G}}_r}}=\mathcal{K}_{\mathbf{M}_{\hat{\mathsf{G}}_l}}$$

and

$$\begin{bmatrix}\mathbf{M}_{\hat{\mathsf{G}}_r} & \mathbf{M}_{\hat{\mathsf{K}}_r}\end{bmatrix}\begin{bmatrix}\mathbf{M}_{\hat{\mathsf{K}}_l}\\ \mathbf{M}_{\hat{\mathsf{G}}_l}\end{bmatrix}=\begin{bmatrix}\mathbf{M}_{\hat{\mathsf{K}}_l}\\ \mathbf{M}_{\hat{\mathsf{G}}_l}\end{bmatrix}\begin{bmatrix}\mathbf{M}_{\hat{\mathsf{G}}_r} & \mathbf{M}_{\hat{\mathsf{K}}_r}\end{bmatrix}=\begin{bmatrix}\mathrm{I} & 0\\ 0 & \mathrm{I}\end{bmatrix},$$

where $\mathcal{H}_1$ is any Hilbert space isomorphic to $\hat{\mathcal{G}}_{\mathsf{P}} \ominus \mathbf{S}\hat{\mathcal{G}}_{\mathsf{P}}$ and $\mathcal{H}_2$ is any Hilbert space isomorphic to $\hat{\mathcal{G}}_{\mathsf{C}}^{\sharp} \ominus \mathbf{S}\hat{\mathcal{G}}_{\mathsf{C}}^{\sharp}$. In fact, $\mathcal{H}_1$ may be taken to be $\mathrm{L}^2_{\mathbb{H}}(\mathcal{U})$ and $\mathcal{H}_2$ to be $\mathrm{L}^2_{\mathbb{H}}(\mathcal{Y})$, as is done in the rest of the proof. That this is possible is shown in Appendix A.

Now define

$$\mathsf{G}_r := \mathbf{Z}^{-1}\mathbf{M}_{\hat{\mathsf{G}}_r}\mathbf{Z}, \quad \mathsf{G}_l := \mathbf{Z}^{-1}\mathbf{M}_{\hat{\mathsf{G}}_l}\mathbf{Z}, \quad \mathsf{K}_l := \mathbf{Z}^{-1}\mathbf{M}_{\hat{\mathsf{K}}_r}\mathbf{Z},$$

and

$$\mathsf{K}_r := \mathbf{Z}^{-1}\mathbf{M}_{\hat{\mathsf{K}}_l}\mathbf{Z}.$$

Then $\mathcal{G}_{\mathsf{P}} = \mathbf{W}\mathcal{G}_P = \mathcal{R}_{\mathsf{G}_r} = \mathcal{K}_{\mathsf{G}_l}$ and

$$\begin{bmatrix}\mathsf{G}_r & \mathsf{K}_r\end{bmatrix}\begin{bmatrix}\mathsf{K}_l \\ \mathsf{G}_l\end{bmatrix} = \begin{bmatrix}\mathsf{K}_l \\ \mathsf{G}_l\end{bmatrix}\begin{bmatrix}\mathsf{G}_r & \mathsf{K}_r\end{bmatrix} = \begin{bmatrix}\mathrm{I} & 0 \\ 0 & \mathrm{I}\end{bmatrix}.$$

Furthermore, given any stable, LSI system

$$\mathsf{Q}_0 : \ell^2_{\mathbb{Z}^+}(\mathrm{L}^2_{\mathbb{H}}(\mathcal{U})) \to \ell^2_{\mathbb{Z}^+}(\mathrm{L}^2_{\mathbb{H}}(\mathcal{Y}))$$

and stable, LSI systems

$$\mathsf{Q}_1 : \ell^2_{\mathbb{Z}^+}(\mathrm{L}^2_{\mathbb{H}}(\mathcal{U})) \to \ell^2_{\mathbb{Z}^+}(\mathrm{L}^2_{\mathbb{H}}(\mathcal{U})) \quad \text{and} \quad \mathsf{Q}_2 : \ell^2_{\mathbb{Z}^+}(\mathrm{L}^2_{\mathbb{H}}(\mathcal{Y})) \to \ell^2_{\mathbb{Z}^+}(\mathrm{L}^2_{\mathbb{H}}(\mathcal{Y}))$$

with stable inverses, observe (as in Remark 3.4) that the stable, LSI systems

$$\mathsf{G}_{r\star} := \mathsf{G}_r\mathsf{Q}_1, \quad \mathsf{G}_{l\star} := \mathsf{Q}_2^{-1}\mathsf{G}_l, \quad \mathsf{K}_{r\star} := (\mathsf{K}_r - \mathsf{G}_r\mathsf{Q}_0)\mathsf{Q}_2,$$

and $\mathsf{K}_{l\star} := \mathsf{Q}_1^{-1}(\mathsf{K}_l + \mathsf{Q}_0\mathsf{G}_l)$ satisfy

$$\mathcal{G}_{\mathsf{P}} = \mathcal{R}_{\mathsf{G}_{r\star}} = \mathcal{K}_{\mathsf{G}_{l\star}}$$

and

$$\begin{aligned}
\begin{bmatrix}\mathsf{G}_{r\star} & \mathsf{K}_{r\star}\end{bmatrix}\begin{bmatrix}\mathsf{K}_{l\star} \\ \mathsf{G}_{l\star}\end{bmatrix} &:= \begin{bmatrix}\mathsf{G}_r\mathsf{Q}_1 & (\mathsf{K}_r - \mathsf{G}_r\mathsf{Q}_0)\mathsf{Q}_2\end{bmatrix}\begin{bmatrix}\mathsf{Q}_1^{-1}(\mathsf{K}_l + \mathsf{Q}_0\mathsf{G}_l) \\ \mathsf{Q}_2^{-1}\mathsf{G}_l\end{bmatrix} \\
&= \begin{bmatrix}\mathsf{Q}_1^{-1}(\mathsf{K}_l + \mathsf{Q}_0\mathsf{G}_l) \\ \mathsf{Q}_2^{-1}\mathsf{G}_l\end{bmatrix}\begin{bmatrix}\mathsf{G}_r\mathsf{Q}_1 & (\mathsf{K}_r - \mathsf{G}_r\mathsf{Q}_0)\mathsf{Q}_2\end{bmatrix} \\
&= \begin{bmatrix}\mathsf{Q}_1^{-1} & \mathsf{Q}_1^{-1}\mathsf{Q}_0 \\ 0 & \mathsf{Q}_2^{-1}\end{bmatrix}\begin{bmatrix}\mathsf{K}_l \\ \mathsf{G}_l\end{bmatrix}\begin{bmatrix}\mathsf{G}_r & \mathsf{K}_r\end{bmatrix}\begin{bmatrix}\mathsf{Q}_1 & -\mathsf{Q}_0\mathsf{Q}_2 \\ 0 & \mathsf{Q}_2\end{bmatrix} \\
&= \begin{bmatrix}\mathrm{I} & 0 \\ 0 & \mathrm{I}\end{bmatrix}. \qquad (4.1)
\end{aligned}$$

Consequently, if it were possible to choose Q_0, Q_1, and Q_2 to ensure that $(\mathsf{G}_r\mathsf{Q}_1)_{[0]}$, $(\mathsf{Q}_2^{-1}\mathsf{G}_l)_{[0]}$, $(\mathsf{Q}_1^{-1}(\mathsf{K}_l + \mathsf{Q}_0\mathsf{G}_l))_{[0]}$, and $((\mathsf{K}_r + \mathsf{G}_r\mathsf{Q}_0)\mathsf{Q}_2)_{[0]}$ are all causal mappings on the finite horizon $\mathbb{H}$, it would follow by Remark 2.2 that

$$G_{r\star} := \mathbf{W}^{-1}\mathsf{G}_{r\star}\mathbf{W}, \quad G_{l\star} := \mathbf{W}^{-1}\mathsf{G}_{l\star}\mathbf{W}, \quad K_{r\star} := \mathbf{W}^{-1}\mathsf{K}_{r\star}\mathbf{W},$$

and $K_{l\star} := \mathbf{W}^{-1}\mathsf{K}_{l\star}\mathbf{W}$ are all stable, LPTV systems satisfying

$$\mathcal{G}_P = \mathcal{R}_{G_{r\star}} = \mathcal{K}_{G_{l\star}}$$

and

$$\begin{bmatrix} G_{r\star} & K_{r\star} \end{bmatrix} \begin{bmatrix} K_{l\star} \\ G_{l\star} \end{bmatrix} = \begin{bmatrix} K_{l\star} \\ G_{l\star} \end{bmatrix} \begin{bmatrix} G_{r\star} & K_{r\star} \end{bmatrix} = \begin{bmatrix} I & 0 \\ 0 & I \end{bmatrix}.$$

Accordingly, the rest of the proof is dedicated to showing that this is possible.

Let

$$\begin{bmatrix} \mathsf{M}_{r[0]} \\ \mathsf{N}_{r[0]} \end{bmatrix} := \mathsf{G}_{r[0]} \quad \text{and} \quad \begin{bmatrix} -\mathsf{N}_{l[0]} & \mathsf{M}_{l[0]} \end{bmatrix} := \mathsf{G}_{l[0]},$$

where the partitioning is conformal with that of G_P and the subscript [0] denotes the first element of the sequence uniquely identifiable with the associated block-Toeplitz representation. Now recall that, by definition, $P \in \mathscr{P}^e_{\mathcal{U},\mathcal{Y}}$ is causally extendible with locally Lipschitz-continuous extension. As such, it follows that both $\mathsf{M}_{r[0]} \in \mathcal{B}_{\mathrm{L}^2_{\mathbb{H}}(\mathcal{U}),\mathrm{L}^2_{\mathbb{H}}(\mathcal{U})}$ and $\mathsf{M}_{l[0]} \in \mathcal{B}_{\mathrm{L}^2_{\mathbb{H}}(\mathcal{Y}),\mathrm{L}^2_{\mathbb{H}}(\mathcal{Y})}$ are bijective and hence, boundedly invertible by the open mapping theorem [2, p. 79]. To see this, first note that since P is causally extendible,

$$\mathcal{R}_{\mathsf{M}_{r[0]}} = \mathbf{T}_h \mathcal{D}_P = \mathrm{L}^2_{\mathbb{H}}(\mathcal{U}).$$

Consequently, $\mathsf{M}_{r[0]}$ is surjective. Suppose now that there exists some nonzero $\overset{\circ}{q} \in \mathrm{L}^2_{\mathbb{H}}(\mathcal{U})$ such that $\mathsf{M}_{r[0]}\, \overset{\circ}{q} = 0$. Then by causality of P it follows that $\mathsf{N}_{r[0]}\, \overset{\circ}{q}$ must also be zero. Since $\mathsf{K}_{l[0]}\mathsf{G}_{r[0]} = \mathrm{I}$, this implies that

$$0 = \mathsf{K}_{l[0]} \begin{bmatrix} \mathsf{M}_{r[0]} \\ \mathsf{N}_{r[0]} \end{bmatrix} \overset{\circ}{q} = \overset{\circ}{q},$$

which is a contradiction. Hence, the linear operator $\mathsf{M}_{r[0]}$ must be injective. This, and the fact that it is also surjective, implies that $\mathsf{M}_{r[0]}$ is bijective as claimed. Now consider $\mathsf{M}_{l[0]}$, and suppose that there exists a nonzero $\overset{\circ}{q} \in \mathrm{L}^2_{\mathbb{H}}(\mathcal{Y})$ such that $\mathsf{M}_{l[0]}\, \overset{\circ}{q} = 0$. Then since

$$\mathcal{K}_{\begin{bmatrix} -\mathsf{N}_{l[0]} & \mathsf{M}_{l[0]} \end{bmatrix}} = \mathbf{T}_h \mathcal{G}_P,$$

it follows that $\begin{bmatrix} 0 \\ \overset{\circ}{q} \end{bmatrix} \in \mathbf{T}_h \mathcal{G}_P$. But this contradicts the causality of P and hence $\overset{\circ}{q}$ must be 0. Consequently, $\mathsf{M}_{l[0]}$ has zero kernel and is thereby injective. Moreover, since $\mathsf{G}_{l[0]}\mathsf{K}_{r[0]} = \mathrm{I}$,

$$\mathcal{R}_{\begin{bmatrix} -\mathsf{N}_{l[0]} & \mathsf{M}_{l[0]} \end{bmatrix}} = \mathrm{L}^2_{\mathbb{H}}(\mathcal{Y}).$$

So for all $\overset{\circ}{e} \in \mathrm{L}^2_{\mathbb{H}}(\mathcal{Y})$, there exists a $\overset{\circ}{g} = \begin{bmatrix} \overset{\circ}{g}_1 \\ \overset{\circ}{g}_2 \end{bmatrix} \in \mathrm{L}^2_{\mathbb{H}}(\mathcal{V})$ such that

$$\overset{\circ}{e} = \begin{bmatrix} -\mathsf{N}_{l[0]} & \mathsf{M}_{l[0]} \end{bmatrix} \begin{bmatrix} \overset{\circ}{g}_1 \\ \overset{\circ}{g}_2 \end{bmatrix}.$$

In fact,

$$\overset{\circ}{e} = \begin{bmatrix} -\mathsf{N}_{l[0]} & \mathsf{M}_{l[0]} \end{bmatrix} \begin{bmatrix} \overset{\circ}{g}_1 + \overset{\circ}{u} \\ \overset{\circ}{g}_2 + \mathsf{P}_{[0]}\, \overset{\circ}{u} \end{bmatrix}$$

for all $\overset{\circ}{u} \in \mathbf{T}_h\mathcal{D}_P$. Since $\mathbf{T}_h\mathcal{D}_P = \mathrm{L}^2_{\mathbb{H}}(\mathcal{U})$, $\overset{\circ}{u}$ can be set to $-\overset{\circ}{g}_1$. Doing this gives $\overset{\circ}{e} = \mathsf{M}_{l[0]}(\overset{\circ}{g}_2 - \mathsf{P}_{[0]}\overset{\circ}{g}_1)$, from which it is evident that $\mathcal{R}_{\mathsf{M}_{l[0]}} = \mathrm{L}^2_{\mathbb{H}}(\mathcal{Y})$. Correspondingly, $\mathsf{M}_{l[0]}$ is bijective as claimed.

Having shown that both $\mathsf{M}_{r[0]}$ and $\mathsf{M}_{l[0]}$ are boundedly invertible, it is possible to construct the required Q_0, Q_1, and Q_2. Let Q_1 satisfy $\mathsf{Q}_{1[0]} = \mathsf{M}^{-1}_{r[0]}$ and $\mathsf{Q}_{1[i]} = 0$ for $i \neq 0$, where $\{\mathsf{Q}_{1[i]}\}_{i=0}^{\infty}$ denotes the sequence uniquely identifiable with the block-Toeplitz representation of Q_1. Then

$$(\mathsf{G}_r\mathsf{Q}_1)_{[0]} = \begin{bmatrix} \mathrm{I} \\ \mathsf{N}_{r[0]}\mathsf{M}^{-1}_{r[0]} \end{bmatrix},$$

which is causal on $\mathrm{L}^2_{\mathbb{H}}(\mathcal{U})$, since P is a causal system. Now let $\begin{bmatrix}\mathsf{Y}_{l[0]} & \mathsf{X}_{l[0]}\end{bmatrix} := \mathsf{K}_{l[0]}$, where the partition is conformal with $\mathcal{G}_{\mathsf{P}}$, and construct Q_0 so that $\mathsf{Q}_{0[0]} = -\mathsf{X}_{l[0]}\mathsf{M}^{-1}_{l[0]}$ and $\mathsf{Q}_{0[i]} = 0$ for $i \neq 0$. Then since $\mathsf{N}_{r[0]}\mathsf{M}^{-1}_{r[0]} = \mathsf{M}^{-1}_{l[0]}\mathsf{N}_{l[0]} = \mathsf{P}_{[0]}$ and $\mathsf{Y}_{l[0]}\mathsf{M}_{r[0]} + \mathsf{X}_{l[0]}\mathsf{N}_{r[0]} = \mathrm{I}$, it follows that

$$\begin{aligned}(\mathsf{Q}_1^{-1}(\mathsf{K}_l + \mathsf{Q}_0\mathsf{G}_l))_{[0]} &= (\mathsf{Q}_1^{-1})_{[0]}\left(\mathsf{K}_{l[0]} + \mathsf{Q}_{0[0]}\begin{bmatrix}-\mathsf{N}_{l[0]} & \mathsf{M}_{l[0]}\end{bmatrix}\right) \\ &= \mathsf{M}_{r[0]}\begin{bmatrix}\mathsf{Y}_{l[0]} + \mathsf{X}_{l[0]}\mathsf{M}^{-1}_{l[0]}\mathsf{N}_{l[0]} & 0\end{bmatrix} \\ &= \mathsf{M}_{r[0]}\begin{bmatrix}\mathsf{M}^{-1}_{r[0]} & 0\end{bmatrix} = \begin{bmatrix}\mathrm{I} & 0\end{bmatrix},\end{aligned}$$

which is obviously causal on $\mathrm{L}^2_{\mathbb{H}}(\mathcal{V})$, as required. Finally, let $\left[\begin{smallmatrix}-\mathsf{X}_{r[0]} \\ \mathsf{Y}_{r[0]}\end{smallmatrix}\right] := \mathsf{K}_{r[0]}$. Then from (4.1), it follows that $\mathsf{M}_{r[0]}\mathsf{X}_{l[0]} - \mathsf{X}_{r[0]}\mathsf{M}_{l[0]} = 0$ and hence, that $\mathsf{Q}_{0[0]} = -\mathsf{X}_{l[0]}\mathsf{M}^{-1}_{l[0]} = \mathsf{M}^{-1}_{r[0]}\mathsf{X}_{r[0]}$. Defining Q_2 to satisfy $\mathsf{Q}_{2[0]} = \mathsf{M}_{l[0]}$ and $\mathsf{Q}_{2[i]} = 0$ for $i \neq 0$, gives $(\mathsf{Q}_2^{-1}\mathsf{G}_l)_{[0]} = [-\mathsf{M}^{-1}_{l[0]}\mathsf{N}_{l[0]} \quad \mathrm{I}]$ and $((\mathsf{K}_r + \mathsf{G}_r\mathsf{Q}_0)\mathsf{Q}_2)_{[0]} = \left[\begin{smallmatrix}0 \\ \mathrm{I}\end{smallmatrix}\right]$. As required, each is respectively causal on $\mathrm{L}^2_{\mathbb{H}}(\mathcal{V})$ and $\mathrm{L}^2_{\mathbb{H}}(\mathcal{Y})$, completing the proof. ∎

REMARK 4.2. *Since G_r and G_l in Theorem 4.1 are, respectively, left and right invertible by stable, LPTV systems, they are called strong-right and strong-left representations of $\mathcal{G}_P$, respectively. Note that they are only unique up to invertible factors. That is, for any stable systems $Q_1(Q_1^{-1}) \in \mathscr{P}^e_{\mathcal{U},\mathcal{U}}$, and $Q_2(Q_2^{-1}) \in \mathscr{P}^e_{\mathcal{Y},\mathcal{Y}}$, the systems G_rQ_1 and $Q_2^{-1}G_l$ are also strong-right and strong-left representations of $\mathcal{G}_P$, respectively.*

5. Closed-loop stability and stabilizing controllers. The next result constitutes a useful characterization of closed-loop stability in terms of range and kernel representations of $\mathcal{G}_P$ and $\mathcal{G}^\sharp_C$. In turn, this result leads to a Youla-style parametrization of stabilizing controllers.

LEMMA 5.1. *Given a well-posed plant/controller pair $P \in \mathscr{P}^e_{\mathcal{U},\mathcal{Y}}$ and $C \in \mathscr{P}^e_{\mathcal{Y},\mathcal{U}}$, let $G_r \in \mathscr{P}^e_{\mathcal{U},\mathcal{V}}$ be any stable system with zero kernel such that $\mathcal{G}_P = \mathcal{R}_{G_r}$, and $K_l \in \mathscr{P}^e_{\mathcal{V},\mathcal{U}}$ be any stable system such that $\mathcal{R}_{K_l} = \mathrm{L}^2_{\mathbb{R}^+}(\mathcal{U})$ and $\mathcal{G}^\sharp_C = \mathcal{K}_{K_l}$. Then the following are equivalent:*

(i) *The closed-loop system $[P, C]$ is stable.*
(ii) *The stable, LPTV system K_lG_r is bijective (and hence boundedly invertible, although the inverse may not be causal).*

Furthermore, if $[P, C]$ is stable and G_r and K_l are, respectively, strong-right and strong-left representations of $\mathcal{G}_P$ and $\mathcal{G}^\sharp_C$, then $(K_lG_r)^{-1}$ is a stable system in $\mathscr{P}^e_{\mathcal{U},\mathcal{U}}$. That is, the inverse is also causal.

Proof. Since the closed-loop is well posed by assumption, it follows by Proposition 2.4 and the identities $\mathcal{G}_P = \mathcal{R}_{G_r}$ and $\mathcal{G}_C^\sharp = \mathcal{K}_{K_l}$ that

$$[P,C] \text{ is stable } \Leftrightarrow \mathcal{R}_{G_r} + \mathcal{K}_{K_l} = \mathrm{L}^2_{\mathbb{R}^+}(\mathcal{V}) \text{ and } \mathcal{R}_{G_r} \cap \mathcal{K}_{K_l} = \{0\}.$$

Since $\mathcal{R}_{K_l} = \mathrm{L}^2_{\mathbb{R}^+}(\mathcal{U})$ and $\mathcal{K}_{G_r} = \{0\}$, it is clear that

$$\begin{gathered}\mathcal{R}_{G_r} + \mathcal{K}_{K_l} = \mathrm{L}^2_{\mathbb{R}^+}(\mathcal{V}) \text{ and } \mathcal{R}_{G_r} \cap \mathcal{K}_{K_l} = \{0\} \\ \Downarrow \\ \mathcal{R}_{K_l G_r} = \mathrm{L}^2_{\mathbb{R}^+}(\mathcal{U}) \text{ and } \mathcal{K}_{K_l G_r} = \{0\}.\end{gathered}$$

To see that the converse is also true, let $\mathcal{R}_{K_l G_r} = \mathrm{L}^2_{\mathbb{R}^+}(\mathcal{U})$ and $\mathcal{K}_{K_l G_r} = \{0\}$. That $\mathcal{R}_{G_r} \cap \mathcal{K}_{K_l} = \{0\}$ is immediate. Now suppose that there exists a $v \in \mathrm{L}^2_{\mathbb{R}^+}(\mathcal{V})$ such that $v \notin \mathcal{R}_{G_r} + \mathcal{K}_{K_l}$, and define $e{:=}K_l v$. Then since $\mathcal{R}_{K_l G_r} = \mathrm{L}^2_{\mathbb{R}^+}(\mathcal{U})$ and $\mathcal{K}_{K_l G_r} = \{0\}$, there exists a unique $q \in \mathrm{L}^2_{\mathbb{R}^+}(\mathcal{U})$ such that $e = K_l G_r q = K_l g$, where $g{:=}G_r q$. Consequently, $K_l(g - v) = 0$, which implies either $v = g$ or $(g - v) \in \mathcal{K}_{K_l}$. But this contradicts $v \notin \mathcal{R}_{G_r} + \mathcal{K}_{K_l}$, since $g \in \mathcal{R}_{G_r}$. Consequently, it follows that $[P,C]$ is stable if and only if the bounded linear operator $K_l G_r$ is bijective (and hence, boundedly invertible by the open mapping theorem [2, p. 79]). Note that $(K_l G_r)^{-1}$ is periodically time-varying but that it may not be locally causal on $\mathrm{L}^2_{\mathbb{H}}(\mathcal{U})$.

It is now shown that if $[P,C]$ is stable and G_r and K_l are, respectively, strong-right and strong-left representations of $\mathcal{G}_P$ and $\mathcal{G}_C^\sharp$, then $(K_l G_r)^{-1} \in \mathscr{P}^e_{\mathcal{U},\mathcal{U}}$. First note that for any K_l and G_r that satisfy the original conditions of the lemma,

$$\mathbf{\Pi}_{\mathcal{G}_P \| \mathcal{G}_C^\sharp} = G_r (K_l G_r)^{-1} K_l. \tag{5.1}$$

To see this, note that since $[P,C]$ is stable, the subspaces $\mathcal{G}_P$ and $\mathcal{G}_C^\sharp$ induce a coordinatization of $\mathrm{L}^2_{\mathbb{R}^+}(\mathcal{V})$. Thereby, any $g \in \mathrm{L}^2_{\mathbb{R}^+}(\mathcal{V})$ can be written uniquely as the sum $g = p + c$, where $p \in \mathcal{G}_P$ and $c \in \mathcal{G}_C^\sharp$. Furthermore, since $\mathcal{K}_{G_r} = \{0\}$ and $\mathcal{R}_{G_r} = \mathcal{G}_P$, there exists a unique $q \in \mathrm{L}^2_{\mathbb{R}^+}(\mathcal{U})$ such that $p = G_r q$. Consequently,

$$\begin{aligned}\left(\mathbf{\Pi}_{\mathcal{G}_P \| \mathcal{G}_C^\sharp} - G_r (K_l G_r)^{-1} K_l\right) g &= p - G_r (K_l G_r)^{-1} K_l p \\ &= p - G_r (K_l G_r)^{-1} K_l G_r q \\ &= p - G_r q = 0\end{aligned}$$

for all $g \in \mathrm{L}^2_{\mathbb{R}^+}(\mathcal{U})$, implying that $\mathbf{\Pi}_{\mathcal{G}_P \| \mathcal{G}_C^\sharp} = G_r (K_l G_r)^{-1} K_l$. Now take G_r and K_l to be *strong-right* and *strong-left* representations. Then, G_r and K_l satisfy the original conditions of the lemma and there exist stable systems $K_{l\star} \in \mathscr{P}^e_{\mathcal{V},\mathcal{U}}$ and $G_{r\star} \in \mathscr{P}^e_{\mathcal{U},\mathcal{V}}$ such that $K_{l\star} G_r = I$ and $K_l G_{r\star} = I$. Now from (5.1),

$$(K_l G_r)^{-1} = K_{l\star} \mathbf{\Pi}_{\mathcal{G}_P \| \mathcal{G}_C^\sharp} G_{r\star},$$

which since $\mathbf{\Pi}_{\mathcal{G}_P \| \mathcal{G}_C^\sharp} \in \mathscr{P}^e_{\mathcal{V},\mathcal{V}}$ (see Proposition 2.4) implies that $(K_l G_r)^{-1} \in \mathscr{P}^e_{\mathcal{U},\mathcal{U}}$. This completes the proof. ∎

REMARK 5.2. *Note that by interchanging the roles of P and C in Lemma 5.1, given any stable system $G_l \in \mathscr{P}^e_{\mathcal{Y},\mathcal{Y}}$ such that $\mathcal{R}_{G_l} = \mathrm{L}^2_{\mathbb{R}^+}(\mathcal{U})$ and $\mathcal{G}_P = \mathcal{K}_{G_l}$, and any stable system $K_r \in \mathscr{P}^e_{\mathcal{Y},\mathcal{V}}$ with zero kernel such that $\mathcal{G}_C^\sharp = \mathcal{R}_{K_r}$, the following are equivalent*:

(i) *The closed-loop system $[P,C]$ is stable.*

(ii) *The stable, LPTV system G_lK_r is bijective (and hence boundedly invertible, although the inverse may not be causal).*

Furthermore, if G_l and K_r are, respectively, strong-left and strong-right representations of $\mathcal{G}_P$ and $\mathcal{G}_C^\sharp$, and $[P, C]$ is stable, then $(G_lK_r)^{-1}$ is a stable system $\mathscr{P}^e_{\mathcal{Y},\mathcal{Y}}$.

Interestingly, given a system $P \in \mathscr{P}^e_{\mathcal{U},\mathcal{Y}}$ and a stabilizing controller $C \in \mathscr{P}^e_{\mathcal{Y},\mathcal{U}}$, strong-right and strong-left representations of $\mathcal{G}_C^\sharp$ can be constructed from strong-right and strong-left representations of $\mathcal{G}_P$ and the respective stable left and right inverses. This is analogous to necessity of the well-known Youla parametrization of stabilizing controllers for certain classes of systems [27, 7, 6, 5], as summarized in the following theorem.

THEOREM 5.3. *Given a stabilizable system $P \in \mathscr{P}^e_{\mathcal{U},\mathcal{Y}}$ and stable systems $G_r \in \mathscr{P}^e_{\mathcal{U},\mathcal{V}}$, $G_l \in \mathscr{P}^e_{\mathcal{V},\mathcal{Y}}$, $K_l \in \mathscr{P}^e_{\mathcal{V},\mathcal{U}}$, and $K_r \in \mathscr{P}^e_{\mathcal{Y},\mathcal{V}}$ such that*

$$\mathcal{G}_P = \mathcal{R}_{G_r} = \mathcal{K}_{G_l}$$

and

$$\begin{bmatrix} G_r & K_r \end{bmatrix} \begin{bmatrix} K_l \\ G_l \end{bmatrix} = \begin{bmatrix} K_l \\ G_l \end{bmatrix} \begin{bmatrix} G_r & K_r \end{bmatrix} = \begin{bmatrix} I & 0 \\ 0 & I \end{bmatrix},$$

a system $C \in \mathscr{P}^e_{\mathcal{Y},\mathcal{U}}$ stabilizes P only if there exists a stable $Q \in \mathscr{P}^e_{\mathcal{Y},\mathcal{U}}$ such that

$$K_r - G_rQ$$

is a strong-right representation of $\mathcal{G}_C^\sharp$ and

$$K_l + QG_l$$

is a strong-left representation of $\mathcal{G}_C^\sharp$.

Proof. Let $C \in \mathscr{P}^e_{\mathcal{Y},\mathcal{U}}$ denote a stabilizing controller for P. Then it follows by Theorem 4.1 that there exist stable systems $G_r \in \mathscr{P}^e_{\mathcal{U},\mathcal{V}}$, $G_l \in \mathscr{P}^e_{\mathcal{V},\mathcal{Y}}$, $K_l \in \mathscr{P}^e_{\mathcal{V},\mathcal{U}}$, and $K_r \in \mathscr{P}^e_{\mathcal{Y},\mathcal{V}}$ such that

$$\mathcal{G}_P = \mathcal{R}_{G_r} = \mathcal{K}_{G_l}$$

and

$$\begin{bmatrix} G_r & K_r \end{bmatrix} \begin{bmatrix} K_l \\ G_l \end{bmatrix} = \begin{bmatrix} K_l \\ G_l \end{bmatrix} \begin{bmatrix} G_r & K_r \end{bmatrix} = \begin{bmatrix} I & 0 \\ 0 & I \end{bmatrix}. \tag{5.2}$$

Furthermore, for any stable system $Q \in \mathscr{P}^e_{\mathcal{Y},\mathcal{U}}$,

$$\begin{aligned} &\begin{bmatrix} G_r & (K_r - G_rQ) \end{bmatrix} \begin{bmatrix} (K_l + QG_l) \\ G_l \end{bmatrix} \\ &= \begin{bmatrix} (K_l + QG_l) \\ G_l \end{bmatrix} \begin{bmatrix} G_r & (K_r - G_rQ) \end{bmatrix} = \begin{bmatrix} I & 0 \\ 0 & I \end{bmatrix}. \end{aligned} \tag{5.3}$$

Now from (5.3),

$$(K_l + QG_l)(K_r - G_rQ) = 0,$$

implying that $\mathcal{R}_{(K_r - G_rQ)} \subset \mathcal{K}_{(K_l + QG_l)}$. To see that the opposite inclusion also holds, note from (5.3) that

$$(K_l + QG_l) \begin{bmatrix} G_r & (K_r - G_rQ) \end{bmatrix} = \begin{bmatrix} I & 0 \end{bmatrix}. \tag{5.4}$$

Now suppose there exists a nonzero $x \in \mathcal{K}_{(K_l+QG_l)}$ that is not in $\mathcal{R}_{(K_r-G_rQ)}$. Then since

$$\mathcal{K}\begin{bmatrix} G_r & (K_r - G_rQ)\end{bmatrix} = \{0\} \quad \text{and} \quad \mathcal{R}\begin{bmatrix} G_r & (K_r - G_rQ)\end{bmatrix} = \mathrm{L}^2_{\mathbb{R}^+}(\mathcal{V}),$$

there exists a unique $u =: \left[\begin{smallmatrix} u_1 \\ u_2 \end{smallmatrix}\right] \in \mathrm{L}^2_{\mathbb{R}^+}(\mathcal{V})$ such that

$$x = \begin{bmatrix} G_r & (K_r - G_rQ)\end{bmatrix}\begin{bmatrix} u_1 \\ u_2 \end{bmatrix}.$$

Furthermore, $0 = (K_l + QG_l)x = u_1$, which using (5.4), implies that $x \in \mathcal{R}_{(K_r-G_rQ)}$. This is a contradiction and, therefore, $\mathcal{K}_{(K_l+QG_l)} \subset \mathcal{R}_{(K_r-G_rQ)}$.

Since $[P, C]$ is stable (or equivalently C is stabilized P), it follows by Theorem 4.1 that there exist strong-right and strong-left representations of $\mathcal{G}_C$. Noting that $\mathcal{G}^\sharp_C := \left[\begin{smallmatrix} 0 & I \\ I & 0 \end{smallmatrix}\right]\mathcal{G}_C$, these give rise to strong-right and strong-left representations of $\mathcal{G}^\sharp_C$. Denote these by $K_{r\star} \in \mathscr{P}^e_{\mathcal{Y},\mathcal{V}}$ and $K_{l\star} \in \mathscr{P}^e_{\mathcal{V},\mathcal{U}}$, respectively. It then follows by Lemma 5.1 that $K_{l\star}G_r$ is boundedly invertible in $\mathscr{P}^e_{\mathcal{U},\mathcal{U}}$. Now define

$$Q := -(K_{l\star}G_r)^{-1}K_{l\star}K_r,$$

noting that Q is a stable system in $\mathscr{P}^e_{\mathcal{Y},\mathcal{U}}$. Then using (5.2),

$$\begin{aligned} K_l - QG_l &= K_l + (K_{l\star}G_r)^{-1}K_{l\star}K_rG_l \\ &= K_l + (K_{l\star}G_r)^{-1}K_{l\star}(I - G_rK_l) \\ &= K_l - K_l + (K_{l\star}G_r)^{-1}K_{l\star} = (K_{l\star}G_r)^{-1}K_{l\star}, \end{aligned}$$

implying that $\mathcal{K}_{(K_l-QG_l)} = \mathcal{K}_{K_{l\star}}$. Since $\mathcal{R}_{(K_r+G_rQ)} = \mathcal{K}_{(K_l-QG_l)}$ and $\mathcal{R}_{K_{r\star}} = \mathcal{K}_{K_{l\star}}$, it also follows that $\mathcal{R}_{(K_r+G_rQ)} = \mathcal{R}_{K_{r\star}}$. That is, the strong-right and strong-left representations of $\mathcal{G}^\sharp_C$ are in the form required. □

REMARK 5.4. *If in Theorem* 5.3 *the choice of* Q *was restricted to the stable systems in* $\mathscr{P}^e_{\mathcal{U},\mathcal{Y}}$ *for which the subspace* $\mathcal{R}_{(K_r-G_rQ)} = \mathcal{K}_{(K_l+QG_l)}$ *corresponds to the inverse graph of a system* $C \in \mathscr{P}^e_{\mathcal{U},\mathcal{Y}}$ *such that the closed-loop* $[P, C]$ *is well posed, then* C *would stabilize* P. *That is, Theorem* 5.3 *would also be sufficient. To see this, note from* (5.3) *that for such a* Q, *the system* $(K_l + QG_l)$ *is a strong-left representation of* $\mathcal{G}^\sharp_C$. *So by arguments similar to those used in the proof of Lemma* 5.1, *and the fact that* $(K_l + QG_l)G_r = I$ *is obviously invertible, it follows that* $[P, C]$ *is stable. Again from* (5.3), *and since* $\mathcal{R}_{(K_r-G_rQ)} = \mathcal{K}_{(K_l+QG_l)} = \mathcal{G}^\sharp_C$, *it also follows that* $(K_r - G_rQ)$ *is a strong-right representation of the inverse graph of a stabilizing controller.*

Appendix A. Supplement to the proof of Theorem 4.1. As required in Theorem 4.1, it is shown here that $\mathrm{L}^2_{\mathbb{H}}(\mathcal{U})$ is isomorphic to $\hat{\mathcal{G}}_{\mathsf{P}} \ominus \mathbf{S}\hat{\mathcal{G}}_{\mathsf{P}}$ and that $\mathrm{L}^2_{\mathbb{H}}(\mathcal{Y})$ is isomorphic to $\hat{\mathcal{G}}^\sharp_{\mathsf{C}} \ominus \mathbf{S}\hat{\mathcal{G}}^\sharp_{\mathsf{C}}$, where all of these objects are defined as in the proof of the theorem. Recall that there exists a function $\hat{\mathsf{G}}_r \in \mathrm{H}^\infty_{\mathbb{D}}(\mathcal{B}_{\mathcal{H}_1,\mathrm{L}^2_{\mathbb{H}}(\mathcal{V})})$ such that

$$\hat{\mathcal{G}}_{\mathsf{P}} = \hat{\mathsf{G}}_r\mathrm{H}^2_{\mathbb{D}}(\mathcal{H}_1),$$

where $\mathcal{H}_1$ is a Hilbert space isomorphic to $\hat{\mathcal{G}}_{\mathsf{P}} \ominus \mathbf{S}\hat{\mathcal{G}}_{\mathsf{P}}$. Note that since $P \in \mathscr{P}_{\mathcal{U},\mathcal{Y}}$ is causally extendible, $\mathbf{T}_h\mathcal{D}_P = \mathrm{L}^2_{\mathbb{H}}(\mathcal{U})$ and hence

$$\mathsf{G}_{r[0]}\mathcal{H}_1 = \begin{bmatrix} \mathrm{I} \\ \mathsf{P}_{[0]} \end{bmatrix}\mathrm{L}^2_{\mathbb{H}}(\mathcal{U}), \tag{A.1}$$

where $\mathsf{G}_{r[0]}(=\hat{\mathsf{G}}_r(0))$ and $\mathsf{P}_{[0]}$ denote the first elements of the sequences uniquely identifiable with the block-Toeplitz representations of $\mathsf{G}_r{:=}\mathbf{Z}^{-1}\mathbf{M}_{\hat{\mathsf{G}}_r}\mathbf{Z}$ and $\mathsf{P}{:=}\mathbf{W}P\mathbf{W}^{-1}$, respectively. Also recall that $\mathbf{M}_{\hat{\mathsf{G}}_r}$ has a left-inverse $\mathbf{M}_{\hat{\mathsf{K}}_l}$ with $\hat{\mathsf{K}}_l \in \mathrm{H}^\infty_{\mathbb{D}}(\mathcal{B}_{\mathrm{L}^2_{\mathbb{H}}(\mathcal{V}),\mathcal{H}_1})$. As such, $\mathsf{K}_{l[0]} = \hat{\mathsf{K}}_l(0)$ is a bounded left-inverse of $\mathsf{G}_{r[0]}$.[6] Define

$$\mathsf{Q}{:=}\mathsf{K}_{l[0]}\begin{bmatrix}\mathrm{I}\\ \mathsf{P}_{[0]}\end{bmatrix} : \mathrm{L}^2_{\mathbb{H}}(\mathcal{U}) \to \mathcal{H}_1.$$

Then since $\begin{bmatrix}\mathrm{I}\\ \mathsf{P}_{[0]}\end{bmatrix}$ has zero kernel and $\mathsf{K}_{l[0]}\mathsf{G}_{r[0]} = \mathrm{I}$, it follows from (A.1) that Q is a bijective mapping. Hence, $\mathsf{Q}(\mathsf{Q}^*\mathsf{Q})^{-\frac{1}{2}}$ is an isomorphism between $\mathrm{L}^2_{\mathbb{H}}(\mathcal{U})$ and $\mathcal{H}_1$. Correspondingly, $\mathrm{L}^2_{\mathbb{H}}(\mathcal{U})$ is isomorphic to $\hat{\mathcal{G}}_{\mathsf{P}} \ominus \mathbf{S}\hat{\mathcal{G}}_{\mathsf{P}} \sim \mathcal{H}_1$. It can be shown in a similar manner that $\mathrm{L}^2_{\mathbb{H}}(\mathcal{Y})$ is isomorphic to $\hat{\mathcal{G}}^\sharp_{\mathsf{C}} \ominus \mathbf{S}\hat{\mathcal{G}}^\sharp_{\mathsf{C}}$.

REFERENCES

[1] B. A. Bamieh, J. B. Pearson, B. A. Francis, and A. R. Tannenbaum, *A lifting technique for linear periodic systems with applications to sampled-data control*, Systems Control Lett., 17 (1991), pp. 79–88.

[2] B. Bollobás, *Linear Analysis: An Introductory Course*, Cambridge Math. Textbooks, Cambridge University Press, Cambridge, UK, 1990.

[3] M. W. Cantoni, *Linear Periodic Systems: Robustness Analysis and Sampled-Data Control*, Ph.D. thesis, Department of Engineering, University of Cambridge, England, 1998.

[4] M. W. Cantoni and K. Glover, *Gap-metric robustness analysis of linear periodically time-varying feedback systems*, SIAM J. Control Optim., 38 (2000, pp. 803–822.

[5] W. N. Dale and M. C. Smith, *Stabilizability and existence of system representations for discrete-time time-varying systems*, SIAM J. Control Optim., 31 (1993), pp. 1538–1557.

[6] C. A. Desoer and C. L. Gustafson, *Algebraic theory of linear multivariable feedback systems*, IEEE Trans. Automat. Control, 29 (1984), pp. 909–917.

[7] C. A. Desoer, R.-W. Liu, J. Murray, and R. Saeks, *Feedback system design: The fractional representation approach to analysis and synthesis*, IEEE Trans. Automat. Control, 25 (1980), pp. 399–412.

[8] J. C. Doyle, T. T. Georgiou, and M. C. Smith, *The parallel projection operators of a nonlinear feedback system*, Systems Control Lett., 20 (1993), pp. 79–85.

[9] C. Foiaş and A. Frazho, *The Commutant Lifting Approach to Interpolation Problems*, Oper. Theory Adv. Appl. 44, Birkhäuser-Verlag, Berlin, 1990.

[10] C. Foiaş, T. T. Georgiou, and M. C. Smith, *Robust stability of feedback systems: A geometric approach using the gap metric*, SIAM J. Control Optim., 31 (1993), pp. 1518–1537.

[11] T. T. Georgiou and M. C. Smith, *Optimal robustness in the gap metric*, IEEE Trans. Automat. Control, 35 (1990), pp. 673–686.

[12] T. T. Georgiou and M. C. Smith, *Graphs, causality and stabilizability: Linear, shift-invariant systems on* $\mathrm{L}^2([0,\infty))$, Math. Control Signals Systems, 6 (1993), pp. 195–223.

[13] P. D. Lax, *Translation invariant subspaces*, Acta Math., 101 (1959), pp. 163–178.

[14] D. C. McFarlane and K. Glover, *Robust Controller Design Using Normalised Coprime Factorisation Plant Descriptions*, Lecture Notes in Control and Inform. Sci. 138, Springer-Verlag, Berlin, 1990.

[15] R. J. Ober and J. A. Sefton, *Stability of control systems and graphs of linear systems*, Systems Control Lett., 17 (1991), pp. 265–280.

[16] R. Ravi, A. M. Pascoal, and P. P. Khargonekar, *Normalized coprime factorizations for linear time-varying systems*, Systems Control Lett., 18 (1992), pp. 455–465.

[17] M. Rosenblum and J. Rovnyak, *Hardy Classes and Operator Theory*, Oxford University Press, London, 1985.

[18] M. A. Rotea and P. P. Khargonekar, *Stabilizability of linear time-varying and uncertain linear systems*, IEEE Trans. Automat. Control, 33 (1988), pp. 884–887.

[6] That $\mathsf{G}_{r[0]}$ has a bounded left-inverse also follows from the fact that $\hat{\mathsf{G}}_r$ may be taken to be inner (see the proof of Theorem 3.3).

[19] J. A. Sefton and R. J. Ober, *On the gap metric and coprime factor perturbations*, Automatica J. IFAC, 29 (1993), pp. 723–734.
[20] M. C. Smith, *On stabilisation and the existence of coprime factorisations*, IEEE Trans. Automat. Control, 34 (1989), pp. 1005–1007.
[21] B. Sz-Nagy and C. Foiaş, *Harmonic Analysis of Operators on Hilbert Space*, North-Holland, Amsterdam, 1970.
[22] M. Vidyasagar, *The graph metric for unstable plants and robustness estimates for feedback stability*, IEEE Trans. Automat. Control, 29 (1984), pp. 403–417.
[23] M. Vidyasagar, *Control System Synthesis: A Factorization Approach*, The MIT Press, Cambridge, MA, 1985.
[24] M. Vidyasagar, H. Schneider, and B. A. Francis, *Algebraic and topological aspects of feedback stabilisation*, IEEE Trans. Automat. Control, 27 (1983), pp. 880–894.
[25] G. Vinnicombe, *Frequency domain uncertainty and the graph topology*, IEEE Trans. Automat. Control, 38 (1993), pp. 1371–1383.
[26] J. C. Willems, *The Analysis of Feedback Systems*, in Research Monographs 62, The MIT Press, Cambridge, MA, 1971.
[27] D. C. Youla, H. A. Jabr, and J. J. Bongiorno, *Modern Wiener-Hopf design of optimal controllers—part* II*: The multivariable case*, IEEE Trans. Automat. Control, 21 (1976), pp. 319–338.

SIAM J. CONTROL OPTIM.
Vol. 38, No. 3, pp. 803–822

GAP-METRIC ROBUSTNESS ANALYSIS OF LINEAR PERIODICALLY TIME-VARYING FEEDBACK SYSTEMS*

MICHAEL CANTONI† AND KEITH GLOVER†

Abstract. Quantitative robust stability results are established in this paper for feedback systems that evolve in continuous-time and exhibit linear, periodically time-varying (LPTV) behavior. The results presented are analogous to results known to hold for linear, time-invariant (LTI) systems, although they do not follow directly from these. System uncertainty is measured using the gap metric, which quantifies the distance between systems in terms of the aperture between their graphs (the subspaces corresponding to all input-output pairs for each system). The main robustness result characterizes the largest gap-ball of LPTV plants stabilized by a nominal LPTV feedback controller known to stabilize a nominal LPTV plant at the center of this ball. A key step in the proof of this result makes use of a formula derived for the directed gap between LPTV systems. This formula is essentially a generalization of Georgiou's for LTI systems. Importantly, all of the results presented apply to a class of sampled-data control systems, as a special case.

Key words. gap metric, robust stability, periodic time-variation, sampled-data systems

AMS subject classifications. 93C05, 93C25, 9C350, 93C60, 93D09, 93D25

PII. S0363012998346591

1. Introduction.

The motivation for this paper stems from a desire to characterize the robustness properties of sampled-data (SD) control systems. Such systems are invariably time-varying, but often the time-variation is periodic [6]. As such, the approach taken here is to study general linear, periodically time-varying (LPTV) feedback systems.

Most systematic frameworks for control system design are model based. Correspondingly, it is imperative to account for the uncertainty that may be associated with any mathematical model of a physical process. Importantly, feedback systems can be made robust to uncertainty of this kind. In fact, this property is the principal reason for using feedback in control systems. The uncertainty that feedback systems can tolerate may be qualified mathematically by introducing a *suitable* topology on the set of systems concerned and quantified in terms of any metric that induces such a topology. Consider the standard closed-loop configuration shown in Figure 1.1. This is said to be stable if $H(P,C) := \left[\begin{smallmatrix} I & C \\ P & I \end{smallmatrix}\right]^{-1}$ is causal and bounded. A topology suitable for robustness analysis is defined in [25, 23] (for quite clear reasons) as one that satisfies the following properties: Given any plant P in some set of interest and a stabilizing controller C, there exists a neighborhood of P (in this topology on the set of plants) such that all plants in this neighborhood are also stabilized by C and the mapping $P \mapsto H(P,C)$ is continuous at P (with respect to this topology and the induced-norm topology on the corresponding set of stable closed loops). That is, for any sequence $\{P_i\}_{i=0}^{\infty}$ converging to P (with respect to a topology with these properties), all but a finite number of the plants in the sequence must be stabilized by C and $H(P_i,C) \to H(P,C)$ in the induced-norm topology. Put more simply,

*Received by the editors October 29, 1998; accepted for publication (in revised form) March 22, 1999; published electronically March 8, 2000. This research was supported in part by the Gledden Studentship of the University of Western Australia and the Engineering and Physical Sciences Research Council (EPSRC), UK.

http://www.siam.org/journals/sicon/38-3/34659.html

†Department of Engineering, University of Cambridge, Cambridge CB2 1PZ, UK (mwc@eng.cam.ac.uk, kg@eng.cam.ac.uk).

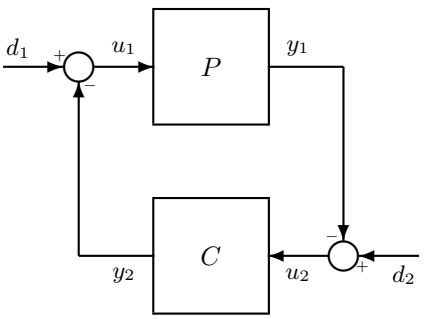


FIG. 1.1. *Standard closed-loop configuration.*

closed-loop stability is maintained in sufficiently small neighborhoods and closed-loop performance changes gradually in the face of tolerable uncertainty.

For stable systems, a topology suitable for studying robustness is the induced-norm topology. This framework lies at the heart of H^∞ control theory as developed in the seminal paper of Zames [30]. A shortcoming of the induced-norm topology is the inability to consider (explicitly) unstable components of a given closed loop. This shortcoming was addressed by defining the so-called graph topology as described below (see also [31, 23, 24]). For certain classes of system, a topology satisfying the properties described above can be characterized explicitly in terms of the graph representation of systems (the subspace of all input-output pairs) [23]. Such a topology is correspondingly referred to as the graph topology. Qualitatively, two systems are "close" in the graph topology if their graph subspaces are "close." For systems known to admit fractional representations that are coprime (such as LTI systems [24, 21], for example), a basis for the graph topology can be constructed by additively perturbing range and kernel representations of the graph, which can be constructed from the coprime factors [24]. It is also shown in [24] that for such systems, the graph topology (defined in terms of the basis described above) is the weakest with respect to which closed-loop performance varies continuously and closed-loop stability is a robust property.

The topological aspects of robustness analysis described so far qualify the types of uncertainty that are tolerable in a feedback sense. To quantify a level of robustness in these terms, a metric to induce the graph topology is required. For LTI systems a number of metrics have this property, including the graph metric [23], the gap metric [13], and the ν-gap metric [26]. Correspondingly, there are many well-known quantitative robustness results expressed in terms of these metrics for LTI systems [23, 31, 9, 13, 14, 19, 26].

All of the robust stability results derived in this paper for LPTV feedback systems are expressed in terms of the gap metric. Given two systems, the gap between them is defined to be the aperture between their corresponding graph subspaces [16]. It is well known that the gap between two subspaces is the maximum of two directed gaps [17] and, correspondingly, the gap between two systems is also the maximum of two directed gaps. In this paper, a formula is derived for the directed gap between two LPTV systems. The formula characterizes the directed gap in terms of a two-block H^∞-optimization problem involving a particular representation of the graph. As such, it is essentially a generalization of that given in [13] for LTI systems. The formula derived is used in a key step of the main robustness result further on, and, importantly,

it can be used to derive an algorithm for computing the gap [4, Chap. 5].

The quantitative robust stability results derived in this paper for LPTV feedback systems are, to a large extent, analogous to the results of [14] for LTI systems and to those of [12] for linear, time-varying (LTV) systems. However, the LPTV results do not follow directly from either the LTI or the LTV results. In particular, problems concerning causality are not addressed in [12], and although the periodic nature of the time-variation considered here imparts a high degree of structure to the analysis, it is somewhat more involved than in the time-invariant case. As for LTI systems, it can be shown that for a general class of LPTV systems, the gap metric induces the weakest topology with respect to which closed-loop performance varies continuously and closed-loop stability is a robust property. This result is not presented explicitly here but can be found in [4, Chap. 4]. All of the results obtained accommodate possibly infinite-dimensional input and output spaces and, importantly, can be applied to LPTV SD control systems as a special case.

The paper has the following structure. In section 2, the notation, terminology, and preliminary mathematics used throughout the paper are introduced. Section 3 is devoted to the gap metric, with the main result being a formula for the directed gap between two LPTV systems. In section 4, quantitative robustness results are obtained in terms of the gap metric. The main result characterizes the largest gap-ball of LPTV plants centered at a nominal plant that a nominal stabilizing LPTV controller is guaranteed to stabilize. In addition to this, robustness to simultaneous gap-perturbations of both the nominal plant and the controller is considered. To conclude this section, the main robustness result is specialized to SD control systems in which the sampling and hold devices are synchronized and periodic.

2. Preliminaries. In this section, general notation and definitions used throughout the paper are introduced. Let $\mathbb{R}$, $\mathbb{Z}$, $\mathbb{R}^+$, $\mathbb{Z}^+$, $\mathbb{D}$, and $\mathbb{T}$ denote the reals, integers, nonnegative reals, nonnegative integers, open unit-disc, and unit-circle, respectively. For convenience, given a real number $h > 0$, the interval $[0, h) \subset \mathbb{R}$ is denoted by $\mathbb{H}$. In any Hilbert space $\mathcal{H}$, the inner-product is denoted by $\langle \cdot, \cdot \rangle_{\mathcal{H}}$ and the norm by $\| \cdot \|_{\mathcal{H}}$. For a subspace $\mathcal{U} \subset \mathcal{H}$, the orthogonal complement of $\mathcal{U}$ in $\mathcal{H}$ is denoted by $\mathcal{H} \ominus \mathcal{U}$ (or $\mathcal{U}^\perp$, when the ambient space $\mathcal{H}$ is clear from context), and the orthogonal projection onto $\mathcal{U}$ is denoted by $\mathbf{\Pi}_{\mathcal{U}}$.

Consider two closed subspaces $\mathcal{G}$ and $\mathcal{F}$ of a Hilbert space $\mathcal{H}$. If $\mathcal{G} \cap \mathcal{F} = \{0\}$ and $\mathcal{G} + \mathcal{F} = \mathcal{H}$, then $\mathcal{G}$ and $\mathcal{F}$ are said to induce a *coordinatization* of $\mathcal{H}$. In this case, any $h \in \mathcal{H}$ can be uniquely decomposed as the sum $h = g + f$, where $g \in \mathcal{G}$ and $f \in \mathcal{F}$ (see [7], for example). The bounded, linear operator $\mathbf{\Pi}_{\mathcal{G}\|\mathcal{F}} : h \mapsto g$ is called the *parallel projection* onto $\mathcal{G}$ along $\mathcal{F}$. Similarly, $\mathbf{\Pi}_{\mathcal{F}\|\mathcal{G}} : h \mapsto f$ is called the parallel projection onto $\mathcal{F}$ along $\mathcal{G}$.

Let $\mathcal{U}$ and $\mathcal{Y}$ be arbitrary Hilbert spaces and consider a linear operator $P : \mathcal{D}_P \subset \mathcal{U} \to \mathcal{Y}$, where $\mathcal{D}_P := \{u \in \mathcal{U} : Pu \in \mathcal{Y}\}$ is called the domain of P. The range of P is defined to be $\mathcal{R}_P := \{Pu : u \in \mathcal{D}_P\}$ and $\mathcal{K}_P := \{u \in \mathcal{D}_P : Pu = 0\}$ is called the kernel of P. The *graph* of P is defined to be the totality of all input-output pairs

$$\mathcal{G}_P := \begin{bmatrix} I \\ P \end{bmatrix} \mathcal{D}_P \subset \mathcal{U} \oplus \mathcal{Y},$$

and for notational convenience the *inverse graph* is denoted by $\mathcal{G}_P^\sharp := \left[\begin{smallmatrix} 0 & I \\ I & 0 \end{smallmatrix}\right] \mathcal{G}_P$. Note that a (linear) subspace $\mathcal{G} \subset \mathcal{U} \oplus \mathcal{Y}$ corresponds to the graph of a linear operator if and only if $\left[\begin{smallmatrix} 0 \\ y \end{smallmatrix}\right] \in \mathcal{G}$ implies that $y = 0$. The symbol $\mathcal{B}_{\mathcal{U},\mathcal{Y}}$ is used to denote the

Banach space of all bounded, linear operators $P : \mathcal{U} \to \mathcal{Y}$; that is, all such operators with $\mathcal{D}_P = \mathcal{U}$ and finite induced-norm

$$\|P\| := \sup_{\substack{u \in \mathcal{D}_P \\ u \neq 0}} \frac{\|Pu\|_{\mathcal{Y}}}{\|u\|_{\mathcal{U}}} < \infty.$$

Given an operator $P \in \mathcal{B}_{\mathcal{U},\mathcal{Y}}$, there exist a unique operator $P^* \in \mathcal{B}_{\mathcal{Y},\mathcal{U}}$ such that for all $u \in \mathcal{U}$ and $y \in \mathcal{Y}$,

$$\langle y, Pu \rangle_{\mathcal{Y}} = \langle P^* y, u \rangle_{\mathcal{U}}.$$

The operator P^* is called the (Hilbert space) adjoint. Note that for any $P \in \mathcal{B}_{\mathcal{U},\mathcal{Y}}$, $P = (P^*)^*$ and $\mathcal{R}_P$ is orthogonal to $\mathcal{K}_{P^*}$. An operator $P \in \mathcal{B}_{\mathcal{U},\mathcal{Y}}$ is called an isometry if $\langle Pu, Pu \rangle_{\mathcal{Y}} = \langle u, u \rangle_{\mathcal{U}}$ for all $u \in \mathcal{U}$ (or equivalently, $P^*P = I$).

2.1. Signals and systems. In this paper, a system is simply considered to be an operator mapping between signal-spaces. Primarily, attention is focused on systems mapping between continuous-time spaces of signals with finite energy. Mathematically such signals can be thought of as functions in $\mathrm{L}^2_{\mathbb{R}^+}(\mathcal{H})$, the Hilbert space of $\mathcal{H}$-valued, (Lebesgue) square-integrable functions $f : \mathbb{R}^+ \to \mathcal{H}$. By virtue of the class of systems considered in what follows and the analysis technique employed to study these systems, the following signal-spaces also play an important role: the discrete-time signal-space $\ell^2_{\mathbb{Z}^+}(\mathcal{H})$ of square-summable sequences $\mathsf{f} : \mathbb{Z}^+ \to \mathcal{H}$, and the frequency-domain signal-space $\mathrm{H}^2_{\mathbb{D}}(\mathcal{H})$ of functions $\varphi : \mathbb{D} \to \mathcal{H}$ that are analytic in $\mathbb{D}$ and satisfy

$$\int_0^{2\pi} \left\langle \varphi(re^{j\omega}), \varphi(re^{j\omega}) \right\rangle_{\mathcal{H}} d\omega < M$$

for some $M < \infty$ and all $0 \leq r < 1$. Note that $\mathrm{H}^2_{\mathbb{D}}(\mathcal{H})$ is isomorphic to $\ell^2_{\mathbb{Z}^+}(\mathcal{H})$ via the $\mathbf{Z}$-transform isomorphism [22, pp. 184–185], defined for all $\mathsf{f} \in \ell^2_{\mathbb{Z}^+}(\mathcal{H})$ by

$$(\mathbf{Z}\mathsf{f})(\lambda) := \sum_{i=0}^{\infty} \mathsf{f}_k \lambda^k, \quad \lambda \in \mathbb{D}.$$

Let $\mathcal{U}$ and $\mathcal{Y}$ be Hilbert spaces. For a linear, continuous-time system $P : \mathcal{D}_P \subset \mathrm{L}^2_{\mathbb{R}^+}(\mathcal{U}) \to \mathrm{L}^2_{\mathbb{R}^+}(\mathcal{Y})$, the standard notion of causality can be expressed as follows: P is *causal* if for all $\tau \in \mathbb{R}^+$, $\mathbf{T}_\tau \mathcal{G}_P$ corresponds to the graph of a linear operator, where $\mathbf{T}_\tau$ is the projection that truncates a signal to zero after time τ. Given a real number $h > 0$, such a system P is called *periodically time-varying* (with period h) if $\mathbf{U}_{kh} \mathcal{G}_P \subset \mathcal{G}_P$ for all $k \in \mathbb{Z}^+$, where $\mathbf{U}_\tau$ denotes the unilateral (forward) shift on $\mathrm{L}^2_{\mathbb{R}^+}(\cdot)$.[1]

The following technical definitions are necessary to facilitate a precise definition of the class of systems considered in what follows. A causal, linear, continuous-time system $P : \mathcal{D}_P \subset \mathrm{L}^2_{\mathbb{R}^+}(\mathcal{U}) \to \mathrm{L}^2_{\mathbb{R}^+}(\mathcal{U})$ is said to be *causally extendible* if $\mathbf{T}_\tau \mathcal{D}_P = \mathbf{T}_\tau \mathrm{L}^2_{\mathbb{R}^+}(\mathcal{U})$ for all $\tau < \infty$. In this way, the input to P can be chosen arbitrarily over any finite interval $[0, \tau)$ and then continued into $\mathcal{D}_P$. Since P is causal, the corresponding output over $[0, \tau)$ is defined *uniquely* by the input up to time τ. Accordingly, when P is causally extendible there is a one-to-one correspondence between P and a system $P_e : \mathrm{L}^{2,\mathrm{e}}_{\mathbb{R}^+}(\mathcal{U}) \to \mathrm{L}^{2,\mathrm{e}}_{\mathbb{R}^+}(\mathcal{Y})$ such that for all $u_e \in \mathrm{L}^{2,\mathrm{e}}_{\mathbb{R}^+}(\mathcal{U})$ and

[1] According to this definition an LTI system is also LPTV, since such a system must, by definition, satisfy $\mathbf{U}_\tau \mathcal{G}_P \subset \mathcal{G}_P \ \forall\ \tau \in \mathbb{R}^+$.

$\tau < \infty$, $\mathbf{T}_\tau P_e u_e = \mathbf{T}_\tau P_e \mathbf{T}_\tau u_e := \mathbf{T}_\tau P u$ for any $u \in \mathcal{D}_P$ that satisfies $\mathbf{T}_\tau u_e = \mathbf{T}_\tau u$, where $\mathrm{L}^{2,\mathrm{e}}_{\mathbb{R}^+}(\mathcal{H}) := \{f : \mathbb{R}^+ \to \mathcal{H} \ : \ \mathbf{T}_\tau f \in \mathrm{L}^2_{\mathbb{R}^+}(\mathcal{H}) \text{ for all } \tau < \infty\}$ denotes the extended space associated with $\mathrm{L}^2_{\mathbb{R}^+}(\mathcal{H})$ for any Hilbert space $\mathcal{H}$. If P and P_e are related in this way, P_e is called the (causal) extension of P, and it is said that P is induced by P_e. When P is causally extendible, its causal extension P_e is said to be *locally Lipschitz-continuous* if for all $\tau \in \mathbb{R}^+$,

$$\sup_{\substack{u_1,u_2 \in \mathrm{L}^{2,\mathrm{e}}_{\mathbb{R}^+}(\mathcal{U}) \\ \mathbf{T}_\tau u_1 \neq \mathbf{T}_\tau u_2}} \frac{\|\mathbf{T}_\tau(P_e u_1 - P_e u_2)\|_{\mathrm{L}^2_{\mathbb{R}^+}(\mathcal{Y})}}{\|\mathbf{T}_\tau(u_1 - u_2)\|_{\mathrm{L}^2_{\mathbb{R}^+}(\mathcal{U})}} < \infty.$$

Furthermore, a causally extendible system P is said to be *strongly causal* if its causal extension P_e satisfies the following property: Given any $\tau \in \mathbb{R}^+$, $\epsilon > 0$ and $\tau_1 \leq \tau$ ($\tau_1 \in \mathbb{R}^+$), there exists a $\delta > 0$ such that for all $u_1, u_2 \in \mathrm{L}^{2,\mathrm{e}}_{\mathbb{R}^+}(\mathcal{U})$ with $\mathbf{T}_\tau u_1 = \mathbf{T}_\tau u_2$,

$$\|\mathbf{T}_{(\tau_1+\delta)}(P_e u_1 - P_e u_2)\|_{\mathrm{L}^2_{\mathbb{R}^+}(\mathcal{Y})} \leq \epsilon \|\mathbf{T}_{(\tau_1+\delta)}(u_1 - u_2)\|_{\mathrm{L}^2_{\mathbb{R}^+}(\mathcal{U})}.$$

Basically, a strongly causal system cannot respond instantaneously to inputs. Naturally, similar definitions to those above hold for systems defined on the discrete-time signal-space $\ell^2_{\mathbb{Z}^+}(\cdot)$.

DEFINITION 2.1. *Given two Hilbert spaces $\mathcal{U}$ and $\mathcal{Y}$, let $\mathscr{P}_{\mathcal{U},\mathcal{Y}}$ denote the set of causal LPTV (with period h, say) continuous-time systems*

$$P : \mathcal{D}_P \subset \mathrm{L}^2_{\mathbb{R}^+}(\mathcal{U}) \to \mathrm{L}^2_{\mathbb{R}^+}(\mathcal{Y})$$

with closed graphs.[2] *Furthermore, let $\mathscr{P}^e_{\mathcal{U},\mathcal{Y}} \subset \mathscr{P}_{\mathcal{U},\mathcal{Y}}$ denote the subset of causally extendible systems with locally Lipschitz-continuous extensions, and let $\mathscr{P}^{e,sc}_{\mathcal{U},\mathcal{Y}}$ denote the subset of systems that are also strongly causal.*

A continuous-time system $P \in \mathscr{P}_{\mathcal{U},\mathcal{Y}}$ is said to be *stable* if it is causal and $\mathcal{D}_P = \mathrm{L}^2_{\mathbb{R}^+}(\mathcal{U})$. Since it is assumed (by definition) that all systems in $\mathscr{P}_{\mathcal{U},\mathcal{Y}}$ have closed graphs, this implies that $\|P\| < \infty$ by the closed graph theorem [3, p. 80]. Note that any stable system $P \in \mathscr{P}_{\mathcal{U},\mathcal{Y}}$ is an element of $\mathscr{P}^e_{\mathcal{U},\mathcal{Y}}$.

Importantly, each system in $\mathscr{P}_{\mathcal{U},\mathcal{Y}}$ is equivalent (via the time-lifting isomorphism defined next) to a discrete-time system that is shift invariant [2]. Let $\mathbf{W} : \mathrm{L}^2_{\mathbb{R}^+}(\mathcal{H}) \to \ell^2_{\mathbb{Z}^+}(\mathrm{L}^2_{\mathbb{H}}(\mathcal{H}))$ denote the *time-lifting isomorphism* defined for each $f \in \mathrm{L}^2_{\mathbb{R}^+}(\mathcal{H})$ by

$$\overset{\circ}{\mathsf{f}}_k(\theta) := (\mathbf{W}f)_k = f(kh + \theta), \quad \theta \in \mathbb{H},$$

where $\mathrm{L}^2_{\mathbb{H}}(\mathcal{H}) := \mathbf{T}_h \mathrm{L}^2_{\mathbb{R}^+}(\mathcal{H})$. Then given any system $P \in \mathscr{P}_{\mathcal{U},\mathcal{Y}}$, the time-lifted equivalent, discrete-time system

$$\mathsf{P} := \mathbf{W} P \mathbf{W}^{-1} : \mathcal{D}_{\mathsf{P}} \subset \ell^2_{\mathbb{Z}^+}(\mathrm{L}^2_{\mathbb{H}}(\mathcal{U})) \to \ell^2_{\mathbb{Z}^+}(\mathrm{L}^2_{\mathbb{H}}(\mathcal{Y}))$$

is causal and linear shift-invariant (LSI) in the sense that its graph is a shift-invariant subspace, meaning $\mathbf{S}\mathcal{G}_{\mathsf{P}} \subset \mathcal{G}_{\mathsf{P}} = \mathbf{W}\mathcal{G}_P$, and $\mathbf{T}_k \mathcal{G}_{\mathsf{P}}$ corresponds to the graph of a linear operator for all $k \in \mathbb{Z}^+$ (implying causality), where $\mathbf{S}$ denotes the unilateral (forward) shift on $\ell^2_{\mathbb{Z}^+}(\cdot)$ and $\mathbf{T}_k$ denotes the truncation to zero after time k.[3] The key here is

[2] That a linear system has a closed graph is necessary for it to be stabilizable [12]. Hence, this is assumed.

[3] Throughout, the sans serif font (e.g., P) is used to distinguish objects associated with discrete-time signals and systems from continuous-time objects, for which italics are used (e.g., P).

that $\mathbf{S}^k\mathbf{W} = \mathbf{W}\mathbf{U}_{kh}$ for all $k \in \mathbb{Z}^+$, so that a shift by h in the continuous-time signal space $\mathrm{L}^2_{\mathbb{R}^+}(\cdot)$ corresponds to a shift by a single time-step in the isomorphic, discrete-time signal-space $\ell^2_{\mathbb{Z}^+}(\mathrm{L}^2_{\mathbb{H}}(\cdot))$. Note that by representing each signal in $\ell^2_{\mathbb{Z}^+}(\mathrm{L}^2_{\mathbb{H}}(\cdot))$ as a column vector with kth entry corresponding to the value of the signal at time k, the time-lifted equivalent system $\mathsf{P}(:=\mathbf{W}P\mathbf{W}^{-1})$ has lower-triangular, block-Toeplitz structure

$$\begin{bmatrix} \mathsf{P}_{[0]} & 0 & \cdots & \cdots & \cdots \\ \mathsf{P}_{[1]} & \mathsf{P}_{[0]} & 0 & \cdots & \cdots \\ \mathsf{P}_{[2]} & \mathsf{P}_{[1]} & \mathsf{P}_{[0]} & 0 & \cdots \\ \vdots & \ddots & \ddots & \ddots & \ddots \end{bmatrix},$$

which can be uniquely identified with the sequence $\{\mathsf{P}_{[i]}\}_{i=0}^{\infty}$. This representation plays a significant role in the sequel. If the original LPTV system P is causally extendible to a locally Lipschitz-continuous system, then each $\mathsf{P}_{[i]} \in \mathcal{B}_{\mathrm{L}^2_{\mathbb{H}}(\mathcal{U}),\mathrm{L}^2_{\mathbb{H}}(\mathcal{Y})}$.

REMARK 2.2. *Given a (causal) LSI system* $\mathsf{P} : \mathcal{D}_P \subset \ell^2_{\mathbb{Z}^+}(\mathrm{L}^2_{\mathbb{H}}(\mathcal{U})) \to \ell^2_{\mathbb{Z}^+}(\mathrm{L}^2_{\mathbb{H}}(\mathcal{Y}))$,[4] *the equivalent continuous-time system* $P:=\mathbf{W}^{-1}\mathsf{P}\mathbf{W}$ *may not be locally causal in each* $[kh,(k+1)h)$ *interval of time. The equivalent system* P *is causal if and only if* $\mathsf{P}_{[0]}$ *(the first element of the sequence uniquely identifiable with the block-Toeplitz representation of* P*) is a causal mapping from* $\mathrm{L}^2_{\mathbb{H}}(\mathcal{U})$ *to* $\mathrm{L}^2_{\mathbb{H}}(\mathcal{Y})$*, in the sense that* $\mathbf{T}_\tau\mathcal{G}_{\mathsf{P}_{[0]}}$ *corresponds to the graph of a linear operator for all* $\tau \in \mathbb{H}$.

A (causal) LSI system $\mathsf{P} : \mathcal{D}_{\mathsf{P}} \subset \ell^2_{\mathbb{Z}^+}(\mathcal{U}) \to \ell^2_{\mathbb{Z}^+}(\mathcal{Y})$ is called stable if $\mathcal{D}_{\mathsf{P}} = \ell^2_{\mathbb{Z}^+}(\mathcal{U})$. If in addition $\mathcal{G}_{\mathsf{P}}$ is closed, then $\|\mathsf{P}\| < \infty$ by the closed-graph theorem. Related to such operators is the Hardy space $\mathrm{H}^\infty_{\mathbb{D}}(\mathcal{B}_{\mathcal{U},\mathcal{Y}})$ of functions $\Phi : \mathbb{D} \to \mathcal{B}_{\mathcal{U},\mathcal{Y}}$ that are bounded and analytic in the open unit-disc. The norm of a function $\Phi \in \mathrm{H}^\infty_{\mathbb{D}}(\mathcal{B}_{\mathcal{U},\mathcal{Y}})$ is defined as

$$\|\Phi\|_\infty := \sup_{\lambda\in\mathbb{D}} \|\Phi(\lambda)\|.$$

Given a function $\Phi \in \mathrm{H}^\infty_{\mathbb{D}}(\mathcal{B}_{\mathcal{U},\mathcal{Y}})$, boundary values can be defined almost everywhere on the unit-circle $\mathbb{T}$. The resulting boundary values $\Phi(e^{j\omega})$ are essentially bounded on $\mathbb{T}$, and $\|\Phi\|_\infty = \operatorname{ess\,sup}_{\omega\in[0,2\pi)}\|\Phi(e^{j\omega})\|$. A function $\Phi \in \mathrm{H}^\infty_{\mathbb{D}}(\mathcal{B}_{\mathcal{U},\mathcal{Y}})$ is called inner if $\Phi(e^{j\omega})$ is an isometry almost everywhere on $\mathbb{T}$. Corresponding to each $\Phi \in \mathrm{H}^\infty_{\mathbb{D}}(\mathcal{B}_{\mathcal{U},\mathcal{Y}})$ is a multiplication operator $\mathbf{M}_\Phi : \mathrm{H}^2_{\mathbb{D}}(\mathcal{U}) \to \mathrm{H}^2_{\mathbb{D}}(\mathcal{Y})$, defined by $(\mathbf{M}_\Phi\varphi)(\lambda):=\Phi(\lambda)\varphi(\lambda)$ for all $\varphi \in \mathrm{H}^2_{\mathbb{D}}(\mathcal{U})$ and $\lambda \in \mathbb{D}$.

PROPOSITION 2.3 (see [11], p. 235). *Let* $\mathcal{U}$ *and* $\mathcal{Y}$ *be arbitrary Hilbert spaces. Then given a stable, LSI system* $\mathsf{P} : \ell^2_{\mathbb{Z}^+}(\mathcal{U}) \to \ell^2_{\mathbb{Z}^+}(\mathcal{Y})$ *with closed graph, there exists a function* $\hat{\mathsf{P}} \in \mathrm{H}^\infty_{\mathbb{D}}(\mathcal{B}_{\mathcal{U},\mathcal{Y}})$ *such that* $\mathsf{P} = \mathbf{Z}^{-1}\mathbf{M}_{\hat{\mathsf{P}}}\mathbf{Z}$*. Furthermore,* P *is an isometry if and only if the corresponding symbol* $\hat{\mathsf{P}}$ *is inner. Moreover, given any* $\Phi \in \mathrm{H}^\infty_{\mathbb{D}}(\mathcal{B}_{\mathcal{U},\mathcal{Y}})$*, the multiplication operator* $\mathbf{M}_\Phi : \mathrm{H}^2_{\mathbb{D}}(\mathcal{U}) \to \mathrm{H}^2_{\mathbb{D}}(\mathcal{Y})$ *is bounded and LSI*[5] *with* $\|\mathbf{M}_\Phi\| = \|\Phi\|_\infty$.

2.2. Feedback systems. Consider the closed-loop configuration shown in Figure 1.1, and suppose that $P \in \mathscr{P}^e_{\mathcal{U},\mathcal{Y}}$ and $C \in \mathscr{P}^e_{\mathcal{Y},\mathcal{U}}$. Denote by P_e and C_e the respective locally Lipschitz-continuous causal extensions, and for notational convenience, let $\mathcal{V}:=\mathcal{U}\oplus\mathcal{Y}$. The closed loop, denoted by $[P,C]$, is said to be well-posed if the following three conditions hold.

[4]Since this system is defined only for positive time, shift-invariance implies causality.

[5]The unilateral shift $\mathbf{S}$ on $\mathrm{H}^2_{\mathbb{D}}(\cdot)$ corresponds to multiplication by λ.

(i) $\left[\begin{smallmatrix} I & C_e \\ P_e & I \end{smallmatrix}\right] : \mathrm{L}^{2,\mathrm{e}}_{\mathbb{R}^+}(\mathcal{V}) \to \mathrm{L}^{2,\mathrm{e}}_{\mathbb{R}^+}(\mathcal{V})$, the system mapping $\left[\begin{smallmatrix} u_1 \\ u_2 \end{smallmatrix}\right]$ to $\left[\begin{smallmatrix} d_1 \\ d_2 \end{smallmatrix}\right]$, is bijective, implying the existence of a *unique* solution to the functional equations that describe the closed loop.

(ii) $H_e(P,C) := \left[\begin{smallmatrix} I & C_e \\ P_e & I \end{smallmatrix}\right]^{-1}$ is causal and locally Lipschitz-continuous.

(iii) The solution to the functional equations that described the closed loop is insensitive to very high frequency modeling errors, such as small transmission delays.

A sufficient condition for well-posedness is that P or C is strongly causal.[6] For further details and a discussion of the physical significance of well-posedness, the reader is referred to the seminal work of Willems [28, Chap. 4]. In addition to well-posedness, it is desirable for the system $H(P,C) : \mathcal{D}_{H(P,C)} \subset \mathrm{L}^2_{\mathbb{R}^+}(\mathcal{V}) \to \mathrm{L}^2_{\mathbb{R}^+}(\mathcal{V})$ induced by $H_e(P,C)$ to satisfy $\mathcal{D}_{H(P,C)} = \mathrm{L}^2_{\mathbb{R}^+}(\mathcal{V})$. In this case, the closed loop is said to be stable and P is said to be stabilized by C. Since P and C are both linear it can be shown that this also implies closed loop stability with finite gain [28, p. 117]; that is, $\|H(P,C)\| < \infty$.

Given a plant $P \in \mathscr{P}^e_{\mathcal{U},\mathcal{Y}}$ and a controller $C \in \mathscr{P}^e_{\mathcal{Y},\mathcal{U}}$, note that

$$\begin{bmatrix} I & C \\ P & I \end{bmatrix} : \mathcal{D}_P \oplus \mathcal{D}_C \to \mathcal{G}_P + \mathcal{G}_C^\sharp.$$

Thus, if $[P,C]$ is well-posed, then $\mathcal{G}_P \cap \mathcal{G}_C^\sharp = \{0\}$. Furthermore, observe that $\mathcal{D}_{H(P,C)} = \mathcal{G}_P + \mathcal{G}_C^\sharp$ and hence that the closed loop is stable if and only if $\mathcal{G}_P + \mathcal{G}_C^\sharp = \mathrm{L}^2_{\mathbb{R}^+}(\mathcal{V})$ (with $\mathcal{G}_P \cap \mathcal{G}_C^\sharp = \{0\}$). This useful geometric characterization of closed-loop stability is summarized in the following proposition, which has also appeared in [12, 15, 19] for other classes of systems.

PROPOSITION 2.4. *Given a well-posed plant/controller pair $P \in \mathscr{P}^e_{\mathcal{U},\mathcal{Y}}$ and $C \in \mathscr{P}^e_{\mathcal{Y},\mathcal{U}}$, the closed-loop $[P,C]$ is stable if and only if the graph of the plant and the inverse graph of the controller induce a coordinatization of $\mathrm{L}^2_{\mathbb{R}^+}(\mathcal{V})$, that is, if and only if $\mathcal{G}_P \cap \mathcal{G}_C^\sharp = \{0\}$ and $\mathcal{G}_P + \mathcal{G}_C^\sharp = \mathrm{L}^2_{\mathbb{R}^+}(\mathcal{V})$. In this case,the parallel projections $\mathbf{\Pi}_{\mathcal{G}_P \| \mathcal{G}_C^\sharp}$ and $\mathbf{\Pi}_{\mathcal{G}_C^\sharp \| \mathcal{G}_P}$ are stable systems in $\mathscr{P}^e_{\mathcal{V},\mathcal{V}}$. In fact,*

$$\mathbf{\Pi}_{\mathcal{G}_P \| \mathcal{G}_C^\sharp} = \begin{bmatrix} I & 0 \\ 0 & -I \end{bmatrix} H(P,C) + \begin{bmatrix} 0 & 0 \\ 0 & I \end{bmatrix}$$

and

$$\mathbf{\Pi}_{\mathcal{G}_C^\sharp \| \mathcal{G}_P} = \begin{bmatrix} -I & 0 \\ 0 & I \end{bmatrix} H(P,C) + \begin{bmatrix} I & 0 \\ 0 & 0 \end{bmatrix}.$$

2.3. Representations of the graph. The following results, concerning the existence of particular representations of the graph, are fundamental to all of the results presented in the sequel. These results can be found in [5], and accordingly they are stated here without proof.

THEOREM 2.5. *If $P \in \mathscr{P}^e_{\mathcal{U},\mathcal{Y}}$ can be stabilized by some $C \in \mathscr{P}^e_{\mathcal{Y},\mathcal{U}}$, then there exist stable systems $G_r \in \mathscr{P}^e_{\mathcal{U},\mathcal{V}}$, $G_l \in \mathscr{P}^e_{\mathcal{V},\mathcal{Y}}$, $K_l \in \mathscr{P}^e_{\mathcal{V},\mathcal{U}}$, and $K_r \in \mathscr{P}^e_{\mathcal{Y},\mathcal{V}}$ such that*

$$\mathcal{G}_P = \mathcal{R}_{G_r} = \mathcal{K}_{G_l}$$

[6] This is not the weakest such condition. For example, that the product of the instantaneous gains is strictly less than 1 is also sufficient [28, Chap. 4].

and

$$\begin{bmatrix} G_r & K_r \end{bmatrix} \begin{bmatrix} K_l \\ G_l \end{bmatrix} = \begin{bmatrix} K_l \\ G_l \end{bmatrix} \begin{bmatrix} G_r & K_r \end{bmatrix} = \begin{bmatrix} I & 0 \\ 0 & I \end{bmatrix}.$$

G_r and G_l in Theorem 2.5 are called *strong-right* and *strong-left* representations of $\mathcal{G}_P$, respectively. Note that they are unique only up to invertible factors. That is, for any stable systems $Q_1(Q_1^{-1}) \in \mathscr{P}^e_{\mathcal{U},\mathcal{U}}$, and $Q_2(Q_2^{-1}) \in \mathscr{P}^e_{\mathcal{Y},\mathcal{Y}}$, the systems G_rQ_1 and $Q_2^{-1}G_l$ are also strong-right and strong-left representations of $\mathcal{G}_P$, respectively.

LEMMA 2.6. *Given a well-posed plant/controller pair $P \in \mathscr{P}^e_{\mathcal{U},\mathcal{Y}}$ and $C \in \mathscr{P}^e_{\mathcal{Y},\mathcal{U}}$, let $G_r \in \mathscr{P}^e_{\mathcal{U},\mathcal{V}}$ be any stable system with zero kernel such that $\mathcal{G}_P = \mathcal{R}_{G_r}$, and let $K_l \in \mathscr{P}^e_{\mathcal{V},\mathcal{U}}$ be any stable system such that $\mathcal{R}_{K_l} = \mathrm{L}^2_{\mathbb{R}^+}(\mathcal{U})$ and $\mathcal{G}^\sharp_C = \mathcal{K}_{K_l}$. Then the following are equivalent.*

(i) *The closed-loop $[P, C]$ is stable.*

(ii) *The stable, linear system K_lG_r is bijective (and hence boundedly invertible, although the inverse may not be causal).*

Furthermore, if $[P, C]$ is stable and G_r and K_l are, respectively, strong-right and strong-left representations of $\mathcal{G}_P$ and $\mathcal{G}^\sharp_C$, then $(K_lG_r)^{-1}$ is a stable system in $\mathscr{P}^e_{\mathcal{U},\mathcal{U}}$.

A similar result to Lemma 2.6 holds with the right representation of $\mathcal{G}_P$ replaced by a left representation, and similarly for the inverse graph of the controller.

REMARK 2.7. *Note that with P, C, G_r, and K_l defined as in the first part of Lemma* 2.6*, the following useful identity holds:*

$$\mathbf{\Pi}_{\mathcal{G}_P \| \mathcal{G}^\sharp_C} = G_r(K_lG_r)^{-1}K_l.$$

Now, let

$$\mathsf{P} := \mathbf{W}P\mathbf{W}^{-1}, \quad \mathsf{C} := \mathbf{W}C\mathbf{W}^{-1}, \quad \mathsf{G}_r := \mathbf{W}G_r\mathbf{W}^{-1}, \quad \textit{and} \quad \mathsf{K}_l := \mathbf{W}K_l\mathbf{W}^{-1}$$

so that $\mathcal{G}_{\mathsf{P}} = \mathcal{R}_{\mathsf{G}_r}$, $\mathcal{R}_{\mathsf{K}_l} = \ell^2_{\mathbb{Z}^+}(\mathrm{L}^2_{\mathbb{H}}(\mathcal{U}))$, and $\mathcal{G}^\sharp_{\mathsf{C}} = \mathcal{K}_{\mathsf{K}_l}$. Then the closed-loop $[P, C]$ is stable if and only if the bounded linear operator $\mathsf{K}_l\mathsf{G}_r$ is boundedly invertible. In this case, $(\mathsf{K}_l\mathsf{G}_r)^{-1}$ is a stable LSI system and hence equivalent (via $\mathbf{Z}$) to a multiplication operator with symbol in $\mathrm{H}^\infty_{\mathbb{D}}(\mathcal{B}_{\mathrm{L}^2_{\mathbb{H}}(\mathcal{U}),\mathrm{L}^2_{\mathbb{H}}(\mathcal{U})})$ (cf. Proposition 2.3*). Moreover, $\mathbf{\Pi}_{\mathcal{G}_{\mathsf{P}} \| \mathcal{G}^\sharp_{\mathsf{C}}} = \mathsf{G}_r(\mathsf{K}_l\mathsf{G}_r)^{-1}\mathsf{K}_l$.*

3. The gap metric. In this section, the gap metric is formally introduced. The main result of the section is a formula for the directed gap between LPTV systems. The formula is essentially a generalization of Georgiou's result for LTI systems [13], and it facilitates the use of function theoretic arguments in a key step of the proof of the main robustness result in the next section.

The gap (or aperture) between two (closed) subspaces $\mathcal{H}_0$ and $\mathcal{H}_1$ of a Hilbert space $\mathcal{H}$ is defined by (see [16, pp. 197–200], for example)

$$\delta(\mathcal{H}_0, \mathcal{H}_1) := \|\mathbf{\Pi}_{\mathcal{H}_0} - \mathbf{\Pi}_{\mathcal{H}_1}\|.$$

From this it follows that $\delta(\cdot, \cdot)$ is a metric. It can be shown, as in [17, Sect. 15.3], that

$$\delta(\mathcal{H}_0, \mathcal{H}_1) = \max\{\vec{\delta}(\mathcal{H}_0, \mathcal{H}_1), \vec{\delta}(\mathcal{H}_1, \mathcal{H}_0)\},$$

where

$$\vec{\delta}(\mathcal{H}_i, \mathcal{H}_j) = \|\mathbf{\Pi}_{\mathcal{H}_j^\perp}\mathbf{\Pi}_{\mathcal{H}_i}\| \quad (i, j = 0, 1)$$

is called the directed gap. Note hence that $0 \le \delta(\mathcal{H}_0, \mathcal{H}_1) \le 1$.

PROPOSITION 3.1 (see [17]). *For $i = 0, 1$ let $\mathcal{H}_i$ be a closed subspace of a Hilbert space $\mathcal{H}$. Then $\mathbf{\Pi}_{\mathcal{H}_1}$ is a bijective mapping from $\mathcal{H}_0$ to $\mathcal{H}_1$ if and only if $\delta(\mathcal{H}_0, \mathcal{H}_1) < 1$. Moreover, if $\delta(\mathcal{H}_0, \mathcal{H}_1) < 1$, then*

$$\delta(\mathcal{H}_0, \mathcal{H}_1) = \vec{\delta}(\mathcal{H}_0, \mathcal{H}_1) = \vec{\delta}(\mathcal{H}_1, \mathcal{H}_0).$$

The gap between two (closed) operators P_0 and P_1 is defined to be the gap between their graphs as follows:

$$\delta_g(P_0, P_1) := \delta(\mathcal{G}_{P_0}, \mathcal{G}_{P_1}),$$

with the directed gap being defined similarly [16, pp. 197–200]. Measuring the distance between systems in this way was first introduced into the control literature in [31] and considered further in the work of several others, notably [13, 14, 20, 12, 10].

Consider a system $P_0 \in \mathscr{P}_{\mathcal{U},\mathcal{Y}}$ and recall that $\mathcal{G}_{P_0}$ is isomorphic to both $\mathcal{G}_{\mathsf{P}_0}$ and $\hat{\mathcal{G}}_{\mathsf{P}_0} := \mathbf{Z}\mathcal{G}_{\mathsf{P}_0}$, where $\mathsf{P}_0 := \mathbf{W}P_0\mathbf{W}^{-1}$. Now given another system $P_1 \in \mathscr{P}_{\mathcal{U},\mathcal{Y}}$,

$$\delta_g(P_0, P_1) = \delta(\mathcal{G}_{P_0}, \mathcal{G}_{P_1}) = \delta(\hat{\mathcal{G}}_{\mathsf{P}_0}, \hat{\mathcal{G}}_{\mathsf{P}_1})$$

holds for the gap and

$$\vec{\delta}_g(P_i, P_j) = \vec{\delta}(\mathcal{G}_{P_i}, \mathcal{G}_{P_j}) = \vec{\delta}(\hat{\mathcal{G}}_{\mathsf{P}_i}, \hat{\mathcal{G}}_{\mathsf{P}_j}) \quad (i, j = 0, 1)$$

holds for the directed gap. The next corollary of Proposition 3.1 constitutes a useful characterization of when the gap between two systems is strictly less than 1 (see [20] and [27] for the LTI analogue).

COROLLARY 3.2. *Given two systems P_0 and P_1 in $\mathscr{P}_{\mathcal{U},\mathcal{Y}}$ for $i = 0, 1$ let $\hat{\mathsf{G}}_{ri}$ be any inner function in $\mathrm{H}^\infty_{\mathbb{D}}(\mathcal{B}_{\mathrm{L}^2_{\mathbb{H}}(\mathcal{U}),\mathrm{L}^2_{\mathbb{H}}(\mathcal{Y})})$ such that $\hat{\mathcal{G}}_{\mathsf{P}_i} := \mathbf{ZW}\mathcal{G}_{P_i} = \hat{\mathsf{G}}_{ri}\mathrm{H}^2_{\mathbb{D}}(\mathrm{L}^2_{\mathbb{H}}(\mathcal{U}))$.[7] Then the following are equivalent.*

(i) $\delta_g(P_0, P_1) = \delta(\hat{\mathcal{G}}_{\mathsf{P}_0}, \hat{\mathcal{G}}_{\mathsf{P}_1}) < 1$.

(ii) *The bounded linear operator $(\mathbf{M}_{\hat{\mathsf{G}}_{r1}})^*\mathbf{M}_{\hat{\mathsf{G}}_{r0}}$ is bijective (and hence boundedly invertible).*

Proof. Since $\mathbf{\Pi}_{\hat{\mathcal{G}}_{\mathsf{P}_i}} = \mathbf{M}_{\hat{\mathsf{G}}_{ri}}(\mathbf{M}_{\hat{\mathsf{G}}_{ri}})^*$, $\hat{\mathcal{G}}_{\mathsf{P}_i} = \mathcal{R}_{\mathbf{M}_{\hat{\mathsf{G}}_{ri}}}$, and $\mathcal{K}_{\mathbf{M}_{\hat{\mathsf{G}}_{ri}}} = \{0\}$, it follows by Proposition 3.1 that $\delta(\hat{\mathcal{G}}_{\mathsf{P}_0}, \hat{\mathcal{G}}_{\mathsf{P}_1}) < 1$ if and only if $\mathbf{M}_{\hat{\mathsf{G}}_{r1}}(\mathbf{M}_{\hat{\mathsf{G}}_{r1}})^*\mathbf{M}_{\hat{\mathsf{G}}_{r0}}$ is a bijective mapping from $\mathrm{H}^2_{\mathbb{D}}(\mathrm{L}^2_{\mathbb{H}}(\mathcal{U}))$ to $\hat{\mathcal{G}}_{\mathsf{P}_1}$ (see also [18, p. 201] for a similar result). Using again the fact that $\hat{\mathcal{G}}_{\mathsf{P}_i} = \mathcal{R}_{\mathbf{M}_{\hat{\mathsf{G}}_{ri}}}$ and $\mathcal{K}_{\mathbf{M}_{\hat{\mathsf{G}}_{ri}}} = \{0\}$, this is equivalent to

$$\mathcal{R}_{(\mathbf{M}_{\hat{\mathsf{G}}_{r1}})^*\mathbf{M}_{\hat{\mathsf{G}}_{r0}}} = \mathrm{H}^2_{\mathbb{D}}(\mathrm{L}^2_{\mathbb{H}}(\mathcal{U})) \text{ and } \mathcal{K}_{(\mathbf{M}_{\hat{\mathsf{G}}_{r1}})^*\mathbf{M}_{\hat{\mathsf{G}}_{r0}}} = \{0\}.$$

That is, $\delta(\hat{\mathcal{G}}_{\mathsf{P}_0}, \hat{\mathcal{G}}_{\mathsf{P}_1}) < 1$ if and only if $(\mathbf{M}_{\hat{\mathsf{G}}_{r1}})^*\mathbf{M}_{\hat{\mathsf{G}}_{r0}}$ is bijective (and hence boundedly invertible by the open mapping theorem [3, p. 79]). ∎

Importantly, it is possible to express the directed gap between two LPTV systems in terms of a two-block optimization problem over $\mathrm{H}^\infty_{\mathbb{D}}$ and a specific representation of the corresponding graphs. This formula is essentially a generalization of Georgiou's [13] for the directed gap between two finite-dimensional LTI systems (see also [29, Thm. 1] and [20, Cor. 4.4]). The resultant formula is fundamental to the robust

[7] Such $\hat{\mathsf{G}}_{ri}$ exist by the Beurling–Lax–Halmos theorem. See [5] for further details.

stability results derived in the next section, and it gives rise to an algorithm for computing the gap between LPTV systems [4, Chap. 5].

THEOREM 3.3. *Given two systems P_0 and P_1 in $\mathscr{P}_{\mathcal{U},\mathcal{Y}}$ for $i = 0, 1$ let $\hat{\mathsf{G}}_{ri}$ be an inner function in $\mathrm{H}^\infty_{\mathbb{D}}(\mathcal{B}_{\mathrm{L}^2_{\mathbb{H}}(\mathcal{U}),\mathrm{L}^2_{\mathbb{H}}(\mathcal{V})})$ such that $\hat{\mathcal{G}}_{\mathsf{P}_i} = \mathbf{ZW}\mathcal{G}_{P_i} = \hat{\mathsf{G}}_{ri}\mathrm{H}^2_{\mathbb{D}}(\mathrm{L}^2_{\mathbb{H}}(\mathcal{U}))$. Then*

$$\vec{\delta}_g(P_0, P_1) := \vec{\delta}(\hat{\mathcal{G}}_{\mathsf{P}_0}, \hat{\mathcal{G}}_{\mathsf{P}_1}) = \inf_{\hat{\mathsf{Q}} \in \mathrm{H}^\infty_{\mathbb{D}}(\mathcal{B}_{\mathrm{L}^2_{\mathbb{H}}(\mathcal{U}),\mathrm{L}^2_{\mathbb{H}}(\mathcal{U})})} \|\hat{\mathsf{G}}_{r0} - \hat{\mathsf{G}}_{r1}\hat{\mathsf{Q}}\|_\infty.$$

Proof. The proof follows that of an H^∞ optimization result in [11, p. 248], using the commutant lifting theorem. For the full proof see Appendix A. □

4. Quantitative robustness analysis. An important property of feedback systems is that closed-loop stability can be desensitized to reasonable perturbations of the plant and the controller. In the subsections to follow, this property is quantitatively analyzed in terms of the gap metric for LPTV, feedback systems that evolve in continuous-time. First gap-metric perturbations to the plant alone are considered. Then robustness in the face of simultaneous gap-metric perturbations to both the plant and the controller is investigated.

4.1. Plant perturbations. In this subsection, the robustness of closed-loop stability is studied within the context of plant perturbations measured in the gap metric. Before going on, however, it is convenient to make the following definition.

DEFINITION 4.1. *Given a stable closed-loop $[P, C]$ with $P \in \mathscr{P}^e_{\mathcal{U},\mathcal{Y}}$ and $C \in \mathscr{P}^e_{\mathcal{Y},\mathcal{U}}$, define*

$$b^{-1}_{P,C} := \|\mathbf{\Pi}_{\mathcal{G}_P \| \mathcal{G}^\sharp_C}\| = \|\mathbf{\Pi}_{\mathcal{G}_\mathsf{P} \| \mathcal{G}^\sharp_\mathsf{C}}\|. \tag{4.1}$$

For reasons that will become apparent shortly, $b_{P,C}$ is called the robust-stability margin.

The robust stability result derived next characterizes the largest gap-ball of LPTV plants centered at a nominal plant that a nominal stabilizing LPTV controller is guaranteed to stabilize. To a large extent, the result is analogous to the LTI result of [14] and the general LTV result of [12]. Note, however, that it does not follow directly from either of these. In particular, considerable effort is required to prove necessity, where the approach taken is to construct a *causal* plant P_1 of appropriate gap distance from the nominal plant so that $[P_1, C]$ is well-posed but not stable.

THEOREM 4.2. *Let $[P_0, C]$ be a stable, closed-loop system, where $P_0 \in \mathscr{P}^e_{\mathcal{U},\mathcal{Y}}$ and $C \in \mathscr{P}^{e,sc}_{\mathcal{Y},\mathcal{U}}$. Then, the following are equivalent.*

(i) $b_{P_0,C} \geq \beta$.

(ii) *$[P_1, C]$ is stable for all $P_1 \in \mathscr{P}^e_{\mathcal{U},\mathcal{Y}}$ that satisfy $\delta_g(P_0, P_1) < \beta$.*

Proof of (i) $\Rightarrow$ (ii). Assume that (i) holds so that $b^{-1}_{P_0,C} \leq \beta^{-1}$. Now consider any system $P_1 \in \mathscr{P}^e_{\mathcal{U},\mathcal{Y}}$ such that $\delta_g(P_0, P_1) := \delta(\mathcal{G}_{\mathsf{P}_0}, \mathcal{G}_{\mathsf{P}_1}) < \beta$, where $\mathsf{P}_i = \mathbf{W}P_i\mathbf{W}^{-1}$ for $i = 0, 1$. Since P_i is LPTV, P_i is LSI and, correspondingly, $\mathcal{G}_{\mathsf{P}_i}$ is a shift-invariant subspace of $\ell^2_{\mathbb{Z}^+}(\mathrm{L}^2_{\mathbb{H}}(\mathcal{V}))$.[8] Thus, it follows by the Beurling–Lax–Halmos theorem (see [11, p. 239]) that for each $i = 0, 1$, there exists an LSI isometry $\mathsf{G}_{ri} : \ell^2_{\mathbb{Z}^+}(\mathrm{L}^2_{\mathbb{H}}(\mathcal{U})) \to \ell^2_{\mathbb{Z}^+}(\mathrm{L}^2_{\mathbb{H}}(\mathcal{V}))$ such that $\mathcal{G}_{\mathsf{P}_i} = \mathcal{R}_{\mathsf{G}_{ri}}$.[9] As G_{ri} is an isometry, $\mathbf{\Pi}_{\mathcal{G}_{\mathsf{P}_i}} = \mathsf{G}_{ri}(\mathsf{G}_{ri})^*$ and hence

[8] Recall that for notational convenience $\mathcal{V} := \mathcal{U} \oplus \mathcal{Y}$.

[9] For a discussion on the existence of such representations of the graph, see [5].

$$\delta(\mathcal{G}_{\mathsf{P}_0}, \mathcal{G}_{\mathsf{P}_1}) = \|\mathbf{\Pi}_{\mathcal{G}_{\mathsf{P}_0}} - \mathbf{\Pi}_{\mathcal{G}_{\mathsf{P}_1}}\| = \|\mathsf{G}_{r0}\mathsf{G}_{r0}^* - \mathsf{G}_{r1}\mathsf{G}_{r1}^*\| < \beta. \tag{4.2}$$

Now since $[P_0, C]$ is stable, it follows by Theorem 2.5 that there exists a stable LSI system $\mathsf{K}_l : \ell^2_{\mathbb{Z}^+}(\mathrm{L}^2_{\mathbb{H}}(\mathcal{Y})) \to \ell^2_{\mathbb{Z}^+}(\mathrm{L}^2_{\mathbb{H}}(\mathcal{U}))$ satisfying $\mathcal{K}_{\mathsf{K}_l} = \mathcal{G}^\sharp_{\mathsf{C}}$ and $\mathcal{R}_{\mathsf{K}_l} = \ell^2_{\mathbb{Z}^+}(\mathcal{U})$. Furthermore, $\mathbf{\Pi}_{\mathcal{G}_{\mathsf{P}_0}\|\mathcal{G}^\sharp_{\mathsf{C}}} = \mathsf{G}_{r0}(\mathsf{K}_l\mathsf{G}_{r0})^{-1}\mathsf{K}_l$ (see Remark 2.7). Correspondingly,

$$b^{-1}_{P_0,C} = \|(\mathsf{K}_l\mathsf{G}_{r0})^{-1}\mathsf{K}_l\| \le \beta^{-1}, \tag{4.3}$$

since G_{r0} is an isometry. Combining equations (4.2) and (4.3) yields

$$\|(\mathsf{K}_l\mathsf{G}_{r0})^{-1}\mathsf{K}_l(\mathsf{G}_{r0}\mathsf{G}_{r0}^* - \mathsf{G}_{r1}\mathsf{G}_{r1}^*)\| < 1.$$

This implies that

$$\|\mathrm{I} - (\mathsf{K}_l\mathsf{G}_{r0})^{-1}\mathsf{K}_l\mathsf{G}_{r1}\mathsf{G}_{r1}^*\mathsf{G}_{r0}\| \le \|\mathsf{G}_{r0}^* - (\mathsf{K}_l\mathsf{G}_{r0})^{-1}\mathsf{K}_l\mathsf{G}_{r1}\mathsf{G}_{r1}^*\| \cdot \|\mathsf{G}_{r0}\| < 1$$

and hence that $(\mathsf{K}_l\mathsf{G}_{r0})^{-1}\mathsf{K}_l\mathsf{G}_{r1}(\mathsf{G}_{r1}^*\mathsf{G}_{r0})$ has bounded inverse. Since $\delta_g(P_0, P_1) < 1$, the operator $\mathsf{G}_{r1}^*\mathsf{G}_{r0}$ has bounded inverse (see Corollary 3.2) and, correspondingly, it follows that $\mathsf{K}_l\mathsf{G}_{r1}$ is also boundedly invertible. By Lemma 2.6 and Remark 2.7, this implies stability of the closed-loop $[P_1, C]$, completing the proof of (i) $\Rightarrow$ (ii).

Proof of (i) $\Leftarrow$ (ii). Suppose that (ii) holds for some $1 \ge \beta > b_{P_0,C}$.[10] Under this hypothesis it is sufficient to construct a system $P_1 \in \mathscr{P}^e_{\mathcal{U},\mathcal{Y}}$ such that $\delta_g(P_0, P_1) < \beta$ and $[P_1, C]$ is well-posed but not stable. Now let G_{r0} and K_l be defined as in the proof of (i) $\Rightarrow$ (ii). Then by Proposition 2.3, G_{r0} and K_l are equivalent (via $\mathbf{Z}$) to multiplication operators with symbols $\hat{\mathsf{G}}_{r0} \in \mathrm{H}^\infty_{\mathbb{D}}(\mathcal{B}_{\mathrm{L}^2_{\mathbb{H}}(\mathcal{U}),\mathrm{L}^2_{\mathbb{H}}(\mathcal{Y})})$ and $\hat{\mathsf{K}}_l \in \mathrm{H}^\infty_{\mathbb{D}}(\mathcal{B}_{\mathrm{L}^2_{\mathbb{H}}(\mathcal{Y}),\mathrm{L}^2_{\mathbb{H}}(\mathcal{U})})$, respectively. Furthermore, $(\hat{\mathsf{K}}_l\hat{\mathsf{G}}_{r0})^{-1} \in \mathrm{H}^\infty_{\mathbb{D}}(\mathcal{B}_{\mathrm{L}^2_{\mathbb{H}}(\mathcal{U}),\mathrm{L}^2_{\mathbb{H}}(\mathcal{U})})$ (see Lemma 2.6). Hence,

$$\hat{\mathsf{K}}_{l\star} := (\hat{\mathsf{G}}_{r0}\hat{\mathsf{K}}_l)^{-1}\hat{\mathsf{K}}_l \in \mathrm{H}^\infty_{\mathbb{D}}\left(\mathcal{B}_{\mathrm{L}^2_{\mathbb{H}}(\mathcal{Y}),\mathrm{L}^2_{\mathbb{H}}(\mathcal{U})}\right)$$

and $\mathsf{K}_{l\star} := \mathbf{Z}^{-1}\mathbf{M}_{\hat{\mathsf{K}}_{l\star}}\mathbf{Z} : \ell^2_{\mathbb{Z}^+}(\mathrm{L}^2_{\mathbb{H}}(\mathcal{Y})) \to \ell^2_{\mathbb{Z}^+}(\mathrm{L}^2_{\mathbb{H}}(\mathcal{U}))$ is a stable LSI system satisfying $\mathcal{K}_{\mathsf{K}_{l\star}} = \mathcal{G}^\sharp_{\mathsf{C}}$ and $\mathcal{R}_{\mathsf{K}_{l\star}} = \ell^2_{\mathbb{Z}^+}(\mathrm{L}^2_{\mathbb{H}}(\mathcal{U}))$. In fact, by definition,

$$b^{-1}_{P_0,C} := \|\mathsf{G}_{r0}\mathsf{K}_{l\star}\| = \|\mathsf{K}_{l\star}\| = \|\hat{\mathsf{K}}_{l\star}\|_\infty = \sup_{\lambda\in\mathbb{D}} \|\hat{\mathsf{K}}_{l\star}(\lambda)\|, \tag{4.4}$$

where the second equality holds because G_{r0} is an isometry. Since the function $\hat{\mathsf{K}}_{l\star}(\lambda)$ is analytic in $\mathbb{D}$, $\|\hat{\mathsf{K}}_{l\star}(\lambda)\|$ satisfies a maximum principle on any connected, open subset of $\mathbb{D}$ (see [1]). Using this property and (4.4), it follows that for any $\epsilon_0 > 0$ and $\epsilon_1 > 0$, there exists a $\lambda_{0,\epsilon_0} \in \mathbb{A}_{\epsilon_0} := \{\lambda : (1-\epsilon_0) \le |\lambda| < 1\}$ such that

$$(b_{P_0,C} + \epsilon_1)^{-1} \le \|\hat{\mathsf{K}}_{l\star}(\lambda_{0,\epsilon_0})\| \le b^{-1}_{P_0,C}.$$

Now for any $\epsilon_2 > \epsilon_1$ it is possible to construct an operator $\Delta_{0,\epsilon_2} \in \mathcal{B}_{\mathrm{L}^2_{\mathbb{H}}(\mathcal{U}),\mathrm{L}^2_{\mathbb{H}}(\mathcal{Y})}$ such that

$$\|\Delta_{0,\epsilon_2}\| \le (b_{P_0,C} + \epsilon_2) \tag{4.5}$$

[10] Since $\delta_g(P_0, P_1) \le 1$ for all $P_1 \in \mathscr{P}^e_{\mathcal{U},\mathcal{Y}}$, it is without loss of generality that β is taken to be less than or equal to 1; otherwise there is an obvious contradiction to what is being proved.

and $(\mathrm{I}+\hat{\mathsf{K}}_{l\star}(\lambda_{0,\epsilon_0})\Delta_{0,\epsilon_2})$ is not invertible in $\mathcal{B}_{\mathrm{L}^2_{\mathbb{H}}(\mathcal{U}),\mathrm{L}^2_{\mathbb{H}}(\mathcal{U})}$.[11] Defining

$$\hat{\Delta}_{\epsilon_2}(\lambda):=\frac{\lambda}{\lambda_{0,\epsilon_0}}\Delta_{0,\epsilon_2}\in \mathrm{H}^\infty_{\mathbb{D}}\left(\mathcal{B}_{\mathrm{L}^2_{\mathbb{H}}(\mathcal{U}),\mathrm{L}^2_{\mathbb{H}}(\mathcal{V})}\right), \tag{4.6}$$

it follows that

$$\|(\mathrm{I}+\hat{\mathsf{K}}_{l\star}(\lambda)\hat{\Delta}_{\epsilon_2}(\lambda))^{-1}\|\to\infty$$

as $\lambda\to\lambda_{0,\epsilon_0}$, since $(\mathrm{I}+\hat{\mathsf{K}}_{l\star}\hat{\Delta})$ is continuous on $\mathbb{D}$ and $(\mathrm{I}+\hat{\mathsf{K}}_{l\star}(\lambda_{0,\epsilon_0})\Delta_{0,\epsilon_2})$ is not invertible.[12] Clearly then, $(\mathrm{I}+\hat{\mathsf{K}}_{l\star}\hat{\Delta}_{\epsilon_2})$ is not invertible in $\mathrm{H}^\infty_{\mathbb{D}}(\mathcal{B}_{\mathrm{L}^2_{\mathbb{H}}(\mathcal{U}),\mathrm{L}^2_{\mathbb{H}}(\mathcal{U})})$.

For some $\hat{\mathsf{Q}}_1(\hat{\mathsf{Q}}_1^{-1})\in \mathrm{H}^\infty_{\mathbb{D}}(\mathcal{B}_{\mathrm{L}^2_{\mathbb{H}}(\mathcal{U}),\mathrm{L}^2_{\mathbb{H}}(\mathcal{U})})$ let

$$\hat{\mathsf{G}}_{r1}:=(\hat{\mathsf{G}}_{r0}+\hat{\Delta}_{\epsilon_2})\hat{\mathsf{Q}}_1,$$

and note that

$$\hat{\mathsf{K}}_{l\star}\hat{\mathsf{G}}_{r1}=(\mathrm{I}+\hat{\mathsf{K}}_{l\star}\hat{\Delta}_{\epsilon_2})\hat{\mathsf{Q}}_1.$$

Clearly, the function $\hat{\mathsf{K}}_{l\star}\hat{\mathsf{G}}_{r1}$ is not invertible in $\mathrm{H}^\infty_{\mathbb{D}}(\mathcal{B}_{\mathrm{L}^2_{\mathbb{H}}(\mathcal{U}),\mathrm{L}^2_{\mathbb{H}}(\mathcal{U})})$. Consequently, with $\mathsf{G}_{r1}:=\mathbf{Z}^{-1}\mathbf{M}_{\hat{\mathsf{G}}_{r1}}\mathbf{Z}$, the stable LSI system $\mathsf{K}_{l\star}\mathsf{G}_{r1}:\ell^2_{\mathbb{Z}^+}(\mathrm{L}^2_{\mathbb{H}}(\mathcal{U}))\to\ell^2_{\mathbb{Z}^+}(\mathrm{L}^2_{\mathbb{H}}(\mathcal{U}))$ is not boundedly invertible. So provided that

(a) $\mathcal{K}_{\mathsf{G}_{r1}}=\{0\}$,
(b) $\mathcal{R}_{\mathsf{G}_{r1}}$ is isomorphic (via $\mathbf{W}$) to the graph of a system $P_1\in\mathscr{P}^e_{\mathcal{U},\mathcal{Y}}$, and
(c) $\delta_g(P_0,P_1)<\beta$,

a contradiction to the initial hypothesis that (ii) holds for some $1\geq\beta>b_{P_0,C}$ can be established using Lemma 2.6 and Remark 2.7. Accordingly, the remainder of the proof is devoted to showing that by appropriate choice of ϵ_0, ϵ_2, and $\hat{\mathsf{Q}}_1$, conditions (a), (b), and (c) can be satisfied.

First note that $\mathsf{G}_{r1}=(\mathsf{G}_{r0}+\Delta_{\epsilon_2})\mathsf{Q}_1$, where

$$\mathsf{G}_{r0}:=\mathbf{Z}^{-1}\mathbf{M}_{\hat{\mathsf{G}}_{r0}}\mathbf{Z},\ \ \Delta_{\epsilon_2}:=\mathbf{Z}^{-1}\mathbf{M}_{\hat{\Delta}_{\epsilon_2}}\mathbf{Z},\ \text{ and }\ \mathsf{Q}_1:=\mathbf{Z}^{-1}\mathbf{M}_{\hat{\mathsf{Q}}_1}\mathbf{Z}.$$

Furthermore, note that $\hat{\Delta}_{\epsilon_2}(\lambda)$ has continuous extension to the closed unit-disc and hence that $\|\hat{\Delta}_{\epsilon_2}\|_\infty=\max_{\omega\in[0,2\pi)}\|\hat{\Delta}_{\epsilon_2}(e^{j\omega})\|$. As such, it follows by Proposition 2.3 and from (4.5) and (4.6) that

$$\|\Delta_{\epsilon_2}\|=\|\hat{\Delta}_{\epsilon_2}\|_\infty=\frac{1}{|\lambda_{0,\epsilon_0}|}\|\Delta_{0,\epsilon_2}\|\leq\frac{b_{P_0,C}+\epsilon_2}{1-\epsilon_0}.$$

[11] To see this, let A denote a bounded, linear operator in $\mathcal{B}_{\mathcal{U},\mathcal{Y}}$. Then for any $\epsilon>0$, it follows by definition of the induced norm that there exists a $u\in\mathcal{U}$ (which is taken to have unit norm) such that with $y:=Au$,

$$\|A\|\geq\|y\|_{\mathcal{Y}}\geq\|A\|-\epsilon.$$

Defining $\Delta:\mathcal{Y}\to\mathcal{U}$ to be the operator that maps αy to $-\alpha u$ for all $\alpha\in\mathbb{C}$ and every other direction orthogonal to this in $\mathcal{Y}$ to 0, it follows that

$$\|\Delta\|\leq\frac{1}{\|A\|-\epsilon},$$

and that $y\in\mathcal{K}_{(I+A\Delta)}$. Correspondingly, $(I+A\Delta)$ is not invertible.

[12] See [8, Lemma 4.3] (only if part) for a similar result.

By hypothesis $(\beta - b_{P_0,C}) > 0$ and setting $\epsilon_2 = (\beta - b_{P_0,C})/2 > 0$, $\epsilon_1 < \epsilon_2$ and $\epsilon_0 < \epsilon_2/2\beta$, gives

$$\|\Delta_{\epsilon_2}\| < \beta \leq 1. \tag{4.7}$$

Hence, since G_{r0} is an isometry, it follows that there exists a constant $c > 0$ such that $\|\mathsf{G}_{r1}\,\overline{\mathsf{q}}^{\circ}\|_{\ell^2_{\mathbb{Z}^+}(\mathrm{L}^2_{\mathbb{H}}(\mathcal{V}))} \geq c\|\,\overline{\mathsf{q}}^{\circ}\|_{\ell^2_{\mathbb{Z}^+}(\mathrm{L}^2_{\mathbb{H}}(\mathcal{U}))}$ for all $\overline{\mathsf{q}}^{\circ} \in \ell^2_{\mathbb{Z}^+}(\mathrm{L}^2_{\mathbb{H}}(\mathcal{U}))$. Correspondingly, $\mathcal{R}_{\mathsf{G}_{r1}}$ is closed and $\mathcal{K}_{\mathsf{G}_{r1}} = \{0\}$. That is, condition (a) is satisfied.

$\hat{\mathsf{Q}}_1 \in \mathrm{H}^{\infty}_{\mathbb{D}}(\mathcal{B}_{\mathrm{L}^2_{\mathbb{H}}(\mathcal{U}),\mathrm{L}^2_{\mathbb{H}}(\mathcal{U})})$ is now constructed so that $\mathcal{R}_{\mathsf{G}_{r1}}$ is isomorphic (via $\mathbf{W}$) to the graph of a system $P_1 \in \mathscr{P}^e_{\mathcal{U},\mathcal{Y}}$. First let

$$\begin{bmatrix}\mathsf{M}_{r1}\\ \mathsf{N}_{r1}\end{bmatrix} := \mathsf{G}_{r1} \text{ and } \begin{bmatrix}\mathsf{M}_{r0}\\ \mathsf{N}_{r0}\end{bmatrix} := \mathsf{G}_{r0},$$

where the partitioning is conformal with $\mathcal{G}_{\mathsf{P}_0}$. By construction, $\Delta_{[0]} = 0$, where the subscript [0] denotes the first term in the sequence uniquely identifiable with the block-Toeplitz representation of Δ. Thus, $\mathsf{G}_{r1[0]} = (\mathsf{G}_{r0}\mathsf{Q}_1)_{[0]}$ and

$$\begin{bmatrix}\mathsf{M}_{r1[0]}\\ \mathsf{N}_{r1[0]}\end{bmatrix} = \left(\begin{bmatrix}\mathsf{M}_{r0}\\ \mathsf{N}_{r0}\end{bmatrix}\mathsf{Q}_1\right)_{[0]} = \begin{bmatrix}\mathsf{M}_{r0[0]}\\ \mathsf{N}_{r0[0]}\end{bmatrix}\mathsf{Q}_{1[0]}. \tag{4.8}$$

Since $\mathcal{R}_{\mathsf{G}_{r0}}$ is isomorphic to $\mathcal{G}_{P_0}$ and $P_0 \in \mathscr{P}^e_{\mathcal{U},\mathcal{Y}}$, it can be shown that $\mathsf{M}_{r0[0]} \in \mathcal{B}_{\mathrm{L}^2_{\mathbb{H}}(\mathcal{U}),\mathrm{L}^2_{\mathbb{H}}(\mathcal{U})}$ is boundedly invertible (for the details of this see the proof of Theorem 4.1 in [5]). So by the assumed invertibility of Q_1 and (4.8), $\mathsf{M}_{r1[0]}$ is also invertible in $\mathcal{B}_{\mathrm{L}^2_{\mathbb{H}}(\mathcal{U}),\mathrm{L}^2_{\mathbb{H}}(\mathcal{U})}$. Consequently, M_{r1} has zero kernel, which confirms that $\mathcal{R}_{\mathsf{G}_{r1}}$ is the graph of a linear operator. In fact, it is isomorphic to the graph of an LPTV system P_1 defined on a subspace of $\mathrm{L}^2_{\mathbb{R}^+}(\mathcal{U})$. It is sufficient therefore to select $\hat{\mathsf{Q}}_1$ to ensure that P_1 is causal and that it has locally Lipschitz-continuous extension. To this end, set $\hat{\mathsf{Q}}_1 = \mathsf{M}^{-1}_{r0[0]}$ (which is clearly an invertible element in $\mathrm{H}^{\infty}_{\mathbb{D}}(\mathcal{B}_{\mathrm{L}^2_{\mathbb{H}}(\mathcal{U}),\mathrm{L}^2_{\mathbb{H}}(\mathcal{U})})$) so that $\mathsf{Q}_{1[0]} = \mathsf{M}^{-1}_{r0[0]}$ and

$$\mathsf{G}_{r1[0]} = \begin{bmatrix}\mathsf{M}_{r1[0]}\\ \mathsf{N}_{r1[0]}\end{bmatrix} = (\mathsf{G}_{r0}\mathsf{Q}_1)_{[0]} = \begin{bmatrix}\mathrm{I}\\ \mathsf{N}_{r0[0]}\mathsf{M}^{-1}_{r0[0]}\end{bmatrix}.$$

Since $\mathsf{P}_{0[0]} = \mathsf{N}_{r0[0]}\mathsf{M}^{-1}_{r0[0]}$ is causal on $\mathrm{L}^2_{\mathbb{H}}(\mathcal{U})$, this implies that $\mathsf{G}_{r1[0]}$ is a causal map on $\mathrm{L}^2_{\mathbb{H}}(\mathcal{U})$. So in view of Remark 2.2, $\left[\begin{smallmatrix}M_{r1}\\ N_{r1}\end{smallmatrix}\right] := G_{r1} = \mathbf{W}^{-1}\mathsf{G}_{r1}\mathbf{W}$ is a stable system in $\mathscr{P}^e_{\mathcal{U},\mathcal{V}}$. This, in turn, implies that P_1 is itself causal. To see this, suppose that P_1 is not causal. Then by definition, there exists a point $\left[\begin{smallmatrix}u\\ y\end{smallmatrix}\right] \in \mathcal{G}_{P_1} = \mathcal{R}_{G_{r1}}$ and a $\tau_1 \in \mathbb{R}^+$ such that

$$\mathbf{T}_{\tau_1}\begin{bmatrix}u\\ y\end{bmatrix} = \begin{bmatrix}0\\ \tilde{y} \neq 0\end{bmatrix}.$$

That is, there exists a $q \in \mathrm{L}^2_{\mathbb{R}^+}(\mathcal{U})$ such that $\tilde{y} = \mathbf{T}_{\tau_1}N_{r1}q$ and $0 = \mathbf{T}_{\tau_1}M_{r1}q$. Since

M_{r1} is causal,

$$\begin{aligned}\mathbf{T}_{\tau_1} M_{r1} q &= \mathbf{T}_{\tau_1}\mathbf{W}^{-1}\mathbf{T}_n \mathsf{M}_{r1}\mathbf{W}q \\ &= \mathbf{T}_{\tau_1}\mathbf{W}^{-1}\begin{bmatrix}\underbrace{\begin{bmatrix}\mathsf{M}_{r1[0]}(=\mathrm{I}) & 0 & \cdots & 0\\ \mathsf{M}_{r1[1]} & \ddots & \ddots & \vdots\\ \vdots & \ddots & \ddots & 0\\ \mathsf{M}_{r1[n]} & \cdots & \mathsf{M}_{r1[1]} & \mathsf{M}_{r1[0]}\end{bmatrix}}_{:=\tilde{\mathsf{M}}_n} \begin{bmatrix}\overset{\circ}{\mathsf{q}}_0\\ \overset{\circ}{\mathsf{q}}_1\\ \vdots\\ \vdots\\ \overset{\circ}{\mathsf{q}}_n\end{bmatrix}\\ 0\\ \vdots\end{bmatrix},\end{aligned} \tag{4.9}$$

where $\overset{\circ}{\mathsf{q}}(=:k\mapsto\overset{\circ}{\mathsf{q}}_k):=\mathbf{W}q$, $n=\lfloor\frac{\tau_1}{h}\rfloor$ and $\lfloor\cdot\rfloor$ denotes the integer part of a real number. Note that $\tilde{\mathsf{M}}_i$ is boundedly invertible for all finite i and, correspondingly, that it has zero kernel and full range.[13] So if $0=\mathbf{T}_{\tau_1}M_{r1}q$, it follows from (4.9) that $\overset{\circ}{\mathsf{q}}_0=\overset{\circ}{\mathsf{q}}_1=\cdots=\overset{\circ}{\mathsf{q}}_{n-1}=0$ and $\mathbf{T}_{(\tau_1-nh)}\overset{\circ}{\mathsf{q}}_n=0$. Equivalently, $\mathbf{T}_{\tau_1}q=0$. Now since N_{r1} is causal, this implies that $\tilde{y}=\mathbf{T}_{\tau_1}N_{r1}q=0$, which is a contradiction. Consequently, P_1 must be causal. This line of reasoning also leads to the conclusion that P_1 is causally extendible to a locally Lipschitz-continuous operator on $\mathrm{L}^{2,\mathrm{e}}_{\mathbb{R}^+}(\mathcal{U})$. To see this note that, since $\tilde{\mathsf{M}}_i$ has full range for all finite i, given any $\tau<\infty$

$$\mathbf{T}_\tau\mathcal{D}_{P_1}=\mathbf{T}_\tau\mathcal{R}_{M_{r1}}=\mathbf{T}_\tau\mathbf{W}^{-1}\begin{bmatrix}\mathcal{R}_{\tilde{\mathsf{M}}_n}\\ 0\end{bmatrix}=\mathbf{T}_\tau\mathrm{L}^2_{[0,(n+1)h)}(\mathcal{U})=\mathbf{T}_\tau\mathrm{L}^2_{\mathbb{R}^+}(\mathcal{U}),$$

where $n=\lfloor\frac{\tau}{h}\rfloor$. Correspondingly, P_1 is causally extendible (by definition). Now denote the extension of P_1 by P_{1e} and observe that

$$\sup_{\substack{u_1,u_2\in\mathrm{L}^{2,\mathrm{e}}_{\mathbb{R}^+}(\mathcal{U})\\ \mathbf{T}_\tau u_1\neq\mathbf{T}_\tau u_2}}\frac{\|\mathbf{T}_\tau(P_{1e}u_1-P_{1e}u_2)\|_{\mathrm{L}^2_{\mathbb{R}^+}(\mathcal{Y})}}{\|\mathbf{T}_\tau(u_1-u_2)\|_{\mathrm{L}^2_{\mathbb{R}^+}(\mathcal{U})}}\le\sup_{\substack{\mathsf{u}\in\mathbf{T}_n\ell^2_{\mathbb{Z}^+}(\mathrm{L}^2_{\mathbb{H}}(\mathcal{U}))\\ \mathsf{u}\neq0}}\frac{\|\tilde{\mathsf{N}}_n\tilde{\mathsf{M}}_n^{-1}\mathsf{u}\|_{\mathbf{T}_n\ell^2_{\mathbb{Z}^+}(\mathrm{L}^2_{\mathbb{H}}(\mathcal{Y}))}}{\|\mathsf{u}\|_{\mathbf{T}_n\ell^2_{\mathbb{Z}^+}(\mathrm{L}^2_{\mathbb{H}}(\mathcal{U}))}},$$

where

$$\tilde{\mathsf{N}}_n:=\begin{bmatrix}\mathsf{N}_{r1[0]} & 0 & \cdots & 0\\ \mathsf{N}_{r1[1]} & \ddots & \ddots & \vdots\\ \vdots & \ddots & \ddots & 0\\ \mathsf{N}_{r1[n]} & \cdots & \mathsf{N}_{r1[1]} & \mathsf{N}_{r1[0]}\end{bmatrix}.$$

Then, since $\tilde{\mathsf{N}}_i$ and $\tilde{\mathsf{M}}_i^{-1}$ are both bounded for all finite i, it follows (by definition) that P_{1e} is locally Lipschitz-continuous as claimed.

It remains to show that $\delta_g(P_0,P_1)<\beta$. First it is shown that $\vec{\delta}_g(P_0,P_1)<\beta$ and then that $\delta_g(P_0,P_1)=\vec{\delta}_g(P_0,P_1)<\beta$. Note that $\hat{\mathcal{G}}_{\mathsf{P}_1}:=\mathbf{Z}\mathbf{W}\mathcal{G}_{\mathsf{P}_1}$ is a shift-invariant subspace of $\mathrm{H}^2_{\mathbb{D}}(\mathrm{L}^2_{\mathbb{H}}(\mathcal{V}))$. Correspondingly, by the Beurling–Lax–Halmos theorem [11, p. 239], there exists an inner function $\hat{\mathsf{G}}_{r\star}\in\mathrm{H}^\infty_{\mathbb{D}}(\mathcal{B}_{\mathrm{L}^2_{\mathbb{H}}(\mathcal{U}),\mathrm{L}^2_{\mathbb{H}}(\mathcal{V})})$ such that $\hat{\mathcal{G}}_{\mathsf{P}_1}=\hat{\mathsf{G}}_{r1\star}\mathrm{H}^2_{\mathbb{D}}(\mathrm{L}^2_{\mathbb{H}}(\mathcal{U}))$. Moreover, since $\mathcal{R}_{\mathbf{M}_{\hat{\mathsf{G}}_{r1}}}=\mathcal{R}_{\mathbf{M}_{\hat{\mathsf{G}}_{r1\star}}}$ and $\mathcal{K}_{\mathbf{M}_{\hat{\mathsf{G}}_{r1}}}=\{0\}$, there

[13] In fact $\tilde{\mathsf{M}}_i^{-1}$ here is block-lower-triangular with identities down the diagonal.

exists a $\hat{\mathsf{Q}}_\star(\hat{\mathsf{Q}}_\star^{-1}) \in \mathrm{H}^\infty_{\mathbb{D}}(\mathcal{B}_{\mathrm{L}^2_{\mathbb{H}}(\mathcal{U}),\mathrm{L}^2_{\mathbb{H}}(\mathcal{U})})$ such that $\hat{\mathsf{G}}_{r1\star}\hat{\mathsf{Q}}_\star = \hat{\mathsf{G}}_{r1} = (\hat{\mathsf{G}}_{r0} + \hat{\Delta})\hat{\mathsf{Q}}_1$ (see Corollary IX.2.2 of the Beurling–Lax–Halmos theorem in [11]). Now by Theorem 3.3,

$$\begin{aligned}\vec{\delta}_g(P_0,P_1) = \vec{\delta}(\hat{\mathcal{G}}_{\mathsf{P}_0},\hat{\mathcal{G}}_{\mathsf{P}_1}) &= \inf_{\hat{\mathsf{Q}}\in \mathrm{H}^\infty_{\mathbb{D}}(\mathcal{B}_{\mathrm{L}^2_{\mathbb{H}}(\mathcal{U}),\mathrm{L}^2_{\mathbb{H}}(\mathcal{U})})} \|\hat{\mathsf{G}}_{r0} - \hat{\mathsf{G}}_{r1\star}\hat{\mathsf{Q}}\|_\infty \\ &= \inf_{\hat{\mathsf{Q}}\in \mathrm{H}^\infty_{\mathbb{D}}(\mathcal{B}_{\mathrm{L}^2_{\mathbb{H}}(\mathcal{U}),\mathrm{L}^2_{\mathbb{H}}(\mathcal{U})})} \|\hat{\mathsf{G}}_{r0} - (\hat{\mathsf{G}}_{r0} + \hat{\Delta}_{\epsilon_2})\hat{\mathsf{Q}}_1\hat{\mathsf{Q}}_\star^{-1}\hat{\mathsf{Q}}\|_\infty \\ &\le \|\hat{\Delta}_{\epsilon_2}\|_\infty < \beta,\end{aligned}$$

where the third inequality follows from the fact that $\hat{\mathsf{Q}}_\star\hat{\mathsf{Q}}_1^{-1} \in \mathrm{H}^\infty_{\mathbb{D}}(\mathcal{B}_{\mathrm{L}^2_{\mathbb{H}}(\mathcal{U}),\mathrm{L}^2_{\mathbb{H}}(\mathcal{U})})$ and the last inequality follows from (4.7).

Now for any $\varphi \in \mathrm{H}^2_{\mathbb{D}}(\mathrm{L}^2_{\mathbb{H}}(\mathcal{U}))$, consider the projection of $\mathbf{M}_{\hat{\mathsf{G}}_{r1}}\varphi \in \hat{\mathcal{G}}_{\mathsf{P}_1}$ onto $\hat{\mathcal{G}}_{\mathsf{P}_0}$. Since $\mathbf{M}_{\hat{\mathsf{G}}_{r0}}$ is an isometry, $\mathbf{\Pi}_{\hat{\mathcal{G}}_{\mathsf{P}_0}} = \mathbf{M}_{\hat{\mathsf{G}}_{r0}}(\mathbf{M}_{\hat{\mathsf{G}}_{r0}})^*$ and the projection described above can be expressed as

$$\mathbf{M}_{\hat{\mathsf{G}}_{r0}}\left(\mathbf{M}_{\hat{\mathsf{G}}_{r0}}\right)^*\left(\mathbf{M}_{\hat{\mathsf{G}}_{r0}} + \mathbf{M}_{\hat{\Delta}_{\epsilon_2}}\right)\mathbf{M}_{\hat{\mathsf{Q}}_1}\varphi = \mathbf{M}_{\hat{\mathsf{G}}_{r0}}\zeta,$$

where

$$\zeta := \left(\mathrm{I} + \left(\mathbf{M}_{\hat{\mathsf{G}}_{r0}}\right)^*\mathbf{M}_{\hat{\Delta}_{\epsilon_2}}\right)\mathbf{M}_{\hat{\mathsf{Q}}_1}\varphi.$$

Furthermore,

$$\|\left(\mathbf{M}_{\hat{\mathsf{G}}_{r0}}\right)^*\mathbf{M}_{\hat{\Delta}_{\epsilon_2}}\| \le \|\mathbf{M}_{\hat{\mathsf{G}}_{r0}}\| \cdot \|\mathbf{M}_{\hat{\Delta}_{\epsilon_2}}\| < \beta \le 1.$$

Thus, $(\mathrm{I} + (\mathbf{M}_{\hat{\mathsf{G}}_{r0}})^*\mathbf{M}_{\hat{\Delta}_\epsilon})$ is a *one-to-one* mapping *onto* $\mathrm{H}^2_{\mathbb{D}}(\mathrm{L}^2_{\mathbb{H}}(\mathcal{U}))$ and since $\mathbf{M}_{\hat{\mathsf{Q}}_1}$ is bijective, it follows that the projection described above is also bijective. So by Proposition 3.1,

$$\vec{\delta}(\hat{\mathcal{G}}_{\mathsf{P}_0},\hat{\mathcal{G}}_{\mathsf{P}_1}) = \vec{\delta}(\hat{\mathcal{G}}_{\mathsf{P}_1},\hat{\mathcal{G}}_{\mathsf{P}_0}) = \delta(\hat{\mathcal{G}}_{\mathsf{P}_0},\hat{\mathcal{G}}_{\mathsf{P}_1})$$

and hence $\delta_g(P_0,P_1) = \delta(\hat{\mathcal{G}}_{\mathsf{P}_0},\hat{\mathcal{G}}_{\mathsf{P}_1}) = \vec{\delta}(\hat{\mathcal{G}}_{\mathsf{P}_0},\hat{\mathcal{G}}_{\mathsf{P}_1}) < \beta$, as required. This completes the proof. □

Consider a closed-loop $[P_0, C]$ with $P_0 \in \mathscr{P}^{e,sc}_{\mathcal{U},\mathcal{Y}}$ and $C \in \mathscr{P}^{e}_{\mathcal{Y},\mathcal{U}}$. Since the plant P_0 is strongly causal, the closed-loop $[P_0, C]$ is well-posed. Similarly, if the set of permissible perturbed plants were restricted to be a subset of $\mathscr{P}^{e,sc}_{\mathcal{U},\mathcal{Y}}$, well-posedness would be maintained. Now note that by the way the destabilizing plant P_1 is constructed in the proof of Theorem 4.2, $\mathsf{P}_{0[0]} = \mathsf{P}_{1[0]}$. Correspondingly, the instantaneous behavior of the nominal plant and the plant constructed to destabilize the closed loop is the same. As such, if P_0 were strongly causal, then P_1 would also be strongly causal. Consequently, (ii) ⇒ (i) in Theorem 4.2 would still hold if permissible perturbed plants were restricted to strongly causal ones only. This is summarized below.

COROLLARY 4.3. *Given a stable, closed-loop system* $[P_0, C]$ *with* $P_0 \in \mathscr{P}^{e,sc}_{\mathcal{U},\mathcal{Y}}$ *and* $C \in \mathscr{P}^{e}_{\mathcal{Y},\mathcal{U}}$, *the following are equivalent.*

(i) $b_{P_0,C} \ge \beta$.

(ii) $[P_1, C]$ *is stable for all* $P_1 \in \mathscr{P}^{e,sc}_{\mathcal{U},\mathcal{Y}}$ *that satisfy* $\delta_g(P_0,P_1) < \beta$.

Similarly, although it is very difficult to characterize analytically, restricting the set of permissible perturbations to all of those which do not cause ill-posedness[14] leads to the following restatement of Theorem 4.2.

[14]This is necessary for (i) ⇒(ii) to hold.

COROLLARY 4.4. *Given a stable, closed-loop system* $[P_0, C]$ *with* $P_0 \in \mathscr{P}^e_{\mathcal{U},\mathcal{Y}}$ *and* $C \in \mathscr{P}^e_{\mathcal{Y},\mathcal{U}}$, *the following are equivalent.*

(i) $b_{P_0,C} \geq \beta$.

(ii) $[P_1, C]$ *is stable for all* $P_1 \in \mathscr{P}^e_{\mathcal{U},\mathcal{Y}}$ *such that* $\delta_g(P_0, P_1) < \beta$ *and* $[P_1, C]$ *is well-posed.*

4.2. Simultaneous plant and controller perturbations. The maximum combined perturbation to both P and C that does not cause instability is quantified in terms of the gap metric in this subsection. As a corollary of Theorem 4.2, it is first shown that a given controller C stabilizes a gap-ball centered at P if and only if P is stabilized by all controllers in a gap-ball of the same radius and centered at C. The maximum combined gap-perturbation (as a sum) to P and C that the closed loop can tolerate is then considered. The results presented are analogous to the LTI results of [14, Sect. VI] and the LTV results of [12, Sect. 4]. Before continuing, it is instructive to recall the following results.

PROPOSITION 4.5 (see [7]). *Suppose that two (closed) manifolds* $\mathcal{F}$ *and* $\mathcal{G}$ *induce a coordinatization of a Hilbert space* $\mathcal{H}$. *If* $\mathcal{F}$ *is linear, then*

$$\|\mathbf{\Pi}_{\mathcal{F}\|\mathcal{G}}\| = \|\mathbf{\Pi}_{\mathcal{G}\|\mathcal{F}}\|.$$

COROLLARY 4.6. *Given a stable closed-loop* $[P, C]$ *with* $P \in \mathscr{P}_{\mathcal{U},\mathcal{Y}}$ *and* $C \in \mathscr{P}_{\mathcal{Y},\mathcal{U}}$,

$$b_{P,C} = b_{C,P}.$$

Proof. Note that (see Definition 4.1)

$$b_{C,P} := \|\mathbf{\Pi}_{\mathcal{G}_C\|\mathcal{G}_P^\sharp}\| = \left\| \begin{bmatrix} 0 & I \\ I & 0 \end{bmatrix} \mathbf{\Pi}_{\mathcal{G}_C^\sharp\|\mathcal{G}_P} \begin{bmatrix} 0 & I \\ I & 0 \end{bmatrix} \right\| = \|\mathbf{\Pi}_{\mathcal{G}_C^\sharp\|\mathcal{G}_P}\|.$$

Now since $[P, C]$ is stable, $\mathcal{G}_P$ and $\mathcal{G}_C^\sharp$ induce a coordinatization of $\mathrm{L}^2_{\mathbb{R}^+}(\mathcal{V})$ (see Proposition 2.4). So by Proposition 4.5, $b_{C,P} = \|\mathbf{\Pi}_{\mathcal{G}_C^\sharp\|\mathcal{G}_P}\| = \|\mathbf{\Pi}_{\mathcal{G}_P\|\mathcal{G}_C^\sharp}\| = b_{P,C}$. ∎

In view of Corollary 4.6, the following result, which characterizes robustness to gap-metric perturbations of the controller, is an immediate consequence of Theorem 4.2.

COROLLARY 4.7. *Let* $[P_0, C_0]$ *be a stable, closed-loop system with* $P_0 \in \mathscr{P}^{e,sc}_{\mathcal{U},\mathcal{Y}}$ *and* $C_0 \in \mathscr{P}^e_{\mathcal{Y},\mathcal{U}}$. *Then the following are equivalent.*

(i) $[P_1, C_0]$ *is stable for all* $P_1 \in \mathscr{P}^{e,sc}_{\mathcal{U},\mathcal{Y}}$ *such that* $\delta_g(P_0, P_1) < \beta$.

(ii) $[P_0, C_1]$ *is stable for all* $C_1 \in \mathscr{P}^e_{\mathcal{Y},\mathcal{U}}$ *such that* $\delta_g(C_0, C_1) < \beta$.

REMARK 4.8. *As with Corollaries* 4.3 *and* 4.4, *Corollary* 4.7 *can be restated with various constraints on the strength of causality of the plants and the controllers.*

Now, with this established, it is possible to quantitatively characterize robustness to perturbations of both P and its stabilizing controller C.

THEOREM 4.9. *Consider a stable closed-loop system* $[P_0, C_0]$ *with* $P_0 \in \mathscr{P}^{e,sc}_{\mathcal{U},\mathcal{Y}}$ *and* $C_0 \in \mathscr{P}^e_{\mathcal{Y},\mathcal{U}}$. *The following are equivalent.*

(i) $\beta \leq b_{P_0,C_0}$.

(ii) $[P_1, C_1]$ *is stable for all* $P_1 \in \mathscr{P}^{e,sc}_{\mathcal{U},\mathcal{Y}}$ *and* $C_1 \in \mathscr{P}^e_{\mathcal{Y},\mathcal{U}}$ *such that*

$$\delta_g(P_0, P_1) + \delta_g(C_0, C_1) < \beta.$$

Proof. The proof follows directly from Corollary 4.3 of Theorem 4.2 and Corollary 4.7, in exactly the same way as the corresponding result [14, Thm. 7]. Assume that

(i) holds, so that by Corollaries 4.3 and 4.7, $[P_0, C_1]$ is stable for all $C_1 \in \mathscr{B}^{\mathrm{e}}_{\delta_g}(C_0, \beta)$, where

$$\mathscr{B}^{\mathrm{e}}_{\delta_g}(F, \epsilon) := \{\grave{F} \in \mathscr{P}^{e}_{\mathcal{Y},\mathcal{U}} \ : \ \delta_g(F, \grave{F}) < \epsilon\}.$$

Consider a controller $C_1 \in \mathscr{P}^{e}_{\mathcal{Y},\mathcal{U}}$, which satisfies $\tau := \delta_g(C_0, C_1) < \beta$. By the metric property of the gap, it follows that $\mathscr{B}^{\mathrm{e}}_{\delta_g}(C_1, \beta - \tau) \subset \mathscr{B}^{\mathrm{e}}_{\delta_g}(C_0, \beta)$ and hence that $[P_0, C_2]$ is stable for all $C_2 \in \mathscr{B}^{\mathrm{e}}_{\delta_g}(C_1, \beta - \tau)$. Using Corollary 4.7, it is then possible to conclude that $[P_1, C_1]$ is stable for all $P_1 \in \mathscr{B}^{\mathrm{e,sc}}_{\delta_g}(P_0, \beta - \tau)$, where

$$\mathscr{B}^{\mathrm{e,sc}}_{\delta_g}(F, \epsilon) := \{\grave{F} \in \mathscr{P}^{e,sc}_{\mathcal{Y},\mathcal{U}} \ : \ \delta_g(F, \grave{F}) < \beta\}.$$

Since this holds for any $C_1 \in \mathscr{P}^{e}_{\mathcal{U},\mathcal{Y}}$ satisfying $\delta_g(C_0, C_1) < \beta$, it is clear that (ii) holds. That (ii) $\Rightarrow$ (i) is immediate, since (ii) here implies (i) in Corollary 4.3. This is precisely $\beta \le b_{P_0,C_0}$. ∎

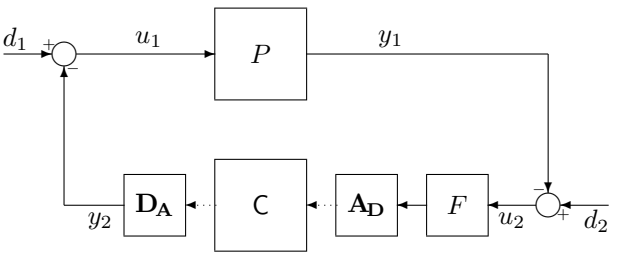


FIG. 4.1. *Sampled-data control system.*

4.3. Robustness of SD control systems. In this subsection, the robustness results obtained in preceding subsections are specialized to the SD case. Set $\mathcal{U} = \mathbb{R}^m$ and $\mathcal{Y} = \mathbb{R}^p$, and consider the feedback configuration shown in Figure 4.1. Let the plant P be in $\mathscr{P}^{e}_{\mathcal{U},\mathcal{Y}}$; the controller $\mathsf{C} : \mathcal{D}_{\mathsf{C}} \subset \ell^2_{\mathbb{Z}_+}(\mathcal{Y}) \to \ell^2_{\mathbb{Z}_+}(\mathcal{U})$ be an LSI, discrete-time system; $\mathbf{A_D}$ be an ideal h-periodic sampling-device; $\mathbf{D_A}$ be a zero-order hold device synchronized with $\mathbf{A_D}$; and F be a stable, low-pass LTI filter required to ensure boundedness of the sampling device. Note that $F \in \mathscr{P}^{e,sc}_{\mathcal{Y},\mathcal{Y}}$ and hence that $C^{sd} := \mathbf{D_A}\mathsf{C}\mathbf{A_D}F \in \mathscr{P}^{e,sc}_{\mathcal{Y},\mathcal{U}}$. Accordingly, Theorem 4.2 applies as summarized in the following corollary.

COROLLARY 4.10. *Consider the stable closed-loop system $[P_0, C^{sd}]$ described above. The following are equivalent.*

(i) $b_{P_0,C^{sd}} \ge \beta$.

(ii) $[P_1, C^{sd}]$ *is stable for all* $P_1 \in \mathscr{P}^{e}_{\mathcal{U},\mathcal{Y}}$ *that satisfy* $\delta_g(P_0, P_1) < \beta$.

REMARK 4.11. *The results presented in subsection* 4.2 *concerning simultaneous perturbations to the plant and controller also apply. Such a result is a useful for characterizing the effect of approximating an LTI continuous-time controller with an SD controller, for example.*

Appendix A. Proof of Theorem 3.3. As mentioned, the proof of Theorem 3.3 follows that of an H^∞ optimization result in [11, p. 248], using the commutant lifting theorem. Note that (for $i = 0, 1$) $\hat{\mathsf{G}}_{\mathsf{P}_i}$ is a shift-invariant subspace of $\mathrm{H}^2_{\mathbb{D}}(\mathrm{L}^2_{\mathbb{H}}(\mathcal{V}))$. As such, it follows by the Beurling–Lax–Halmos theorem (see [11, p. 239]) that there exist inner functions $\hat{\mathsf{G}}_{ri} \in \mathrm{H}^\infty_{\mathbb{D}}(\mathcal{B}_{\mathrm{L}^2_{\mathbb{H}}(\mathcal{U}),\mathrm{L}^2_{\mathbb{H}}(\mathcal{V})})$ such that $\hat{\mathsf{G}}_{\mathsf{P}_i} = \hat{\mathsf{G}}_{ri}\mathrm{H}^2_{\mathbb{D}}(\mathrm{L}^2_{\mathbb{H}}(\mathcal{U}))$ ($i =$

$0,1$).[15] Correspondingly (for $i=0,1$), $\mathbf{M}_{\hat{\mathsf{G}}_{ri}}$ is an isometry and hence it follows that $\mathbf{\Pi}_{\hat{\mathcal{G}}_{\mathsf{P}_i}} = \mathbf{M}_{\hat{\mathsf{G}}_{ri}}(\mathbf{M}_{\hat{\mathsf{G}}_{ri}})^*$. Correspondingly,

$$\vec{\delta}(\hat{\mathcal{G}}_{\mathsf{P}_0}, \hat{\mathcal{G}}_{\mathsf{P}_1}) = \|\mathbf{\Pi}_{\hat{\mathcal{G}}_{\mathsf{P}_1}^{\perp}} \mathbf{M}_{\hat{\mathsf{G}}_{r0}} \left(\mathbf{M}_{\hat{\mathsf{G}}_{r0}}\right)^* \| = \|\mathbf{\Pi}_{\hat{\mathcal{G}}_{\mathsf{P}_1}^{\perp}} \mathbf{M}_{\hat{\mathsf{G}}_{r0}}\|,$$

where the last equality follows from the fact that $\mathbf{M}_{\hat{\mathsf{G}}_{r0}}$ is an isometry. Now recall that $\hat{\mathcal{G}}_{\mathsf{P}_1}^{\perp} := \mathrm{H}^2_{\mathbb{D}}(\mathrm{L}^2_{\mathbb{H}}(\mathcal{V})) \ominus \hat{\mathcal{G}}_{\mathsf{P}_1}$ and take any $\eta \in \mathrm{H}^2_{\mathbb{D}}(\mathrm{L}^2_{\mathbb{H}}(\mathcal{V}))$, which may be expressed as $\eta = \gamma + \varphi$, where $\gamma = \mathbf{\Pi}_{\hat{\mathcal{G}}_{\mathsf{P}_1}} \eta$ and $\varphi = \mathbf{\Pi}_{\hat{\mathcal{G}}_{\mathsf{P}_1}^{\perp}} \eta$. Note that $\mathbf{S}\eta = \mathbf{S}\gamma + \mathbf{S}\varphi$ and that

$$\begin{aligned} \mathbf{\Pi}_{\hat{\mathcal{G}}_{\mathsf{P}_1}^{\perp}} \mathbf{S}\eta &= \mathbf{\Pi}_{\hat{\mathcal{G}}_{\mathsf{P}_1}^{\perp}} \mathbf{S}\gamma + \mathbf{\Pi}_{\hat{\mathcal{G}}_{\mathsf{P}_1}^{\perp}} \mathbf{S}\varphi \\ &= \mathbf{\Pi}_{\hat{\mathcal{G}}_{\mathsf{P}_1}^{\perp}} \mathbf{S}\varphi, \end{aligned}$$

where the second equality follows from the fact that $\mathbf{S}\hat{\mathcal{G}}_{\mathsf{P}_1} \subset \hat{\mathcal{G}}_{\mathsf{P}_1}$. That is,

$$\mathbf{\Pi}_{\hat{\mathcal{G}}_{\mathsf{P}_1}^{\perp}} \mathbf{S}\eta = \mathbf{\Pi}_{\hat{\mathcal{G}}_{\mathsf{P}_1}^{\perp}} \mathbf{S}\mathbf{\Pi}_{\hat{\mathcal{G}}_{\mathsf{P}_1}^{\perp}} \eta = \mathbf{S}_\star \mathbf{\Pi}_{\hat{\mathcal{G}}_{\mathsf{P}_1}^{\perp}} \eta, \tag{A.1}$$

where

$$\mathbf{S}_\star := \mathbf{\Pi}_{\hat{\mathcal{G}}_{\mathsf{P}_1}^{\perp}} \mathbf{S}|_{\hat{\mathcal{G}}_{\mathsf{P}_1}^{\perp}} \ : \ \hat{\mathcal{G}}_{\mathsf{P}_1}^{\perp} \to \hat{\mathcal{G}}_{\mathsf{P}_1}^{\perp}.$$

The unilateral shift $\mathbf{S}$ is an isometric dilation (lifting) of $\mathbf{S}_\star$ in that $\mathbf{S}_\star^n = \mathbf{\Pi}_{\hat{\mathcal{G}}_{\mathsf{P}_1}^{\perp}} \mathbf{S}^n|_{\hat{\mathcal{G}}_{\mathsf{P}_1}^{\perp}}$ for all $n \geq 0$. In fact, by the way $\hat{\mathcal{G}}_{\mathsf{P}_1}^{\perp}$ is defined and since $\hat{\mathcal{G}}_{\mathsf{P}_1}$ is shift-invariant, it follows that $\mathbf{S}$ is the minimal, isometric dilation of $\mathbf{S}_\star$ in the sense that $\mathrm{H}^2_{\mathbb{D}}(\mathrm{L}^2_{\mathbb{H}}(\mathcal{V}))$ is the smallest shift-invariant space that contains $\hat{\mathcal{G}}_{\mathsf{P}_1}^{\perp}$.[16] Furthermore, from (A.1) it follows that

$$\mathbf{S}_\star \mathbf{\Pi}_{\hat{\mathcal{G}}_{\mathsf{P}_1}^{\perp}} = \mathbf{\Pi}_{\hat{\mathcal{G}}_{\mathsf{P}_1}^{\perp}} \mathbf{S}.$$

Consequently, with the generalized Hankel operator $\mathbf{H}_{\hat{\mathsf{G}}_{r0}}$ defined to be

$$\mathbf{H}_{\hat{\mathsf{G}}_{r0}} := \mathbf{\Pi}_{\hat{\mathcal{G}}_{\mathsf{P}_1}^{\perp}} \mathbf{M}_{\hat{\mathsf{G}}_{r0}}, \tag{A.2}$$

it follows that

$$\mathbf{S}_\star \mathbf{H}_{\hat{\mathsf{G}}_{r0}} = \mathbf{\Pi}_{\hat{\mathcal{G}}_{\mathsf{P}_1}^{\perp}} \mathbf{S}\mathbf{M}_{\hat{\mathsf{G}}_{r0}} = \mathbf{H}_{\hat{\mathsf{G}}_{r0}} \mathbf{S}.$$

Since $\mathbf{S}$ is a minimal, isometric dilation of $\mathbf{S}_\star$, it follows by the commutant lifting theorem (cf. [11, Chap. VII] and [22]) that there exists an LSI operator $\mathsf{H} : \mathrm{H}^2_{\mathbb{D}}(\mathrm{L}^2_{\mathbb{H}}(\mathcal{U})) \to \mathrm{H}^2_{\mathbb{D}}(\mathrm{L}^2_{\mathbb{H}}(\mathcal{V}))$, which satisfies $\mathbf{\Pi}_{\hat{\mathcal{G}}_{\mathsf{P}_1}^{\perp}} \mathsf{H} = \mathbf{H}_{\hat{\mathsf{G}}_{r0}}$ and

$$\|\mathsf{H}\| = \|\mathbf{H}_{\hat{\mathsf{G}}_{r0}}\|. \tag{A.3}$$

Furthermore, by Proposition 2.3, the operator H can be expressed as a multiplication operator with symbol $\hat{\mathsf{H}} \in \mathrm{H}^\infty_{\mathbb{D}}(\mathcal{B}_{\mathrm{L}^2_{\mathbb{H}}(\mathcal{U}), \mathrm{L}^2_{\mathbb{H}}(\mathcal{V})})$, and

$$\|\mathsf{H}\| = \|\hat{\mathsf{H}}\|_\infty. \tag{A.4}$$

[15] See [5] for a more detailed discussion concerning representations of the graph.

[16] See [11, Chap. VI] or [22, Chaps. I–II] for a precise definition and treatment of minimal dilations.

Thus,

$$\mathbf{\Pi}_{\hat{\mathcal{G}}_{\mathsf{P}_1}^{\perp}} \mathbf{M}_{\hat{\mathsf{H}}} = \mathbf{H}_{\hat{\mathsf{G}}_{r0}} = \mathbf{\Pi}_{\hat{\mathcal{G}}_{\mathsf{P}_1}^{\perp}} \mathbf{M}_{\hat{\mathsf{G}}_{r0}},$$

which implies that $\mathbf{\Pi}_{\hat{\mathcal{G}}_{\mathsf{P}_1}^{\perp}} (\mathbf{M}_{\hat{\mathsf{H}}} - \mathbf{M}_{\hat{\mathsf{G}}_{0r}}) = 0$. Equivalently, since $\hat{\mathcal{G}}_{\mathsf{P}_1} = \hat{\mathsf{G}}_{r1} \mathrm{H}^2_{\mathbb{D}}(\mathrm{L}^2_{\mathbb{H}}(\mathcal{U}))$

$$(\hat{\mathsf{H}} - \hat{\mathsf{G}}_{r0}) \mathrm{H}^2_{\mathbb{D}} \left(\mathrm{L}^2_{\mathbb{H}}(\mathcal{U}) \right) \subset \hat{\mathsf{G}}_{r1} \mathrm{H}^2_{\mathbb{D}} \left(\mathrm{L}^2_{\mathbb{H}}(\mathcal{U}) \right).$$

Correspondingly, by Corollary IX.2.2 in [11, pp. 239–240] (a corollary of the Beurling–Lax–Halmos theorem) and since $\hat{\mathsf{G}}_{r1}$ is inner, there exists a $\hat{\mathsf{Q}}_\star \in \mathrm{H}^\infty_{\mathbb{D}}(\mathcal{B}_{\mathrm{L}^2_{\mathbb{H}}(\mathcal{U}), \mathrm{L}^2_{\mathbb{H}}(\mathcal{U})})$ such that

$$\hat{\mathsf{H}} - \hat{\mathsf{G}}_{r0} = -\hat{\mathsf{G}}_{r1} \hat{\mathsf{Q}}_\star.$$

Hence, $\hat{\mathsf{H}} = \hat{\mathsf{G}}_{r0} - \hat{\mathsf{G}}_{r1} \hat{\mathsf{Q}}_\star$ and

$$\|\hat{\mathsf{G}}_{r0} - \hat{\mathsf{G}}_{r1} \hat{\mathsf{Q}}_\star\|_\infty = \|\mathbf{H}_{\hat{\mathsf{G}}_{r0}}\| = \|\mathbf{H}_{\hat{\mathsf{G}}_{r0} - \hat{\mathsf{G}}_{r1} \hat{\mathsf{Q}}}\| \leq \inf_{\hat{\mathsf{Q}} \in \mathrm{H}^\infty_{\mathbb{D}}(\mathcal{B}_{\mathrm{L}^2_{\mathbb{H}}(\mathcal{U}), \mathrm{L}^2_{\mathbb{H}}(\mathcal{U})})} \|\hat{\mathsf{G}}_{r0} - \hat{\mathsf{G}}_{r1} \hat{\mathsf{Q}}\|_\infty,$$

where the first equality follows from (A.3) and (A.4), the second by the fact that $\mathbf{H}_{\hat{\mathsf{G}}_{r1} \hat{\mathsf{Q}}} = \mathbf{\Pi}_{\hat{\mathcal{G}}_{\mathsf{P}_1}^{\perp}} \mathbf{M}_{\hat{\mathsf{G}}_{r1} \hat{\mathsf{Q}}} = 0$ for all $\hat{\mathsf{Q}}$, and (finally) the third inequality follows by the fact that $\|\mathbf{\Pi}_{\hat{\mathcal{G}}_{\mathsf{P}_1}^{\perp}}\| \leq 1$. Clearly, equality holds when $\hat{\mathsf{Q}} = \hat{\mathsf{Q}}_\star$. In conclusion,

$$\|\mathbf{\Pi}_{\hat{\mathcal{G}}_{\mathsf{P}_1}^{\perp}} \mathbf{\Pi}_{\hat{\mathcal{G}}_{\mathsf{P}_0}}\| = \|\mathbf{\Pi}_{\hat{\mathcal{G}}_{\mathsf{P}_1}^{\perp}} \mathbf{M}_{\hat{\mathsf{G}}_{r0}}\| = \|\mathbf{H}_{\hat{\mathsf{G}}_{r0}}\| = \inf_{\hat{\mathsf{Q}} \in \mathrm{H}^\infty_{\mathbb{D}}(\mathcal{B}_{\mathrm{L}^2_{\mathbb{H}}(\mathcal{U}), \mathrm{L}^2_{\mathbb{H}}(\mathcal{U})})} \|\hat{\mathsf{G}}_{r0} - \hat{\mathsf{G}}_{r1} \hat{\mathsf{Q}}\|_\infty,$$

which completes the proof. □

REFERENCES

[1] L. V. Ahlfors, *Complex Analysis*, Internat. Ser. Pure Appl. Math., McGraw–Hill, Singapore, 1979.

[2] B. A. Bamieh, J. B. Pearson, B. A. Francis, and A. R. Tannenbaum, *A lifting technique for linear periodic systems with applications to sampled-data control*, Systems Control Lett., 17 (1991), pp. 79–88.

[3] B. Bollobás, *Linear Analysis: An Introductory Course*, Cambridge Mathematical Textbooks, Cambridge University Press, Cambridge, UK, 1990.

[4] M. W. Cantoni, *Linear Periodic Systems: Robustness Analysis and Sampled-Data Control*, Ph.D. thesis, Department of Engineering, University of Cambridge, England, 1998.

[5] M. W. Cantoni and K. Glover, *Existence of right and left representations of the graph for linear periodically time-varying systems*, SIAM J. Control Optim., 38 (2000), pp. 786–802.

[6] T. Chen and B. A. Francis, *Optimal Sampled-Data Control Systems*, Comm. Control Engrg. Ser., Springer-Verlag, London, 1995.

[7] J. C. Doyle, T. T. Georgiou, and M. C. Smith, *The parallel projection operators of a nonlinear feedback system*, Systems Control Lett., 20 (1993), pp. 79–85.

[8] G. E. Dullerud, *Control of Uncertain Sampled-Data Systems*, Ph.D. thesis, University of Cambridge, England, 1994.

[9] A. K. El-Sakkary, *The gap metric: Robustness of stabilisation of feedback systems*, IEEE Trans. Automat. Control, 30 (1985), pp. 240–247.

[10] A. Feintuch, *The time-varying gap and coprime factor perturbations*, Math. Control Signals Systems, 8 (1995), pp. 352–37.

[11] C. Foiaş and A. Frazho, *The Commutant Lifting Approach to Interpolation Problems*, Oper. Theory Adv. Appl. 44, Birkhäuser, Berlin, 1990.

[12] C. Foiaş, T. T. Georgiou, and M. C. Smith, *Robust stability of feedback systems: A geometric approach using the gap metric*, SIAM J. Control Optim., 31 (1993), pp. 1518–1537.

[13] T. T. Georgiou, *On the computation of the gap metric*, Systems Control Lett., 11 (1988), pp. 253–257.
[14] T. T. Georgiou and M. C. Smith, *Optimal robustness in the gap metric*, IEEE Trans. Automat. Control, 35 (1990), pp. 673–686.
[15] T. T. Georgiou and M. C. Smith, *Graphs, causality and stabilizability: Linear, shift-invariant systems on* $L^2([0,\infty))$, Math. Control Signals Systems, 6 (1993), pp. 195–223.
[16] T. Kato, *Perturbation Theory for Linear Operators*, Springer-Verlag, New York, 1966.
[17] M. A. Krasnosel'skii, G. M. Vainikko, P. P. Zabreiko, Y. B. Rutitskii, and V. Y. Stetsenko, *Approximate Solution of Operator Equations*, Wolters-Noordhoff, Groningen, The Netherlands, 1972.
[18] N. K. Nikol'skii, *Treatise on the Shift Operator*, Springer-Verlag, New York, 1986.
[19] R. J. Ober and J. A. Sefton, *Stability of control systems and graphs of linear systems*, Systems Control Lett., 17 (1991), pp. 265–280.
[20] J. A. Sefton and R. J. Ober, *On the gap metric and coprime factor perturbations*, Automatica J. IFAC, 29 (1993), pp. 723–734.
[21] M. C. Smith, *On stabilisation and the existence of coprime factorisations*, IEEE Trans. Automat. Control, 34 (1989), pp. 1005–1007.
[22] B. Sz-Nagy and C. Foiaş, *Harmonic Analysis of Operators on Hilbert Space*, North–Holland, Amsterdam, 1970.
[23] M. Vidyasagar, *The graph metric for unstable plants and robustness estimates for feedback stability*, IEEE Trans. Automat. Control, 29 (1984), pp. 403–417.
[24] M. Vidyasagar, *Control System Synthesis: A Factorization Approach*, MIT Press, Cambridge, MA, 1985.
[25] M. Vidyasagar, H. Schneider, and B. A. Francis, *Algebraic and topological aspects of feedback stabilisation*, IEEE Trans. Automat. Control, 27 (1983), pp. 880–894.
[26] G. Vinnicombe, *Frequency domain uncertainty and the graph topology*, IEEE Trans. Automat. Control, 38 (1993), pp. 1371–1383.
[27] G. Vinnicombe, *Measuring the Robustness of Feedback Systems*, Ph.D. thesis, University of Cambridge, England, 1993.
[28] J. C. Willems, *The Analysis of Feedback Systems*, Res. Monographs 62, MIT Press, Cambridge, MA, 1971.
[29] N. Young, *The Nevanlinna-Pick problem for matrix-valued functions*, J. Operator Theory, 15 (1986), pp. 239–265.
[30] G. Zames, *Feedback and optimal sensitivity: Model reference transformations, multiplicative seminorms and approximate inverses*, IEEE Trans. Automat. Control, 26 (1981), pp. 301–320.
[31] G. Zames and A. K. El-Sakkary, *Unstable systems and feedback: The gap metric*, in Proceedings of the Allerton Conference on Communication, Control, and Computing, University of Illinois, Monticello, IL, 1980, pp. 380–385.

SIAM J. CONTROL OPTIM.
Vol. 38, No. 3, pp. 823–854

A HIGH-ORDER GENERALIZED LOCAL MAXIMUM PRINCIPLE*

URSZULA LEDZEWICZ† AND HEINZ SCHÄTTLER‡

Abstract. We derive a generalized local maximum principle which gives necessary conditions for optimality of abnormal trajectories in optimal control problems. In this theorem the multiplier associated with the objective is nonzero. Our results provide a complete hierarchy of primal constructions of high-order approximating cones (consisting of tangent directions for equality constraints, feasible directions for inequality constraints, and directions of decrease for the objective) and dual characterizations of empty intersection properties of these cones. The essential tool in our construction is a generalization of the Lyusternik theorem which describes the structure of high-order tangent directions to an operator equality constraint in a Banach space at a point where the operator is not regular (i.e., its Fréchet derivative is not onto). The results and procedure are illustrated with several examples.

Key words. optimal control, maximum principle, high-order approximations, nonregular equality constraints, abnormal trajectories

AMS subject classifications. Primary, 49K15, 93C10; Secondary, 49M05, 46N10

PII. S036301299833820X

1. Introduction.

We consider an optimal control problem on $\mathbb{R}^n$ with fixed terminal time and terminal constraints. The control set is a closed and convex subset of $\mathbb{R}^m$ with nonempty interior. We develop necessary conditions for optimality of *abnormal extremals*. Recall that an abnormal extremal is an admissible input-trajectory pair which satisfies the conditions of the Pontryagin maximum principle [28] for which the multiplier at the objective vanishes. Although the word "abnormal," which has its origins in the calculus of variations [7], seems to suggest that these types of extremals are an aberration, for optimal control problems this is not the case. The phenomenon is quite general and can be observed in a multitude of problems, and abnormal extremals cannot be excluded from optimality a priori. For instance, there exist optimal abnormal trajectories for the standard textbook problem of stabilizing the harmonic oscillator time optimally in minimum time, a simple time-invariant linear system. Recently there has been a strong interest in locally length minimizing abnormal geodesics in sub-Riemannian manifolds, and it has been found that abnormal minimizers are ubiquitous rather than exceptional. These have been investigated from an optimal control perspective for instance by Liu and Sussmann [23, 24] and by Agrachev and Sarychev [1, 2].

In the abnormal case conventional necessary conditions for optimality provide conditions which only describe the structure of the constraints. For example, if there are no control constraints, then these conditions only involve the equality constraint defined by the dynamics and terminal conditions as zero set of an operator $F : Z \to$

*Received by the editors May 4, 1998; accepted for publication (in revised form) May 10, 1999; published electronically March 8, 2000.

http://www.siam.org/journals/sicon/38-3/33820.html

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

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

Y between Banach spaces. If the Fréchet derivative $F'(z_*)$ at a point $z_* \in Z$ is onto (the so-called surjectivity or regularity condition), then the classical Lyusternik theorem describes the tangent space to the equality constraint at z_* as the kernel of $F'(z_*)$. Together with approximations of the inequality constraints (feasible cones) and the directions of decrease for the functional to be minimized (cones of decrease), Lagrange multiplier type necessary conditions for optimality in terms of a generalized Euler–Lagrange function can be derived (see, for instance [12]). If $F'(z_*)$ is not onto but closed (and this is always the case for the optimal control problem), then these multiplier rules can be satisfied trivially by choosing a multiplier which annihilates the image of $F'(z_*)$ and setting all other multipliers to zero. However, the corresponding necessary conditions are independent of the objective and describe only the structure of the constraint, yielding little information about the optimality of the abnormal trajectory. Although in some cases these conditions may be helpful, in general this situation is unsatisfactory, and it seems relevant to find other necessary conditions for optimality which are specifically tailored to abnormal processes and involve the objective in a nontrivial way.

The main reason why conventional necessary conditions fail for abnormal cases is that they, like the maximum principle, use only linear approximations for the equality constraint. This is woefully inadequate near abnormal points z_* when the kernel of the operator $F'(z_*)$ contains many directions which are *not* tangent to the equality constraint. Much of the difficulty in analyzing abnormal points in extremum problems can be traced back to the fact that the equality constraint is typically no longer a manifold near abnormal points. For illustrative purposes, simply consider the problem to minimize a functional $I : \mathbb{R}^n \to \mathbb{R}$, $z \mapsto I(z)$, subject to $F(z) = 0$, where $F : \mathbb{R}^n \to \mathbb{R}^k$. The trivial example to minimize a function $I(z_1, z_2)$ subject to $z_1 z_2 = 0$ illustrates the point perfectly well. Clearly the origin is a local minimizer if and only if zero is a local minimizer for both functions $I(z_1, 0)$ and $I(0, z_2)$. Hence, if one aims to develop necessary and/or sufficient conditions for optimality of abnormal extremals, it is imperative to analyze different branches of the zero-set of F. In an ideal situation one can hope that the subproblems of minimizing I over the branches are regular and thus can be analyzed with standard methods. But finding these branches precisely is at the heart of the matter.

In our research we have pursued this direction. Generalizing a result of Avakov [4, 5] in [21], we derived a high-order generalization of the classical Lyusternik theorem which for general $p \in \mathbb{N}$ describes the structure of p-order tangent directions to an operator equality constraint in a Banach space also for nonregular operators under a more general surjectivity assumption involving the first p derivatives of the operator. The underlying idea is simply to analyze the operator F further using higher-order derivatives and thus to replace linear approximations by polynomial approximations of degree p if the Fréchet derivative $F'(z_*)$ is not onto. Based on the results in [21], p-order tangent cones to the equality constraint can explicitly be calculated along critical directions which comprise the low-order terms. Combining these cones with standard constructions of high-order cones of decrease for the functional and high-order feasible cones to inequality constraints, all taken along critical directions, we can derive generalized necessary conditions for optimality for extremum problems in Banach spaces which allow incorporation of the objective with a nonzero multiplier. In [22] an abstract formulation of these results was presented for minimization problems in Banach spaces. Characteristic of these results is that they are parametrized by critical directions as it is *natural* near abnormal points. The main result of [22,

Theorem 7.1] gives a dual characterization for the empty intersection property of the various approximating cones along critical directions, but primal arguments using the cones themselves are often equally effective. The dependence of our results on critical directions, however, is viewed as a necessity of the problem rather than a weakness of the results.

In this paper we apply the abstract results of [22] to the optimal control problem, but we only consider the so-called weak or local version of the maximum principle. This result is weaker than the Pontryagin maximum principle [28] in the sense that the Pontryagin maximum principle asserts that the Hamiltonian of the control problem is indeed minimized over the control set at every time along the reference trajectory by the reference control. The local version only gives the necessary conditions for optimality for this property. However, it is well known how to use an argument of Dubovitskii to derive the Pontryagin maximum principle from the local version [12, Lecture 13], and in this sense the local maximum principle can be considered a first step in the derivation of the complete result. We will consider strong formulations of our results elsewhere. (A preliminary version is given in [20].)

Our research is part of a currently active research direction which aims at developing theories which are tailored to abnormal processes. Milyutin [25] introduced a method known as "weakening equality constraints" which has been developed further by Dmitruk [8]. In the papers by Avakov [4, 5] and by Arutyunov [3] both necessary and sufficient conditions for optimality of abnormal extremals are given based on quadratic approximations. In earlier papers we embedded Avakov's work into a Dubovitskii–Milyutin framework using the concept of second-order approximating cones [17] and derived a second-order generalized maximum principle [18]. Izmailov [15] also considers quadratic approximations, but for problems with inequality constraints. While mostly optimization related techniques are used in these papers, on a different level Agrachev and Sarychev [1] use differential geometric techniques to develop a theory of the second variation for abnormal extremals. They give both necessary and sufficient conditions for so-called corank 1 abnormal extremals (extremals for which there exists a unique multiplier) in terms of the Jacobi equation and related Morse indices and nullity theorems. While our results are not as far reaching, Agrachev and Sarychev's underlying necessary condition for optimality [1, Theorem 3.4] naturally fits into our framework, and we show how it can be rederived with our methods. Also, our results do apply to arbitrary abnormal extremals, and we give examples to show how our main result provides nontrivial conditions also in cases when the multiplier is no longer unique. In these examples the reference extremals have both normal and abnormal multipliers, but this is of no consequence for our approximations. In one of the examples we derive the classical accessory type necessary condition for optimality for the normal multiplier, i.e., $\mathcal{L}''_{xx}(\lambda_0, \lambda, z_*) \geq 0$, where $\mathcal{L}(\lambda_0, \lambda, z_*) = \lambda_0 I'(z_*) + \lambda F'(z_*)$ denotes the standard Lagrangian, but only on a subspace of $\ker F'(z_*)$ which consists of actual tangent directions. In general, it is easily seen that it is *not* a necessary condition for optimality of z_* that $\mathcal{L}''_{xx}$ is positive semidefinite on $\ker F'(z_*)$ when the multiplier is not unique. Second-order necessary conditions for optimality in the type of accessory problem results without normality assumptions have first been given by Gilbert and Bernstein [11]. Stefani and Zezza [29] also derive results without making normality assumptions.

This brief survey of the literature given focuses on research on the maximum principle related to normality conditions. Obviously the Pontryagin maximum principle has literally hundreds of extensions too numerous to be discussed here. We want

to mention only a few, among them a recently developed general formulation of the theorem by Sussmann which is valid under very weak differentiability assumptions and applies to many even classical problems for which the standard version is inadequate [30]. High-order approximations also have a long history in connection with the maximum principle and we mention only the papers by Krener [16] and Gabasov and Kirillova [9] for the smooth case or the recent papers by Páles and Zeidan [26, 27] under nonsmooth assumptions.

The main result of this paper, a p-order local maximum principle has been announced in [19], but no proof is given there. Also, the formulation in [19] allows for p-order abnormality in the sense that it includes the more degenerate case when the multiplier for the objective is allowed to vanish. Here we give conditions under which this multiplier will be nonzero which guarantee the nontriviality of this p-order extension. Our results provide a complete hierarchy of primal constructions of high-order approximating directions and dual characterizations of empty intersection properties of these approximating cones. Thus the theory developed in [22] and here gives necessary conditions for optimality for increasingly more degenerate structures.

2. Problem formulation. We consider an optimal control problem in Bolza form with fixed terminal time: (OC) Minimize the functional

$$I(x,u) = \int_0^T L(x(t),u(t),t)dt + \ell(x(T)) \tag{2.1}$$

subject to the constraints

$$\dot{x}(t) = f(x(t),u(t),t), \tag{2.2}$$

$$x(0) = 0, \qquad q(x(T)) = 0, \tag{2.3}$$

$$u(\cdot) \in \mathcal{U} = \{u \in L^r_\infty(0,T) : \ u(t) \in U \text{ almost everywhere (a.e.)}\}. \tag{2.4}$$

The terminal time T is fixed and we make the following *regularity assumptions* on the data: $L : \mathbb{R}^n \times \mathbb{R}^m \times [0,T] \to \mathbb{R}$ and $f : \mathbb{R}^n \times \mathbb{R}^m \times [0,T] \to \mathbb{R}^n$ are C^∞ in (x,u) for every $t \in [0,T]$; both functions and their derivatives are measurable in t for every (x,u) and the functions and all partial derivatives are bounded on compact subsets of $\mathbb{R}^n \times \mathbb{R}^m \times [0,T]$; $\ell : \mathbb{R}^n \to \mathbb{R}$ and $q : \mathbb{R}^n \to \mathbb{R}^k$ are C^∞ and the rows of the Jacobian matrix q_x (i.e., the gradients of the equations defining the terminal constraint) are linearly independent; $U \subset \mathbb{R}^m$ is a closed and convex set with nonempty interior.

We model this problem in the framework of optimization theory as a minimization problem in a Banach space under equality and inequality constraints. Let $W^n_{11}(0,T)$ denote the Banach space of all absolutely continuous functions $x : [0,T] \to \mathbb{R}^n$ with norm $|x| = ||x(0)|| + \int_0^T ||\dot{x}(s)||ds$ and let $\overline{W}^n_{11}(0,T) = W^n_{11}(0,T) \cap \{x \in W^n_{11}(0,T)$: $x(0) = 0\}$. Then the problem is to minimize the functional I over a class $\mathcal{A}$ of input-trajectory pairs $(x,u) \in \overline{W}^n_{11}(0,T) \times L^m_\infty(0,T)$ which is defined by equality constraints and the convex inequality constraint $u \in \mathcal{U}$. The equality constraints can be modelled as $\mathcal{F} = \{(x,u) \in \overline{W}^n_{11}(0,T) \times L^m_\infty(0,T)$: $F(x,u) = 0\}$, where F is the operator

$$\begin{aligned} F &: \overline{W}^n_{11}(0,T) \times L^m_\infty(0,T) \to \overline{W}^n_{11}(0,T) \times \mathbb{R}^k, \\ &(x,u) \mapsto F(x,u) = \left(x(\cdot) - \int_0^{(\cdot)} f(x(s),u(s),s)ds, \ q(x(T)) \right). \end{aligned} \tag{2.5}$$

It is easy to see that the operator F has continuous Fréchet derivatives of arbitrary order. For instance,

$$F'(x,u)(\eta,\xi) = \left(\eta(\cdot) - \int_0^{(\cdot)} f_x(x,u,s)\eta + f_u(x,u,s)\xi ds, q_x(x(T))\eta(T)\right) \tag{2.6}$$

acting on $(\eta,\xi) \in \overline{W}^n_{11}(0,T) \times L^m_\infty(0,T)$. All partial derivatives of f are evaluated along a reference input-trajectory pair $(x,u) \in \mathcal{A}$. The formulas for higher-order derivatives are given by equally straightforward multilinear forms.

3. Regularity of the operator. We first describe the image of the operator $F'(x_*,u_*)$ for a reference input-trajectory pair (x_*,u_*). For notational simplicity, let

$$A(t) = f_x(x_*(t),u_*(t),t), \qquad B(t) = f_u(x_*(t),u_*(t),t), \tag{3.1}$$

and denote the fundamental matrix of the variational equation by $\Phi(t,s)$, i.e.,

$$\frac{\partial}{\partial t}\Phi(t,s) = A(t)\Phi(t,s), \qquad \Phi(s,s) = \mathrm{Id}. \tag{3.2}$$

Furthermore, let $R \subset \mathbb{R}^n$ denote the reachable subspace of the linearized system

$$\dot{h}(t) = A(t)h + B(t)v, \quad h(0) = 0, \tag{3.3}$$

at time T, i.e.,

$$R = \left\{\eta(T) = \int_0^T A(s)\eta(s) + B(s)\xi(s)ds = \int_0^T \Phi(T,s)B(s)\xi(s)ds : \xi \in L^m_\infty(0,T)\right\}. \tag{3.4}$$

It is well known that R is a linear subspace of $\mathbb{R}^n$ and that $R = \mathbb{R}^n$ if and only if (3.3) is completely controllable. In general we have the following.

LEMMA 3.1.

$$\operatorname{Im} F'(x_*,u_*) = \left\{(a,b) \in \overline{W}^n_{11}(0,T) \times \mathbb{R}^k : b \in q_x(x_*(T))\left(\int_0^T \Phi(T,s)\dot{a}(s)ds + R\right)\right\}. \tag{3.5}$$

$\operatorname{Im} F'(x_*,u_*)$ is closed and of finite codimension given by

$$\operatorname{co\,dim} \operatorname{Im} F'(x_*,u_*) = k - \dim\left[R \cap \ker q_x(x_*(T))^\perp\right].$$

In particular, $F'(x_*,u_*)$ is onto if and only if

$$\ker q_x(x_*(T)) + R = \mathbb{R}^n. \tag{3.6}$$

Proof. Given an arbitrary function $a \in \overline{W}^n_{11}(0,T)$, the unique solution $\overline{h} \in \overline{W}^n_{11}(0,T)$ to the equation

$$\overline{h}(t) - \int_0^t A(s)\overline{h}(s)ds = a(t) = \int_0^t \dot{a}(s)ds$$

is given by $\overline{h}(t) = \int_0^t \Phi(t,s)\dot{a}(s)ds$. This function can be superimposed with any other solution h to (3.3), say

$$h(t) = \int_0^t A(s)h(s) + B(s)v(s)ds,$$

without changing the first component of $\operatorname{Im} F'(x_*, u_*)$. Thus, if $\eta = \overline{h} + h$ and $\xi = v$, then

$$F'(x_*, u_*)(\eta, \xi) = \left(a, q_x(x_*(T))\left(\int_0^T \Phi(T,s)\dot{a}(s)ds + h(T)\right)\right)$$

and $h(T) \in R$. Conversely, if $(a, b) = F'(x_*, u_*)(\eta, \xi)$, then $b = q_x(x_*(T))\eta(T)$, and

$$\begin{aligned}\eta(T) - \int_0^T \Phi(T,s)\dot{a}(s)ds &= \int_0^T [A(t)\eta(t) + B(t)\xi(t)]dt + a(T) \\ &\quad - \int_0^T [A(t)\overline{h}(t) + \dot{a}(t)]dt \\ &= \int_0^T A(t)[\eta(t) - \overline{h}(t)] + B(t)\xi(t)dt \in R,\end{aligned}$$

verifying (3.5). It follows that $\operatorname{Im} F'(x_*, u_*)$ is the direct sum of the closed subspace

$$\widetilde{W} = \left\{\left(a, q_x(x_*(T))\int_0^T \Phi(T,s)\dot{a}(s)ds\right) \in \overline{W}^n_{11}(0,T) \times \mathbb{R}^k : a \in \overline{W}^n_{11}(0,T)\right\},$$

which is isomorphic to $\overline{W}^n_{11}(0,T)$ and the finite-dimensional subspace $\widetilde{R} = \{0\} \times q_x(x_*(T))R$. Thus $\operatorname{Im} F'(x_*, u_*)$ is closed and of finite codimension given by the codimension of $q_x(x_*(T))R$. Since the linear map $q_x(x_*(T)) : \mathbb{R}^n \to \mathbb{R}^k$ is an isomorphism when it is restricted to the orthogonal complement of $\ker q_x(x_*(T))$, it follows that the codimension is given by

$$\operatorname{co\,dim} \operatorname{Im} F'(x_*, u_*) = k - \dim(R \cap \ker q_x(x_*(T))^\perp). \tag{3.7}$$

In particular, $F'(x_*, u_*)$ is onto if and only if R contains $\ker \psi_x(x_*(T))^\perp$ which is equivalent to (3.6). □

The operator F gives the state-space representation of the dynamical system for the optimal control problem. Equivalently, and this is used for instance by Agrachev and Sarychev in [1], one could use the input-output map

$$Q : L^m_\infty(0,T) \to \mathbb{R}^k, \qquad u \mapsto q(x(T)), \tag{3.8}$$

where x is the solution to $\dot{x} = f(x, u(t), t)$, $x(0) = 0$. These formulations are equivalent and it is easily seen that the codimension of $\operatorname{Im} Q'(u)$ is given by (3.7). The input-output map has the advantage that it is clear that the image of $Q'(u)$ is closed, in fact finite dimensional. The state-representation leads to calculations which, although not intrinsic, are quite transparent and require only simple differentiations of data directly given.

It is well known how to characterize that (3.3) is not completely controllable. If h is a solution of (3.3) corresponding to control v and λ is a solution to the corresponding adjoint equation

$$\dot{\lambda}(t) = -\lambda(t)A(t), \tag{3.9}$$

(which we write as row-vector $\lambda \in (\mathbb{R}^n)^*$ denoting the space of row-vectors in $\mathbb{R}^n$ by $(\mathbb{R}^n)^*$), then

$$\lambda(T)h(T) = \int_0^T \dot{\lambda}(t)h(t) + \lambda(t)\dot{h}(t)dt = \int_0^T \lambda(t)B(t)v(t)dt. \tag{3.10}$$

Hence, if (3.3) is not completely controllable, then any vector $\lambda(T) \perp R$ defines a solution λ of (3.9) which satisfies $\lambda(t)B(t) \equiv 0$ on $[0,T]$, and, conversely, any solution λ with this property satisfies $\lambda(T) \perp R$. Thus the codimension of R is given by the number of linearly independent solutions to $\dot{\lambda}(t) = -\lambda(t)A(t)$ which satisfy $\lambda(t)B(t) \equiv 0$ on $[0,T]$. Consequently, we have the following well-known characterization.

PROPOSITION 3.2. *The codimension of $F'(x_*, u_*)$ is given by the number of linearly independent solutions to $\dot{\lambda}(t) = -\lambda(t)A(t)$ which satisfy $\lambda(t)B(t) \equiv 0$ on $[0,T]$ and for which $\lambda(T)$ is orthogonal to $\ker q_x(x_*(T))$.*

The nonregularity of the operator F at an input-trajectory pair (x_*, u_*) can be characterized in terms of the conditions of the so-called local maximum principle. Let

$$H(\lambda_0, \lambda, x, u, t) = \lambda_0 L(x, u, t) + \lambda f(x, u, t) \tag{3.11}$$

be the Hamiltonian for the control problem. If the input-trajectory pair (x_*, u_*) is optimal for problem (OC), then the local maximum principle [12] states that there exists a constant $\lambda_0 \geq 0$, an absolutely continuous function $\lambda : [0,T] \to (\mathbb{R}^n)^*$, and a row-vector $\nu \in (\mathbb{R}^k)^*$ such that the following conditions hold:

1. nontriviality of the multipliers: $(\lambda_0, \lambda(t)) \neq 0$ for all $t \in [0,T]$;
2. adjoint equation:

$$\begin{aligned} \dot{\lambda}(t) &= -\lambda_0 L_x(x_*(t), u_*(t), t) - \lambda(t) f_x(x_*(t), u_*(t), t) \\ &= -H_x(\lambda_0, \lambda(t), x_*(t), u_*(t), t); \end{aligned} \tag{3.12}$$

3. transversality condition:

$$\lambda(T) = \lambda_0 \ell_x(x_*(T)) + \nu q_x(x_*(T)); \tag{3.13}$$

4. local minimum condition:

$$\langle H_u(\lambda_0, \lambda(t), x_*(t), u_*(t), t), v - u_*(t) \rangle \geq 0 \qquad \text{for all } v \in U. \tag{3.14}$$

We call input-trajectory pairs (x_*, u_*) for which multipliers λ_0, λ, and ν exist such that these conditions are satisfied *(weak) extremals*. If $\lambda_0 > 0$, then it is possible to normalize $\lambda_0 = 1$ and the extremal is called normal while extremals with $\lambda_0 = 0$ are called *abnormal*.

PROPOSITION 3.3. *The operator F is nonregular at $\Gamma = (x_*, u_*)$ if and only if Γ is an abnormal weak extremal which satisfies $H_u(0, \lambda(t), x_*(t), u_*(t), t) \equiv 0$ on $[0,T]$.*

Proof. If F is nonregular at $\Gamma = (x_*, u_*)$, then there exists a nontrivial solution λ to the equation $\dot{\lambda}(t) = -\lambda(t)A(t)$ which satisfies $\lambda(t)B(t) \equiv 0$ on $[0,T]$ and $\lambda(T)$ is orthogonal to $\ker q_x(x_*(T))$. Since the rows of $q_x(x_*(T))$ are linearly independent, it follows that $\lambda(T)$ is a linear combination of the rows of $q_x(x_*(T))$, i.e., there exists a row-vector $\nu \in (\mathbb{R}^k)^*$ such that $\lambda(T) = \nu q_x(x_*(T))$. If we choose $\lambda_0 = 0$, it follows that Γ is an abnormal weak extremal which satisfies $H_u(0, \lambda(t), x_*(t), u_*(t), t) \equiv 0$ on $[0,T]$. Conversely, if Γ is an abnormal weak extremal which satisfies $H_u(0, \lambda(t), x_*(t), u_*(t), t) \equiv 0$ on $[0,T]$, then λ satisfies the conditions of Proposition 3.2 since $\nu q_x(x_*(T)$ is orthogonal to $\ker q_x(x_*(T))$. ∎

For variational type problems where the control set is the full space, $U = \mathbb{R}^m$, or more generally when the extremal control takes values in the interior of the control set, the condition $H_u \equiv 0$ is automatically satisfied and nonregular equality constraints thus correspond to abnormal extremals. For example, if the problem is also in Mayer form, i.e., $L \equiv 0$, then every terminal point x_* where $\ell_x(x_*)$ is a linear combination of the rows of $q_x(x_*)$ generates an abnormal extremal with nonregular equality constraint.

4. High-order directional derivatives. We introduce a formalism to describe higher derivatives [21]. Let $F : Z \to Y$ be an operator between Banach spaces which is sufficiently often continuously Fréchet differentiable in a neighborhood of $z_* \in Z$, and consider the Taylor expansion of F along a curve $\gamma(\varepsilon) = z_* + \sum_{i=1}^m \varepsilon^i h_i$. We have

$$F\left(z_* + \sum_{i=1}^m \varepsilon^i h_i\right) = F(z_*) + \sum_{i=1}^m \varepsilon^i \nabla^i F(z_*)(h_1, \dots, h_i) + \widetilde{r}(\varepsilon), \tag{4.1}$$

where

$$\nabla^i F(z_*)(h_1, \dots, h_i) \doteq \sum_{r=1}^i \frac{1}{r!} \left(\sum_{j_1+\cdots+j_r=i} F^{(r)}(z_*)(h_{j_1}, \dots, h_{j_r}) \right) \tag{4.2}$$

and $\widetilde{r}(\varepsilon)$ is a function of order $o(\varepsilon^m)$ as $\varepsilon \to 0$. Note that $\nabla^i F(z_*)(h_1, \dots, h_i)$ simply collects the ε^i terms in this expansion. These terms, which we call the *ith-order directional derivatives of F along the sequence* $H_i = (h_1 \dots, h_i)$, $1 \le i \le m$, are easily calculated by straightforward Taylor expansions. For example,

$$\nabla^1 F(z_*)(H_1) = F'(z_*)h_1, \quad \nabla^2 F(z_*)(H_2) = F'(z_*)h_2 + \frac{1}{2}F''(z_*)(h_1, h_1),$$
$$\nabla^3 F(z_*)(H_3) = F'(z_*)h_3 + F''(z_*)(h_1, h_2) + \frac{1}{6}F'''(z_*)(h_1, h_1, h_1),$$

and so on. The higher-order directional derivative $\nabla^i F(z_*)$ is homogeneous of degree i in the directions in the sense that

$$\nabla^i F(z_*)(\varepsilon h_1, \dots, \varepsilon^i h_i) = \varepsilon^i \nabla^i F(z_*)(h_1, \dots, h_i). \tag{4.3}$$

In particular, no indices j_1 and j_2 with $j_1 + j_2 > i$ can occur together as arguments in any of the terms in $\nabla^i F(z_*)$. Thus all the vectors h_j whose index satisfies $2j > i$ appear linearly in $\nabla^i F(z_*)$ and are multiplied by terms which are homogeneous of degree $i - j$. In fact, there exist linear operators $G_k = G_k[F](z_*; H_{k-1})$, $k \in \mathbb{N}$, depending on the derivatives up to order k of F in the point z_* (the k-jet of F in z_*) and on the vectors $H_{k-1} = (h_1, \dots, h_{k-1})$, which describe the contributions of these components. We have $G_1[F](z_*) = F'(z_*)$ and in general $G_k = G_k[F](z_*; H_{k-1})$: $Z \to Y$, $v \longmapsto G_k(v)$ is given by

$$G_k[F](z_*; H_{k-1})(v) = \sum_{r=1}^{k-1} \frac{1}{r!} \left(\sum_{j_1+\cdots+j_r=k-1} F^{(r+1)}(z_*)(h_{j_1}, \dots, h_{j_r}, v) \right). \tag{4.4}$$

These operators $G_k[F](z_*; H_{k-1})$ are simply the Fréchet derivatives of the $(k-1)$th directional derivative of F at z_* along H_{k-1}. Note that these terms are homogeneous

of degree $k-1$. For simplicity of notation we often suppress the arguments. For example, we write

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

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

$$\nabla^{p-1+i}F(z_*)(H_{p-1+i}) = \sum_{k=1}^{i} G_k[F](z_*; H_{k-1})h_{p+i-k} + R_{p-1+i}[F](z_*; H_{p-1}). \tag{4.5}$$

Here, amongst all terms which are homogeneous of degree $p-1+i$, the sum gives the terms which contain one of the vectors $h_p, \dots, h_{p-1+i}$, and the remainder R combines all other terms which only include vectors of index $\le p-1$. Similar to the operators G_k, the remainders R depend on the $(p-1+i)$-jet of the operator F at z_*, but if the map is clear we omit it in the notation. The general structure of these remainders is given by

$$R_q[F](z_*; H_\ell) = \sum_{r=2}^{q} \frac{1}{r!} \left(\sum_{\substack{j_1+\cdots+j_r=q,\\ 1\le j_k\le \ell,\ 1\le k\le r}} F^{(r)}(z_*)(h_{j_1}, \dots, h_{j_r}) \right). \tag{4.6}$$

Thus $R_q(H_\ell)$ consists of the terms which are homogeneous of degree q, but only involve vectors from H_ℓ. For example,

$$R_4[F](z_*; H_2) = \frac{1}{2}F''(z_*)(h_2, h_2) + \frac{1}{2}F^{(3)}(z_*)(h_1, h_1, h_2) + \frac{1}{24}F^{(4)}(z_*)(h_1, h_1, h_1, h_1).$$

Note that the remainders only have contributions from derivatives of at least order two. These operators allow to formalize high-order approximations to an equality constraint at nonregular points [22]. In the next section we give the conditions under which such approximations exist.

5. Critical directions. We first describe the set of critical directions along which high-order tangent approximations to the equality constraint $\mathcal{F}$ can be set up. For a given admissible process $z_* = (x_*, u_*) \in \mathcal{A}$ and a finite sequence $H_{p-1} = (h_1, \dots, h_{p-1}) \in Z^{p-1}$, let

$$Y_i = \sum_{k=1}^{i} \operatorname{Im} G_k[F](z_*; H_{k-1}), \quad i = 1, \dots, p. \tag{5.1}$$

We need to impose the following conditions:

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

$$\nabla^i F(z_*)(H_i) = 0 \quad \text{for all } \ i = 1, \dots, p-1; \tag{5.2}$$

(ii) and the compatibility conditions

$$R_{p-1+i}[F](z_*; H_{p-1}) \in Y_i, \quad i = 1, \dots, p-1, \tag{5.3}$$

are satisfied.

In these equations all partial derivatives of f are evaluated along the reference trajectory. Both conditions (i) and (ii) are necessary for the existence of a p-order tangent vector along H_{p-1}. For instance, if $p = 2$, then (i) says $F'(z_*)h_1 = 0$ and (ii) states that $F''(z_*)(h_1, h_1) \in \operatorname{Im} F'(z_*)$. If the second condition does not hold, then the quadratic term in

$$F(z_* + \varepsilon h_1 + \varepsilon^2 h_1 + \cdots) = F(z_*) + \varepsilon F'(z_*)h_1 + \varepsilon^2 \left(F'(z_*)h_2 + \frac{1}{2} F''(z_*)(h_1, h_1) \right) + \cdots$$

cannot be made zero with any choice of $h_2 \in Z$. Conditions (i) and (ii) are sufficient for the existence of p-order approximations along H_{p-1} under the following regularity condition.

DEFINITION 5.1. *Let $F : Z \to Y$ be an operator between Banach spaces. We say the operator F is* p-regular *at z_* in direction of the sequence $H_{p-1} \in Z^{p-1}$ if the following conditions are satisfied:*

(A1) *$F : Z \to Y$ is $(2p-1)$ times continuously Fréchet differentiable in a neighborhood of z_*;*

(A2) *the subspaces Y_i, $i = 1, \ldots, p$, are closed;*

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

$$\mathcal{G}_p : Z \to Y_1 \times Y_2/Y_1 \times \cdots \times Y/Y_{p-1} \qquad v \mapsto \mathcal{G}_p(v) = (G_1(v), \pi_1 G_2(v), \ldots, \pi_{p-1} G_p(v)), \tag{5.4}$$

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

In the sense of this definition, 1-regularity corresponds to the classical Lyusternik condition while 2-regularity is similar to Avakov's definition [5]. It is shown in [21, Theorem 1] and [22, Corollary 3.3] that the set of all directions h_p for which $H_p = (h_1, \ldots, h_{p-1}; h_p)$ is tangent to $\{z \in Z : F(z) = 0\}$ at z_* is a nonempty affine subspace if F is p-regular in direction of $H_{p-1} = (h_1, \ldots, h_{p-1})$ and H_{p-1} satisfies (i) and (ii).

Example (see [1]). Suppose $\operatorname{Im} F'(z_*)$ has codimension 1 and let λ_* be the (up to nonzero multiples) unique annihilator of $\operatorname{Im} F'(z_*)$, $\lambda_* F'(z_*)h = 0$ for all $h \in Z$. If the quadratic form $\lambda_* F''(z_*)(h, h)$ is indefinite on $\ker F'(z_*)$, then there exist linearly independent vectors h_+ and h_- in $\ker F'(z_*)$ such that F is 2-regular in direction of $h_\pm$ and such that the compatibility condition $F''(z_*)(h_\pm, h_\pm) \in \operatorname{Im} F'(z_*)$ is satisfied.

For, since $\lambda_* F''(z_*)(h, h)$ is indefinite on $\ker F'(z_*)$, there exist vectors h_α and h_β in $\ker F'(z_*)$ so that $\lambda_* F''(z_*)(h_\alpha, h_\alpha) = -1$ and $\lambda_* F''(z_*)(h_\beta, h_\beta) = 1$. If we take nontrivial convex combinations of h_α and h_β and of h_α and $-h_\beta$, it follows that there exist two linearly independent directions h_+ and h_- in $\ker F'(z_*)$ such that $\lambda_* F''(z_*)(h_\pm, h_\pm) = 0$. Since λ_* is the unique annihilator of $\operatorname{Im} F'(z_*)$, this is equivalent to $F''(z_*)(h_\pm, h_\pm) \in \operatorname{Im} F'(z_*)$. Hence $h_\pm$ satisfy conditions (i) and (ii) for $p = 2$. Furthermore, F is 2-regular along $h_\pm$ if there exists a vector $h \in Z$ (not necessarily in $\ker F'(z_*)$) such that $\pi G_2(z_*)(h_\pm, h) \neq 0$, where $\pi : Z \to Y/\operatorname{Im} F'(z_*)$ is the canonical projection. But again, since $\operatorname{Im} F'(z_*)$ has codimension 1, this is equivalent to $\lambda_* F''(z_*)(h_\pm, h) \neq 0$. This condition is satisfied since indeed we have $\lambda_* F''(z_*)(h_+, h_-) \neq 0$. For, if this were not true, then it would follow that $\lambda_* F''(z_*)(h, h) = 0$ for all h in the linear span of h_+ and h_-. But this is not possible since h_α and h_β lie in this span. □

It therefore follows from [22, Corollary 3.3] that h_+ and h_- in $\ker F'(z_*)$ are tangent directions to the set $\{z \in Z : F(z) = 0\}$. (In fact, except for scalar multiples

these are the only tangent directions in the two-dimensional subspace of $\ker F'(z_*)$ spanned by h_α and h_β.) Hence, if the objective is to minimize some functional $I : Z \to \mathbb{R}$ over $\{z \in Z : F(z) = 0\}$, then it is a necessary condition for optimality of z_* that $I'(z_*)h_\pm = 0$. This in its essence (combined with a perturbation argument) is the reasoning of Agrachev and Sarychev, who proved the result below.

PROPOSITION 5.2 (see [1, Theorem 3.4]). *Consider the problem to minimize a functional $I : Z \to \mathbb{R}$ over the set $\{z \in Z : F(z) = 0\}$. Suppose $\operatorname{Im} F'(z_*)$ has codimension 1 and let λ_* be a nonzero annihilator of $\operatorname{Im} F'(z_*)$. If the quadratic form $\lambda_* F''(z_*)(h,h)$ is indefinite on $\ker F'(z_*)$, then it is a necessary condition for optimality of z_* that the codimension of $\operatorname{Im}(F'(z_*), I'(z_*))$ is 2, i.e., that $I'(z_*)h = 0$ for all $h \in \ker F'(z_*)$.*

As described above, this result naturally fits into our framework of high-order approximations near nonregular points. In a certain sense these are still the least degenerate cases.

We now proceed to include the critical directions for the objective I. We will focus on the least degenerate critical case and therefore make the following assumption:

(iii) $I'(z_*)$ is not identically zero and $\nabla^i I(z_*)(H_i) = 0$ for $i = 1, \dots, p-1$.

Assuming $I'(z_*) \neq 0$ excludes some cases in which Theorem 6.1 below could be satisfied with a trivial choice of multipliers. However, rather than including this as a degenerate case, in this case the better way is to analyze the objective further by considering higher derivatives of I. This can be done easily, but the formulas for the cones of decrease and their duals (which enter into the formulation of Theorem 6.1) become different. Here we just give the nondegenerate formulation. The other assumption that the first $p-1$ directional derivatives vanish is directly tied in with optimality. If there exists a first nonzero directional derivative $\nabla^j I(z_*)(H_j)$ with $j < i$ which is positive, then z_* indeed is a local minimum for any curve $z(\varepsilon) = z_* + \sum_{i=1}^{p} \varepsilon^i h_i + o(\varepsilon^p)$, $\varepsilon > 0$, and none of the directions H_{p-1} is of any use in improving the value. In other words, the p-order cone of decrease along H_{p-1} is empty and the empty intersection property of approximating cones holds trivially. Note that we restrict to $\varepsilon \geq 0$ since we also want to include inequality constraints. On the other hand, if $\nabla^j I(z_*)(H_j) < 0$, then H_j is indeed a direction of decrease, and arbitrary high-order extensions of this sequence will give better values. Thus the p-order cone of decrease will be the full space. In this case our conditions below apply, but they simplify in the sense that the multiplier ν_0 in Theorem 6.1 must vanish. The truly critical case is when all these derivatives vanish.

We also need to define the critical directions for the inequality constraint $\mathcal{U}$ in the optimal control problem. More generally, we define a p-order feasible set to an inequality constraint in a Banach space.

DEFINITION 5.3. *Let $S \subset Z$ be a subset with nonempty interior. We call v a p-order feasible vector for S at z_* in direction of $H_{p-1} = (h_1, \dots, h_{p-1}) \in Z^{p-1}$ if there exist an $\varepsilon_0 > 0$ and a neighborhood V of v so that for all $0 < \varepsilon \leq \varepsilon_0$,*

$$z_* + \sum_{i=1}^{p-1} \varepsilon^i h_i + \varepsilon^p V \subset S.$$

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

It follows from this definition that $FS^{(p)}(S; z_*, H_{p-1})$ is open (since any vector in the neighborhood V of v also lies in $FS^{(p)}(S; z_*, H_{p-1})$). It is also clear that

$FS^{(p)}(S; z_*, H_{p-1})$ is convex, if S is. Furthermore, if $h_j \in FS^{(j)}(S; z_*, H_{j-1})$ for some integer $j < p$, then any vector v is allowed as a p-order feasible direction and thus trivially $FS^{(p)}(S; z_*, H_{p-1}) = X$.

For the optimal control problem and $H_{p-1} = ((\eta_1, \xi_1); \dots, (\eta_{p-1}, \xi_{p-1}))$ let $V_{p-1} = (\xi_1, \dots, \xi_{p-1}) \in L^m_\infty(0,1)^p$ denote the sequence of controls. Then the critical feasible directions for the convex inequality constraint $\mathcal{U}$ in $L^m_\infty(0,1)$ consist of all H_{p-1} for which

(iv) $FS^{(p)}(\mathcal{U}; u_*, V_{p-1})$ is nonempty.

DEFINITION 5.4. *We call a direction H_{p-1} a* p-regular critical direction *for the extremum problem at z_* if the operator F is p regular at z_* along H_{p-1} and if conditions* (i)–(iv) *are satisfied.*

6. The *p*-order local maximum principle. Theorem 6.1 below gives a generalized p–order version of the maximum principle obtained from a dual characterization of the fact that if (x_*, u_*) is optimal, then the p-order tangent cones to the set $\{F = 0\}$, the p-order feasible cone to $\mathcal{U}$ and the p-order cone of decrease for the functional I cannot intersect. This theorem generalizes a result which was derived from quadratic approximations in [18, Theorem 4.1] to the case of general p-order approximations. Notice that we write covectors like ψ as row vectors. This is consistent with a multiplier interpretation of the adjoint variable. We also denote partial derivatives by subscripts. For instance, if $\nabla^i f(H_i)$ denotes the ith directional derivative of $f = f(x,u,t)$ with respect to the sequence H_i, then $\left(\nabla^i f(H_i)\right)_x$ denotes its partial derivative in x. For example, suppose $H_1 = (\eta_1, \xi_1)$. Then

$$\nabla^1 f(H_1) = f_x(x,u,t)\eta_1 + f_u(x,u,t)\xi_1,$$

and thus

$$\left(\nabla^1 f(H_1)\right)_x = f_{xx}(x,u,t)\eta_1 + f_{ux}(x,u,t)\xi_1$$

and

$$\left(\nabla^1 f(H_1)\right)_u = f_{xu}(x,u,t)\eta_1 + f_{uu}(x,u,t)\xi_1.$$

THEOREM 6.1 (p-order local maximum principle). *Suppose the admissible process (x_*, u_*) is optimal for the optimal control problem (OC). Then for every p-regular critical direction H_{p-1} there exist a number $\nu_0 = \nu_0(H_{p-1}) \geq 0$, vectors $a_i = a(H_{p-1}) \in (\mathbb{R}^k)^*$, $i = 0, 1, \dots, p-1$, and absolutely continuous functions $\psi(\cdot) = \psi(H_{p-1})(\cdot)$ and $\rho_i(\cdot) = \rho_i(H_{p-1})(\cdot)$, $i = 1, \dots, p-1$, from $[0,T]$ into $(\mathbb{R}^n)^*$, which satisfy the following conditions along the optimal trajectory $(x_*(t), u_*(t), t)$:*

(a) *nontriviality condition: ν_0 and the functional $\lambda : L^m_\infty(0,T) \to \mathbb{R}$ given by*

$$\lambda(\xi) = \int_0^T \left\langle \nu_0 L_u + \psi f_u + \sum_{i=1}^{p-1} \rho_i \left(\nabla^i f(H_i)\right)_u, \xi \right\rangle dt \tag{6.1}$$

do not both vanish identically.

(b) *extended adjoint equation:*

$$\dot{\psi}(t) = -\nu_0 L_x - \psi(t) f_x - \sum_{i=1}^{p-1} \rho_i(t) \left(\nabla^i f(H_i)\right)_x \tag{6.2}$$

with terminal condition

$$\psi(T) = \nu_0 \ell_x(x_*(T)) + a_0 q_x(x_*(T)) + \sum_{i=1}^{p-1} a_i \left(\nabla^i q(x_*(T); H_i)\right)_x . \tag{6.3}$$

(c) *orthogonality conditions on the additional multipliers: The functions* $\rho_i(\cdot)$, $i = 1, \dots, p-1$, *satisfy*

$$\dot\rho_i(t) = -\rho_i(t) f_x, \quad \rho_i(t) f_u \equiv 0, \quad \rho_i(T) = a_i q_x(x_*(T)), \tag{6.4}$$

and for $j = 1, \dots, i-1$, *the following conditions are satisfied for a.e.* $t \in [0,T]$*:*

$$\rho_i(t) \left(\nabla^j f(H_j)\right)_x = 0, \quad \rho_i(t) \left(\nabla^j f(H_j)\right)_u = 0, \quad a_i \left(\nabla^j q((x_*(1); H_j)\right)_x = 0. \tag{6.5}$$

(d) *separation condition: for all* $\xi \in FS^{(p)}(\mathcal{U}; u_*, V_{p-1})$, *we have that*

$$\begin{aligned} 0 \le{}& \nu_0 R_p[\ell](H_{p-1}) + a_0 R_p[q](H_{p-1}) + \sum_{i=1}^{p-1} a_i R_{p+i}[q](H_{p-1}) \\ &+ \int_0^T \left\langle \nu_0 L_u + \psi f_u + \sum_{i=1}^{p-1} \rho_i(t) \left(\nabla^i f(H_i)\right)_u , \xi \right\rangle dt \\ &+ \int_0^T \left(\nu_0 R_p[L](H_{p-1}) + \psi(t) R_p[f](H_{p-1}) + \sum_{i=1}^{p-1} \rho_i(t) R_{p+i}[f](H_{p-1}) \right) dt. \end{aligned} \tag{6.6}$$

COROLLARY 6.2. *The separation condition* (d) *implies the following p-order local minimum condition: along* $(x_*(t), u_*(t), t)$ *we have for every* $u \in U$ *and a.e.* $t \in [0,T]$

$$\left\langle \nu_0 L_u(x_*, u_*) + \psi(t) f_u + \sum_{i=1}^{p-1} \rho_i(t) \left(\nabla^i f(H_i)\right)_u , u - u_*(t) \right\rangle \ge 0. \tag{6.7}$$

7. Proof. Let $z_* = (x_*, u_*)$. Since H_{p-1} is a p-regular critical direction, Theorem 7.1 of [22] applies in its nondegenerate form. The control constraint has the form $\mathcal{Z} = \overline{W}^n_{11}(0,T) \times \mathcal{U}$, where $\mathcal{U} = \{u \in L^m_\infty(0,T) : u(t) \in U \text{ a.e. on } (0,T)\}$ is convex. Since the constraint is active only in the control variable u, $FS^{(p)}(\mathcal{Z}; z_*, H_{p-1})$ is trivially the full space $\overline{W}^n_{11}(0,T)$ in the first coordinate. Thus the p-order feasible set to $\mathcal{Z}$ at z_* in direction of H_{p-1} takes the form

$$FS^{(p)}(\mathcal{Z}; z_*, H_{p-1}) = \overline{W}^n_{11}(0,T) \times FS^{(p)}(\mathcal{U}; u_*, V_{p-1}).$$

Hence by (iv) this set is nonempty. Note also that, if $(\overline\lambda, \mu)$ defines a supporting hyperplane to $FS^{(p)}(\mathcal{Z}; z_*, H_{p-1})$, i.e.,

$$\left\langle \overline\lambda, (\eta, \xi) \right\rangle + \mu \ge 0 \quad \text{for} \quad (\eta, \xi) \in FS^{(p)}(\mathcal{Z}; z_*, H_{p-1}),$$

then $\overline\lambda$ is of the form $\overline\lambda = (0, \lambda)$ and (λ, μ) defines a supporting hyperplane to $FS^{(p)}(\mathcal{U}; u_*, V_{p-1})$.

It therefore follows from [22, Theorem 7.1] that there exists a Lagrange multiplier $\nu_0 \geq 0$, functionals $(y_i^*, w_i^*) \in Y_{i-1}^{\perp_i},\ i = 1, \dots, p$, and a supporting hyperplane (λ, μ) to $FS^{(p)}(\mathcal{U}; u_*, V_{p-1})$, all depending on the sequence H_{p-1}, such that the multipliers ν_0 and λ do not both vanish, and

$$(0, \lambda) \equiv \nu_0 I'(z_*) + \sum_{i=1}^{p} G_i^*[F_1](z_*; H_{i-1}) y_i^* + \sum_{i=1}^{p} G_i^*[F_2](z_*; H_{i-1}) w_i^*, \tag{7.1}$$

$$\begin{aligned} \mu \leq \nu_0 R_p[I](z_*; H_{p-1}) &+ \sum_{i=1}^{p} \langle y_i^*, R_{p-1+i}[F_1](z_*; H_{p-1}) \rangle \\ &+ \sum_{i=1}^{p} \langle w_i^*, R_{p-1+i}[F_2](z_*; H_{p-1}) \rangle. \end{aligned} \tag{7.2}$$

In these formulas F_1 and F_2 denote the components of the operator F, $[F_i]$, respectively, $[I]$ indicates that the operators G or R are taken on F_i, respectively, I, and $*$ denotes the dual map. Also $Y_{i-1}^{\perp_i}$ denotes the annihilator of Y_{i-1} in Y_i^*, i.e., $Y_{i-1}^{\perp_i} = \{z^* \in Y_i^* :\ \langle z^*, v \rangle = 0\ \forall\ v \in Y_{i-1}\}$ and we formally set $Y_0 = \{0\}$, so that $Y_0^{\perp_1} \cong Y_1^*$.

We now analyze these equations. Conditions (b) and (c) of the theorem follow from the Euler–Lagrange equation (7.1). For simplicity of notation we omit the arguments $(z_*; H_i)$. We then have for arbitrary $(\eta, \xi) \in \overline{W}_{11}^n(0,T) \times L_\infty^m(0,T)$ that

$$\lambda(\xi) \equiv \nu_0 I'(\eta, \xi) + \sum_{i=1}^{p} \langle y_i^*, G_i[F_1](\eta, \xi) \rangle + \sum_{i=1}^{p} \langle w_i^*, G_i[F_2](\eta, \xi) \rangle\,. \tag{7.3}$$

The (y_i^*, w_i^*) are continuous linear functionals on the spaces Y_i which extend to continuous linear functionals $(\overline{y}_i^*, \overline{w}_i^*)$ in the full space $\overline{W}_{11}^n(0,T)^* \times (\mathbb{R}^k)^*$. Using the representations of continuous linear functionals in $\overline{W}_{11}^n(0,T)^*$ [14, p. 21], it follows that there exist functions $\psi,\ \rho_i \in L_\infty^m(0,T),\ i = 1, \dots, p-1$, such that for all $(\eta, \xi) \in \overline{W}_{11}^n(0,T) \times L_\infty^m(0,T)$,

$$\langle y_1^*, G_1[F_1](\eta, \xi) \rangle = \langle \overline{y}_1^*, G_1[F_1](\eta, \xi) \rangle = -\int_0^T \psi(t) \frac{d}{dt} \left(G_1[F_1](\eta, \xi) \right) dt, \tag{7.4}$$

and for $i = 2, \dots, p$,

$$\langle y_i^*, G_i[F_1](\eta, \xi) \rangle = \langle \overline{y}_i^*, G_i[F_1](\eta, \xi) \rangle = -\int_0^T \rho_{i-1}(t) \frac{d}{dt} \left(G_i[F_1](\eta, \xi) \right) dt. \tag{7.5}$$

Similarly, we represent the second factor by row-vectors a_{i-1}(note the shift in index),

$$\langle w_i^*, G_i[F_2](\eta, \xi) \rangle = \langle \overline{w}_i^*, G_i[F_2](\eta, \xi) \rangle = a_{i-1} G_i[F_2](\eta, \xi). \tag{7.6}$$

Since we consider extensions of the functionals (y_i^*, w_i^*) from Y_i to the full space Y, the multipliers ψ, ρ_i, and a_i are not unique but depend on these extensions except for (ρ_{p-1}, a_{p-1}) which corresponds to the functional (y_p^*, w_p^*) which already is defined on the full space. We shall show below that the particular extension away from Y_i, although it may give rise to different multipliers in the statement of the theorem, still has no influence in the conditions of the theorem. For the moment, we just pick one

representation. Thus we have for all $(\eta,\xi)\in \overline{W}^n_{11}(0,T)\times L^m_\infty(0,T)$ that

(7.7)
$$\begin{aligned}\lambda(\xi) &= \nu_0\int_0^T (L_x\eta + L_u\xi)ds + \nu_0\ell_x(x_*(T))\eta(T) - \int_0^T \psi(t)\frac{d}{dt}\left(G_1[F_1](\eta,\xi)\right)dt\\ &\quad - \sum_{i=1}^{p-1}\int_0^T \rho_i(t)\frac{d}{dt}\left(G_{i+1}[F_1](\eta,\xi)\right)dt + a_0G_1[F_2](\eta,\xi) + \sum_{i=1}^{p-1}\langle a_i, G_{i+1}[F_2](\eta,\xi)\rangle.\end{aligned}$$

We need to evaluate the time derivatives $\frac{d}{dt}G_i[F](z_*;H_{p-1})$ of the G-operators along the reference trajectory. Since the operator F splits as $F=(F_1,F_2)$, also the operators G_i defined in (4.4) are applied componentwise. Note that we have $F_1(x,u)=x-I(f)(x,u)$, where $I(f)$ is the integral operator $I(f)(x,u)=\int_0^{(\cdot)} f(x(s),u(s),s)ds$. The linear term x will enter only the first operator G_1 but not the operators $G_i,\ i\geq 2$, which depend only on higher derivatives. And the time-derivative will simply cancel the integration. We can use the definition of the ith order directional derivative at a point $\varsigma\in Z$ along the sequence H_i for an arbitrary operator $\Phi: Z\to Y$ between Banach spaces to write the general formula concisely: let $\nabla^i\Phi: Z\to Y,\ i\in\mathbb{N}$,

$$\begin{aligned}&\nabla^i\Phi(\varsigma;H_i): Z\to Y,\\ &\varsigma\mapsto\ \nabla^i\Phi(\varsigma;H_i) = \sum_{r=1}^{i}\frac{1}{r!}\left(\sum_{j_1+\cdots+j_r=i}\Phi^{(r)}(\varsigma)(h_{j_1},\dots,h_{j_r})\right).\end{aligned}\tag{7.8}$$

It then follows for $i=2,\dots,p$ that

$$\begin{aligned}&\frac{d}{dt}\left(G_i[F_1](z_*;H_{i-1})\cdot(\eta,\xi)\right)\\ &\quad= \frac{d}{dt}\left[\sum_{r=1}^{i-1}\frac{1}{r!}\sum_{j_1+\cdots+j_r=i-1}\left(-\int_0^t\frac{\partial f^{(r)}}{\partial x}(z_*)(\eta_{j_1},\xi_{j_1};\dots;\eta_{j_r},\xi_{j_r})\eta\right.\right.\\ &\qquad\left.\left.+\frac{\partial f^{(r)}}{\partial u}(z_*)(\eta_{j_1},\xi_{j_1};\dots;\eta_{j_r},\xi_{j_r})\xi ds\right)\right]\\ &\quad= -\left(\nabla^{i-1}f\right)_x(z_*;H_{i-1})\eta - \left(\nabla^{i-1}f\right)_u(z_*;H_{i-1})\xi.\end{aligned}\tag{7.9}$$

With this notation for $i=2,\dots,p$ we also have that

$$G_i[F_2](z_*;H_{i-1})(\eta,\xi) = \left(\nabla^{i-1}q(x_*(T);H_{i-1})\right)_x\cdot\eta(T).\tag{7.10}$$

Using these formulas we thus obtain for all $(\eta,\xi)\in\overline{W}^n_{11}(0,T)\times L^m_\infty(0,T)$ that

$$\begin{aligned}\lambda(\xi) &= \int_0^T\left(\left(\nu_0L_x+\psi f_x+\sum_{i=1}^{p-1}\rho_i\left(\nabla^i f(H_i)\right)_x\right)\eta-\psi\dot\eta\right)dt\\ &\quad+\int_0^T\left\langle\nu_0L_u+\psi f_u+\sum_{i=1}^{p-1}\rho_i\left(\nabla^i f(H_i)\right)_u,\xi\right\rangle dt\\ &\quad+\nu_0\ell_x(x_*(T))\eta(T)+a_0q_x(x_*(T))\eta(T)+\sum_{i=1}^{p-1}\langle a_i\left(\nabla^i q(H_i)\right)_x,\eta(T)\rangle.\end{aligned}\tag{7.11}$$

Conditions (b)–(c) of the theorem follow from here using the following version of the classical DuBois–Raymond lemma. A proof is given, for instance, in [18].

LEMMA 7.1 (see [18]). *Suppose $\alpha, \beta \in L_1^n(0,T), c \in \mathbb{R}^n$, and for all $h \in \overline{W}_{11}^n(0,T)$ we have*

$$\int_0^T \alpha(t)h(t) + \beta(t)\dot{h}(t)dt = \langle c, h(T)\rangle. \tag{7.12}$$

Then

$$\beta(t) = c - \int_t^T \alpha(s)ds. \tag{7.13}$$

In particular, if necessary after changing β on a set of measure zero, $\beta \in \overline{W}_{11}^n(0,T)$.

Equations (6.2) and (6.3) directly follow if we take $\xi \equiv 0$ in (7.11). The extra conditions on the multipliers ρ_i and a_i follow from the orthogonality relations $(y_{i+1}^*, w_{i+1}^*) \in Y_i^{\perp_{i+1}}$, $i = 1, \dots, p-1$. Since the multiplier (y_{i+1}^*, w_{i+1}^*) is orthogonal to $\operatorname{Im} G_j$ for $j = 1, \dots, i$, we have for all $(\eta, \xi) \in \overline{W}_{11}^n(0,T) \times L_\infty^m(0,T)$ and $j = 1, \dots, i$ that

$$0 = \left\langle y_{i+1}^*, G_j[F_1](\eta,\xi)\right\rangle + \langle w_{i+1}^*, G_j[F_2](\eta,\xi)\rangle. \tag{7.14}$$

For $j = 1$, this yields

$$0 = \int_0^T \rho_i \left(f_x\eta + f_u\xi - \dot{\eta}\right) dt + a_i q_x(x_*(T))\eta(T), \tag{7.15}$$

and for $j = 2, \dots, i$,

$$0 = \int_0^T \rho_i \left(\left(\nabla^{j-1} f(H_{j-1})\right)_x \eta + \left(\nabla^{j-1} f(H_{j-1})\right)_u \xi\right) dt + a_i \left(\nabla^{j-1} q\right)_x (x_*(T))\eta(T). \tag{7.16}$$

Conditions (6.4) and (6.5) are consequences of these relations using Lemma 7.1. For instance, setting $\xi = 0$ in the second equation, we obtain that

$$\int_t^T \rho_i \left(\nabla^{j-1} f(H_{j-1})\right)_x ds = -a_i \left(\nabla^{j-1} q(H_{j-1})\right)_x (x_*(T)) = \text{const}, \tag{7.17}$$

and from this (6.5) follows.

If we set $\eta = 0$ in (7.11), we obtain a representation for the functional λ: for all $\xi \in L_\infty^m(0,T)$,

$$\lambda(\xi) = \int_0^T \left\langle \nu_0 L_u + \psi f_u + \sum_{i=1}^{p-1} \rho_i \left(\nabla^i f(H_i)\right)_u, \xi \right\rangle dt. \tag{7.18}$$

Since (λ, μ) is a supporting hyperplane to $FS^{(p)}(\mathcal{U}; u_*, V_{p-1})$, it follows that $\lambda(\xi) + \mu \geq 0$ for all $\xi \in FS^{(p)}(\mathcal{U}; u_*, V_{p-1})$. But (7.2) states that

$$\mu \leq \nu_0 R_p[I] + \sum_{i=1}^p \langle y_i^*, R_{p-1+i}[F_1]\rangle + \sum_{i=1}^p \langle w_i^*, R_{p-1+i}[F_2]\rangle,$$

and by definition and using the representations of these functionals, we have

$$R_p[I] = \int_0^T R_p[L](H_{p-1})dt + R_p[\ell](H_{p-1}), \tag{7.19}$$

$$\begin{aligned}\langle y_1^*, R_p[F_1]\rangle &= -\int_0^T \left\langle \psi(t), \frac{d}{dt}\left(R_p[F_1](H_{p-1})\right)\right\rangle dt \\ &= \int_0^T \psi(t) R_p[f](H_{p-1})dt, \end{aligned} \tag{7.20}$$

$$\begin{aligned}\langle y_i^*, R_{p-1+i}[F_1]\rangle &= -\int_0^T \left\langle \rho_{i-1}(t), \frac{d}{dt}\left(R_{p-1+i}[F_1](H_{p-1})\right)\right\rangle dt \\ &= \int_0^T \rho_{i-1}(t) R_{p-1+i}[f](H_{p-1})dt, \end{aligned} \tag{7.21}$$

$$\langle w_i^*, R_{p-1+i}[F_2]\rangle = a_{i-1} R_{p-1+i}[q](H_{p-1}). \tag{7.22}$$

Using the representation (7.18) for λ, we therefore obtain condition (6.6). This proves the result. □

The corollary is a direct consequence of the fact that (λ, μ) is a supporting hyperplane to $FS^{(p)}(\mathcal{U}; u_*, V_{p-1})$ since this implies that λ is a supporting functional to $\mathcal{U}$ at u_* [22, Corollary 6.3]. Hence the minimum condition (6.7) follows from Example 10.5 in [12]. We include it since this is the common form of the local minimum principle, but the separation condition is stronger (see the section with examples below).

Remark: Nontriviality of the extension. If H_{p-1} satisfies conditions (i)–(iv) in the definition of critical directions, but the operator F is not p-regular along H_{p-1}, then the conditions of Theorem 6.1 can be satisfied with a trivial choice of multipliers for which both $\nu_0 = 0$ and $\lambda \equiv 0$. Now the nontriviality statement is only on all multipliers which is significantly weaker than the strong nontriviality statement (a) of Theorem 6.1. In fact, in this case the conditions can be satisfied by choosing all the new multipliers zero and for (ν_0, ψ) the multipliers of the local maximum principle. This, of course, defeats the purpose of our extension. The p-regularity eliminates this choice of multipliers and thus gives us a nontrivial extension.

Remark: On the uniqueness of multipliers. There are two sources for nonuniqueness of the multipliers, one essential, the other unimportant. Different multipliers can arise from the existence of linearly independent supporting functionals $(\overline{\lambda} = (0, \lambda), \mu)$ to $FS^{(p)}(\mathcal{Z}; z_*, H_{p-1})$, and this is just one aspect of the particular problem to be considered. However, given λ and ν_0, since the operator $\mathcal{G}_p$ is onto, the functionals $(y_i^*, w_i^*) \in Y_{i-1}^{\perp_i}$, $i = 1, \dots, p$, in the generalized Euler–Lagrange equation (7.1) are uniquely determined. The multipliers ψ, ρ_i, and a_i in Theorem 6.1 are determined by representations obtained for these functionals after extending the (y_i^*, w_i^*) to continuous linear functionals on the full space. Naturally they depend on these extensions, and different independent solutions to conditions (b) and (c) can exist which all represent the same functionals (y_i^*, w_i^*). On the other hand, these extensions should be irrelevant and indeed the nonuniqueness caused in this way has no effect on *the separation condition* (d) *which is invariant* for all these multipliers in the sense that the value of the right-hand side in (6.6) does not depend on the choice of the specific representative from this class of multipliers.

If $\psi + \widehat{\psi}$, $\rho_i + \widehat{\rho}_i$ for $i = 1, \dots, p-2$, and $a_i + \widehat{a}_i$ for $i = 0, \dots, p-2$ are also multipliers which represent these functionals, then $(\widehat{\psi}, \widehat{a}_0)$ defines a functional $(\widehat{y}_1^*, \widehat{w}_1^*)$ in the annihilator of Y_1, and the $(\widehat{\rho}_i, \widehat{a}_i)$ define functionals $(\widehat{y}_{i+1}^*, \widehat{w}_{i+1}^*)$ in

the annihilator of Y_{i+1}, $i = 1, \dots, p-2$. (The last multipliers (ρ_{p-1}, a_{p-1}) are unique since no extensions need to be taken.) It therefore follows for all $(\eta, \xi) \in \overline{W}^n_{11}(0,T) \times L^m_\infty(0,T)$ that

$$
\begin{aligned}
0 &= \langle \widehat{y}_1^*, G_1[F_1](\eta,\xi)\rangle + \langle \widehat{w}_1^*, G_1[F_2](\eta,\xi)\rangle \\
&= -\int_0^T \widehat{\psi}(t)\frac{d}{dt}\left(G_1[F_1](\eta,\xi)\right) dt + \widehat{a_0} G_1[F_2](\eta,\xi)
\end{aligned} \tag{7.23}
$$

and

$$
\begin{aligned}
0 &= \langle \widehat{y}_{i+1}^*, G_{i+1}[F_1](\eta,\xi)\rangle + \langle \widehat{w}_{i+1}^*, G_{i+1}[F_2](\eta,\xi)\rangle \\
&= -\int_0^T \widehat{\rho}_i(t)\frac{d}{dt}\left(G_{i+1}[F_1](\eta,\xi)\right) dt + \widehat{a}_i G_{i+1}[F_2](\eta,\xi).
\end{aligned} \tag{7.24}
$$

In particular, the multipliers $(\widehat{\psi}, \widehat{a_0})$ satisfy analogous orthogonality conditions as the multiplier (ρ_1, a_1) and for $i = 2, \dots, p-2$, $(\widehat{\rho}_i, \widehat{a}_i)$ satisfies analogous orthogonality conditions as (ρ_{i+1}, a_{i+1}). Thus we have

$$
\dot{\widehat{\psi}}(t) = -\widehat{\psi}(t) f_x, \quad \widehat{\psi}(t) f_u \equiv 0, \quad \widehat{\psi}(T) = \widehat{a_0} q_x(x_*(T)), \tag{7.25}
$$

and for $j = 1, \dots, i$, the following conditions are satisfied for a.e. $t \in [0,T]$:

$$
\widehat{\rho}_i(t)\left(\nabla^j f(H_j)\right)_x = 0, \quad \widehat{\rho}_i(t)\left(\nabla^j f(H_j)\right)_u = 0, \quad \widehat{a}_i\left(\nabla^j q((x_*(1); H_j)\right)_x = 0. \tag{7.26}
$$

Note that, different from the orthogonality conditions (c) in Theorem 6.1, here the last index is $j = i$. Therefore we obtain, for example,

$$
\nu_0 L_u + \left(\psi + \widehat{\psi}\right) f_u + \sum_{i=1}^{p-1} (\rho_i + \widehat{\rho}_i)\left(\nabla^i f(H_i)\right)_u = \nu_0 L_u + \psi f_u + \sum_{i=1}^{p-1} \rho_i \left(\nabla^i f(H_i)\right)_u \tag{7.27}
$$

since $\widehat{\psi} f_u \equiv 0$ and also still $\widehat{\rho}_i \left(\nabla^i f(H_i)\right)_u \equiv 0$. As it should be, the extended adjoint equation and its terminal condition are not affected with such a choice. Furthermore, by the compatibility condition (ii) we have $R_{p-1+i}[F](z_*; H_{p-1}) \in Y_i$, and thus it also follows that

$$
\begin{aligned}
0 &= \langle \widehat{y}_1^*, R_p[F_1]\rangle + \langle \widehat{w}_1^*, R_p[F_2]\rangle \\
&= \int_0^T \widehat{\psi}(t) R_p[f](H_{p-1}) dt + \widehat{a_0} R_p[q](H_{p-1}),
\end{aligned} \tag{7.28}
$$

and for $i = 2, \dots, p-1$,

$$
\begin{aligned}
0 &= \langle \widehat{y}_i^*, R_{p-1+i}[F_1]\rangle + \langle \widehat{w}_i^*, R_{p-1+i}[F_2]\rangle \\
&= \int_0^T \widehat{\rho}_i(t) R_{p+i}[f](H_{p-1}) dt + \widehat{a}_i R_{p+i}[q](H_{p-1}).
\end{aligned} \tag{7.29}
$$

Hence we have

$$
\begin{aligned}
0 = {} & \widehat{a}_0 R_p[q](H_{p-1}) + \sum_{i=1}^{p-1} \widehat{a}_i R_{p+i}[q](H_{p-1}) \\
& + \int_0^T \left\langle \widehat{\psi} f_u + \sum_{i=1}^{p-1} \widehat{\rho}_i(t) \left(\nabla^i f(H_i)\right)_u, \xi \right\rangle dt \\
& + \int_0^T \left(\widehat{\psi}(t) R_p[f](H_{p-1}) + \sum_{i=1}^{p-1} \widehat{\rho}_i(t) R_{p+i}[f](H_{p-1}) \right) dt.
\end{aligned} \tag{7.30}
$$

Thus the separation condition (d) is independent of the extensions of the linear functionals $(y_i^*, w_i^*) \in Y_{i-1}^{\perp_i},\ i = 1, \dots, p-1$.

In the case of a Lagrangian minimization problem which has no control constraints, the functional λ vanishes identically. In this case we can normalize $\nu_0 = 1$ and the multipliers $(y_i^*, w_i^*) \in Y_{i-1}^{\perp_i},\ i = 1, \dots, p$, in the generalized Euler–Lagrange equation (7.1) are unique since $\mathcal{G}_p^*$ is one-to-one. In this case we therefore obtain the following corollary.

COROLLARY 7.2 (p-order local maximum principle for Lagrangian problems). *Consider the problem (OC) without control constraints ($U = \mathbb{R}^m$), and suppose the admissible process (x_*, u_*) is optimal. Then for every p-regular critical direction H_{p-1}, there exist vectors $a_i = a(H_{p-1}) \in (\mathbb{R}^k)^*$, $i = 0, 1, \dots, p-1$, and absolutely continuous functions $\psi(\cdot) = \psi(H_{p-1})(\cdot)$ and $\rho_i(\cdot) = \rho_i(H_{p-1})(\cdot)$, $i = 1, \dots, p-1$, from $[0, T]$ into $(\mathbb{R}^n)^*$, which satisfy the conditions* (b)–(d) *of Theorem* 6.1 *along the optimal trajectory $(x_*(t), u_*(t), t)$ for $\nu_0 = 1$. In particular, we thus have*

$$
L_u(x_*, u_*) + \psi(t) f_u + \sum_{i=1}^{p-1} \rho_i(t) \left(\nabla^i f(H_i)\right)_u \equiv 0. \tag{7.31}
$$

Furthermore, for any multipliers which satisfy conditions (b) *and* (c)*, the value of the right-hand side in* (6.6) *is the same.*

8. Approximations of order r in the p-regular case, $r > p$. Theorem 6.1 is based on p-order approximations. If these remain inconclusive, higher-order approximations can now easily be set up. Suppose H_{p-1} is a p-regular critical direction, and consider higher-order approximations H_{r-1}, $r > p$, whose first $p-1$ components are given by H_{p-1}. Several simplifications occur. For instance, since $FS^{(p)}(\mathcal{U}; u_*, V_{p-1})$ is nonempty, arbitrary higher-order approximations for the control constraint can be made, i.e., $FS^{(r)}(\mathcal{U}; u_*, V_{r-1}) = L_\infty^m(0, T)$. More importantly, F is trivially r-regular at (x_*, u_*) in direction of H_{r-1} since already the first p components of the map $\mathcal{G}_r$ are onto. In this case, all the quotient spaces Y_{i+1}/Y_i are $\{0\}$ for $i = p, \dots, r-1$. In particular, all the multipliers associated with these terms are zero, and therefore no additional multipliers arise in higher-order approximations. Thus only the operator $\mathcal{G}_p$ needs to be considered.

We show now that the required higher-order expansions for the tangent directions can easily be made by solving linear equations for the operator $\mathcal{G}_p$. We first analyze the higher-order approximations to the constraint $\mathcal{F} = \{z \in Z : F(z) = 0\}$ in a Banach

space Z. For $i = 0, \dots, p-1$, the term at ε^{p+i} in the expansion of $F(z_* + \sum_{j=1}^{p+i} \varepsilon^j h_j)$ can be expressed as

$$\nabla^{p+i} F(z_*; H_\ell) = \sum_{k=1}^{i+1} G_k[F](z_*; H_{k-1}) h_{p+i+1-k} + R_{p+i}[F](z_*; H_{p-1}), \tag{8.1}$$

and for $\ell \geq 2p$, we get

$$\nabla^{\ell} F(z_*; H_\ell) = \sum_{k=1}^{p} G_k[F](z_*; H_{k-1}) h_{\ell+1-k} + R_{\ell}[F](z_*; H_{\ell-p}). \tag{8.2}$$

($\nabla^\ell F$ collects all terms for which the indices of the directions h sum to ℓ. The operators G_k are used to write contributions of the h vectors with highest index separately. But for our construction the vectors in H_{p-1} are given and only terms with higher indices need to be chosen. Therefore, in the terms of orders p through $2p-1$, we consider only the vectors $h_p, \dots, h_\ell$ as free and restrict the range of the operators G_k, while for terms of orders $\geq 2p$, we always step down the full length p.) There are simple relations between these operators which follow directly from the definition of the R operators. For example, for $\ell = p, \dots, 2p-1$,

$$G_{\ell+1-p}[F](z_*; H_{\ell-p}) h_p + R_\ell[F](z_*; H_{p-1}) = R_\ell[F](z_*; H_p). \tag{8.3}$$

Here $R_\ell[F](z_*; H_p)$ consists of all terms which are homogeneous of degree ℓ, but involve only vectors from H_p. On the left-hand side these terms are split into those which contain h_p and those which don't. Since the total order is less than $2p$, h_p can enter linearly only with terms which must be homogeneous of degree $\ell - p$. These are given by $G_{\ell+1-p}[F](z_*; H_{\ell-p})$.

The construction of higher-order approximating sequences is inductive and requires only solution of *linear* equations in every step. It is shown in [22, Corollary 3.3] that the p-order tangent directions h_p to $\mathcal{F}$ at z_* along a sequence H_{p-1} (which satisfies conditions (i) and (ii) and along which F is p-regular) are given by the solutions to the linear equation

$$\mathcal{G}_p[F](z_*; H_{p-1}) h_p + \mathcal{R}_{p-1}[F](z_*, H_{p-1}) = 0, \tag{8.4}$$

where $\mathcal{R}_{p-1}[F](z_*, H_{p-1}) \in Y_1 \times Y_2/Y_1 \times \cdots \times Y/Y_{p-1}$ is the point with components

$$\left(R_p[F](z_*; H_{p-1}), \pi_1 R_{p+1}[F](z_*; H_{p-1}), \dots, \pi_{p-1} R_{2p-1}[F](z_*; H_{p-1})\right). \tag{8.5}$$

This vector is well defined by the compatibility condition. In the next step we need to choose h_{p+1} so that amongst other conditions we have

$$\begin{aligned} \nabla^{p+1} F(z_*; H_p) &= G_1(z_*) h_{p+1} + \underbrace{G_2(z_*; H_1) h_p + R_{p+1}[F](z_*; H_{p-1})}_{} \\ &= G_1(z_*) h_{p+1} + \qquad R_{p+1}[F](z_*; H_p) = 0. \end{aligned}$$

By the choice of h_p as a p-order tangent direction the required necessary condition $R_{p+1}[F](z_*; H_p) \in Y_1$ is satisfied since h_p already solves the equation $\nabla^{p+1} F(z_*; H_p) = 0$ in the quotient space Y_2/Y_1, i.e., except possibly for a remaining term in Y_1. (See [21] for more detailed explanations on calculating p-order tangent directions.) Similarly, the necessary conditions $R_{p+i}[F](x_*; H_p) \in Y_i$ are satisfied for all the other components, and we can inductively solve for h_{p+1} while preserving the required necessary

conditions for setting up higher-order approximations. Note that these conditions are essential since only in this way can we actually assert that H_{p+1} is a $(p+1)$-order tangent direction. In general we have the following.

PROPOSITION 8.1. *Let H_{p-1} be a $(p-1)$-order tangent direction to $\mathcal{F}$ at z_*, and suppose F is p-regular at z_* in direction of H_{p-1}. Let H_{r-1} be an extension of H_{p-1} to an $(r-1)$-order tangent direction to $\mathcal{F}$ at z_* (i.e., $\nabla^i F(z_*)(H_i) = 0$ for $i = p, \dots, r-1$, and the compatibility conditions $R_{r-1+i}[F](z_*; H_{r-1}) \in Y_i$ for $i = 1, \dots, p-1$ are satisfied), and let $\mathcal{R}_{r-1}[F](z_*; H_{r-1})$ be the point with components*

$$(R_r[F](z_*; H_{r-1}), \pi_1 R_{r+1}[F](z_*; H_{r-1}), \dots, \pi_{p-1} R_{r+p-1}[F](z_*; H_{r-1})). \tag{8.6}$$

Then the set of all r-order tangent directions to $\mathcal{F}$ at z_* along H_{r-1}, $TS^{(r)}(\mathcal{F}; z_*, H_{r-1})$, is a nonempty affine variety given by the solutions to the equation

$$\mathcal{G}_p[F](z_*; H_{p-1}) h_r + \mathcal{R}_{r-1}[F](z_*; H_{r-1}) = 0. \tag{8.7}$$

Proof. In the state-space Z the vectors h_r which extend H_{r-1} to r-order tangent directions are the solutions to the following equations:

$$\begin{aligned}
\nabla^r F(z_*; H_{r-1}) &= G_1(z_*) h_r + R_r[F](z_*; H_{r-1}) = 0, \\
\nabla^{r+1} F(z_*; H_{r-1}) &= G_1(z_*) h_{r+1} + G_2(z_*; h_1) h_r + R_{r+1}[F](z_*; H_{r-1}) \in Y_1 \\
&\dots \\
\nabla^{r+i} F(z_*; H_{r-1}) &= \sum_{k=1}^{i+1} G_k(z_*; H_{k-1}) h_{r+i+1-k} + R_{r+i}[F](z_*; H_{r-1}) \in Y_i
\end{aligned}$$

for $i = 0, \dots, p-1$, and the vectors h_j for $j > r$ are left undetermined in this step. By assumption (guaranteed by the inductive steps of the construction) we have

$$R_{r+i}[F](z_*; H_{r-1}) \in Y_{i+1},$$

and therefore, since the operator $\mathcal{G}_p$ is onto, we can choose h_r so that

$$G_{i+1}(z_*; H_i) h_r + R_{r+i}[F](z_*; H_{r-1}) \in Y_i. \tag{8.8}$$

These vectors are precisely the solutions to (8.7). It follows similar to the proof of the main approximation lemma in [21] that any such vector defines an r-order tangent direction. The only change is that the order of approximation needs to be raised to r. ∎

Note that we need only to solve *linear* equations, but in every step a system of p equations needs to be solved since the operator F is only p-regular. As in [22, section 2] we define the r-order tangent cone to $\mathcal{F}$ at z_* in direction of H_{r-1}, $TC^{(r)}(\mathcal{F}; z_*, H_{r-1})$, as the cone in $Z \times \mathbb{R}$ generated by the vectors $(v, 1)$ with $v \in TS^{(r)}(\mathcal{F}; z_*, H_{r-1})$. The dual or polar cone consists of all continuous linear functionals which are nonnegative on $TC^{(r)}(\mathcal{F}; z_*, H_{r-1})$. Since the operator $\mathcal{G}_p$ is used in each of these higher-order approximations, the structure of the dual cone directly follows from [22, Proposition 3.6].

PROPOSITION 8.2. *The dual or polar r-order tangent cone consists of all linear functionals* $(\lambda, \mu) \in X^* \times \mathbb{R}$ *which can be represented in the following form: there exist functionals* $y_i^* \in Y_{i-1}^{\perp_i}$, $i = 1, \dots, p$, *and a number* $s \geq 0$ *such that*

$$\lambda = \sum_{i=1}^{p} G_i^*[F](z_*; H_{i-1}) y_i^* \ , \tag{8.9}$$

$$\mu = \sum_{i=1}^{p} \langle y_i^*, R_{r-1+i}[F](z_*; H_{r-1}) \rangle + s. \tag{8.10}$$

Thus, given a p-regular tangent direction, it is possible to set up higher-order approximations of arbitrary order. Now consider the problem to minimize a functional $I : Z \to \mathbb{R}$ over the set $\mathcal{F}$. If there exists an integer $j \leq r$ such that $\nabla^j I(z_*; H_j) < 0$ while $\nabla^i I(z_*; H_i) = 0$ for $i = 1, \dots, j-1$, then z_* is not optimal since there exist j-order tangent directions to $\mathcal{F}$ for which I will produce better values than at z_*. Thus it is a necessary condition for optimality of z_* that the first nonzero derivative of $\nabla^i I(z_*; H_i)$ is positive. Once this happens, it is clear that no higher-order approximations which extend H_i are useful in obtaining necessary conditions for optimality. Therefore the r-order critical directions for the objective are those which satisfy

(iii-r) $\nabla^i I(z_*)(H_i) = 0$ for $i = p, \dots, r-1$.

THEOREM 8.3. *Suppose the admissible process* (x_*, u_*) *is optimal for the optimal control problem* (OC), *and let* H_{p-1} *be a p-regular critical direction. Let* H_{r-1} *be an extension of* H_{p-1} *to an* $(r-1)$*-order tangent direction to the equality constraint which is r-order critical for the objective. Then there exist vectors* $a_i = a(H_{p-1}) \in (\mathbb{R}^k)^*$, $i = 0, 1, \dots, p-1$, *and absolutely continuous functions* $\psi(\cdot) = \psi(H_{p-1})(\cdot)$ *and* $\rho_i(\cdot) = \rho_i(H_{p-1})(\cdot)$, $i = 1, \dots, p-1$, *from* $[0, T]$ *into* $(\mathbb{R}^n)^*$, *which satisfy conditions* (b) *and* (c) *of Theorem* 6.1 *with* $\nu_0 = 1$ *and are such that*

$$\begin{aligned} 0 \leq\ & R_r[\ell](H_{r-1}) + a_0 R_r[q](H_{r-1}) + \sum_{i=0}^{p-1} a_i R_{r+i}[q](H_{r-1}) \\ & + \int_0^T \left(R_r[L](H_{r-1}) + \psi(t) R_r[f](H_{r-1}) + \sum_{i=1}^{p-1} \rho_i(t) R_{r+i}[f](H_{r-1}) \right) dt. \end{aligned} \tag{8.11}$$

The value of the right-hand side in (8.11) *is the same independent of the choice of multipliers* a_i, ψ, *and* ρ_i *which satisfy conditions* (b) *and* (c).

Proof. Let $z_* = (x_*, u_*)$. Under these conditions the r-order tangent cone and the r-order cone of decrease are nonempty. Since the feasible cone is the full space, it is therefore a necessary condition for optimality that these cones do not intersect. The r-order cone of decrease of the functional I at z_* in direction of H_{r-1}, $DC^{(r)}(I; z_*, H_{r-1})$, is given by [22, Proposition 4.1] as

$$DC^{(r)}(I; z_*, H_{r-1}) = \{(w, \gamma) \in X \times \mathbb{R} : \gamma > 0, I'(z_*)w + \gamma R_r[I](z_*; H_{r-1}) < 0\} \tag{8.12}$$

and thus is nonempty, open, and convex. Its dual cone is given by [22, Proposition 4.2] as

$$\Big(DC^{(r)}(I;z_*,H_{r-1})\Big)^* = \Big\{(\lambda,\mu)\in X^*\times\mathbb{R}:\ \exists\ \alpha_1\le 0,\ \alpha_2\ge 0 \text{ such that } \begin{pmatrix}\lambda\\ \mu\end{pmatrix} = \begin{pmatrix} I'(z_*) & 0\\ R_r[I](z_*;H_{r-1}) & 1\end{pmatrix}\begin{pmatrix}\alpha_1\\ \alpha_2\end{pmatrix}\Big\}. \tag{8.13}$$

Hence it follows from the Dubovitskii–Milyutin lemma [12] that there exist *nontrivial* continuous linear functionals in the dual cones, $\big(TC^{(r)}(\mathcal{F};z_*,H_{r-1})\big)^*$ and $\big(DC^{(r)}(I;z_*,H_{r-1})\big)^*$, whose sum vanishes identically. Thus there exist functionals $(y_i^*,w_i^*)\in Y_{i-1}^{\perp_i},\ i=1,\dots,p$, such that

(8.14)
$$0\equiv I'(z_*)+\sum_{i=1}^{p} G_i^*[F_1](z_*;H_{i-1})y_i^*+\sum_{i=1}^{p} G_i^*[F_2](z_*;H_{i-1})w_i^*,$$

(8.15)
$$0\le R_r[I](z_*;H_{r-1})+\sum_{i=1}^{p}\langle y_i^*,R_{r-1+i}[F_1](z_*;H_{r-1})\rangle+\sum_{i=1}^{p}\langle w_i^*,R_{r-1+i}[F_2](z_*;H_{r-1})\rangle.$$

The first equation is the Euler–Lagrange equation (7.1) for the case $\lambda\equiv 0$ and $\nu_0=1$. Thus conditions (b) and (c) of Theorem 6.1 follow. The second-order condition (8.15) gives (8.11). In this derivation only the index changes from p to r. Furthermore, as in the case of a Lagrangian extremum problem the multipliers $(y_i^*,w_i^*)\in Y_{i-1}^{\perp_i}$ are unique, and the degrees of freedom brought into the representation by using extensions to the full space do not affect the values (since $R_{r-1+i}[F](z_*;H_{r-1})\in Y_i$ for $i=1,\dots,p-1$, and the freedom in the multipliers is only within functionals which annihilate Y_i). This proves the result. □

Remark. Note that we cannot assert in general that the multipliers in Theorem 8.3 are the same ones as the multipliers which arise in Theorem 6.1. In fact, if λ is not identically zero, they are not. If, however, $\lambda\equiv 0$, then (except for the possible but inconsequential freedom in the extensions) these multipliers are the same since $\mathcal{G}_p^*$ is $1-1$. This applies to the Lagrangian problem.

Corollary 8.4 (Lagrangian problems). *Consider the optimal control problem* (OC) *without control constraints ($U=\mathbb{R}^m$), and suppose the admissible process (x_*,u_*) is optimal. Let H_{p-1} be p-regular critical direction and let $a_i=a(H_{p-1})$, $i=0,1,\dots,p-1$, $\psi(\cdot)$ and $\rho_i(\cdot)$, $i=1,\dots,p-1$, be multipliers which satisfy the conditions of Theorem* 6.1 *along the optimal trajectory $(x_*(t),u_*(t),t)$ for $\nu_0=1$. Then* (8.11) *holds for any extension of H_{p-1} to an $(r-1)$-order tangent direction to the equality constraint which is r-order critical.*

Our theorems equally apply to the regular case, i.e., $p=1$, but then they reduce to classical results. Yet on this level the relations between these results are easiest explained. For $p=1$, Theorem 6.1 is the local maximum principle (e.g., [12, Theorem 12.1]) and the separation condition is simply the statement that there exists a supporting hyperplane to $\mathcal{U}$ at u_*. For the Lagrangian minimization problem and

$r = 2$, Theorem 8.3 gives the accessory problem. For, in this case, (8.11) becomes

$$
\begin{aligned}
0 &\le R_2[\ell](H_1) + a_0 R_2[q](H_1) + \int_0^T R_2[L](H_1) + \psi(t) R_2[f](H_1) dt \\
&= \frac{1}{2}\ell_{xx}(x_*(T))(\eta(T),\eta(T)) + \frac{1}{2}a_0 q_{xx}(x_*(T))(\eta(T),\eta(T)) \\
&\quad + \int_0^T (\eta,\xi) \begin{pmatrix} L_{xx} & L_{xu} \\ L_{ux} & L_{uu} \end{pmatrix} \begin{pmatrix} \eta \\ \xi \end{pmatrix} + \left\langle \psi, (\eta,\xi) \begin{pmatrix} f_{xx} & f_{xu} \\ f_{ux} & f_{uu} \end{pmatrix} \begin{pmatrix} \eta \\ \xi \end{pmatrix} \right\rangle dt \qquad (8.16) \\
&= \frac{1}{2}\ell_{xx}(x_*(T))(\eta(T),\eta(T)) + \frac{1}{2}a_0 q_{xx}(x_*(T))(\eta(T),\eta(T)) \\
&\quad + \int_0^T (\eta,\xi) \begin{pmatrix} H_{xx} & H_{xu} \\ H_{ux} & H_{uu} \end{pmatrix} \begin{pmatrix} \eta \\ \xi \end{pmatrix} dt,
\end{aligned}
$$

where $H = L + \psi f$. This holds for all $(\eta,\xi) \in \ker F'(x_*,u_*)$ for which $I'(x_*,u_*) = 0$. But $(\eta,\xi) \in \ker F'(x_*,u_*)$ if and only if η is a solution to the variational equation

$$
\dot\eta = f_x(x_*(t),u_*(t),t)\eta + f_u(x_*(t),u_*(t),t)\xi, \quad \eta(0) = 0, \tag{8.17}
$$

with terminal constraint $q_x(x_*(T))\eta(T) = 0$. These are the conditions for the domain of the accessory problem. In its typical formulation, however, no reference is made to the condition $I'(x_*,u_*) = 0$. But this is badly misleading. For, if this does not hold, then it follows immediately that (x_*,u_*) cannot be an extremal since it violates the local maximum principle. Thus the condition $I'(x_*,u_*) = 0$ is hidden in the statement that the accessory problem is taken along an extremal. In this sense our primal approach of determining high-order admissible directions results in the proper set of necessary conditions which these directions have to satisfy to be critical of a certain level. Then, if we choose $r = 3$, Theorem 8.3 gives additional necessary conditions for optimality, but only for those directions for which linear or quadratic approximations have been inconclusive, namely second-order critical directions. In this sense, our results provide a complete hierarchy of results which through primal constructions of higher-order approximating directions and dual characterizations of empty intersection properties of approximating cones give necessary conditions for optimality for increasingly more degenerate structures.

9. Examples. In this section we give several examples which illustrate how Theorem 6.1 can be used to eliminate abnormal candidates from optimality. We consider two cases. The first illustrates the case when $\mathrm{Im}(F'(z_*), I'(z_*))$ has codimension 1 and in the second example $\mathrm{Im}(F'(z_*), I'(z_*))$ has codimension 2. In these examples the order p will be arbitrary. However, we use the same equality constraint so that the analysis of high-order tangent directions only needs to be done once.

We consider the problems to minimize the functional

$$
I_\pm(x,u) = \int_0^T \left[(x_1 - 1)^2 + x_2^p + (x_3 \pm 1)^2 - 2\right] dt \tag{9.1}
$$

over all $(x,u) \in \overline{W}^3_{11}(0,T) \times L^2_\infty(0,T)$ subject to the dynamics

$$
\dot x(t) = A(x) + Bu = \begin{pmatrix} 0 \\ x_1^p \\ \alpha x_2^{p-1} x_3 \end{pmatrix} + \begin{pmatrix} 0 & 1 \\ -1 & 0 \\ 0 & -1 \end{pmatrix} \begin{pmatrix} u_1 \\ u_2 \end{pmatrix}, \tag{9.2}
$$

initial condition $x(0) = 0$, and terminal constraints $x_1(T) = 0$ and $x_3(T) = 0$. Here p is an integer, $p \geq 2$, and α is an arbitrary real number. For simplicity we have not imposed any control constraints. A related example which has bounded controls is given in [20].

We first show that the reference trajectory $\Gamma = (x_*, u_*) \equiv (0, 0)$ is an abnormal extremal for each problem. If Γ is optimal, then by the weak local maximum principle there exist a constant $\lambda_0 \geq 0$ and an absolutely continuous function $\lambda : [0, T] \to (\mathbb{R}^3)^*$ such that $(\lambda_0, \lambda(t)) \neq 0$ for all $t \in [0, T]$, $\dot\lambda(t) = \lambda_0(2, 0, \mp 2)$, $\lambda(T) = (\nu_1, 0, \nu_3)$, and the local minimum conditions $\lambda_2(t) \equiv 0$ and $\lambda_1(t) \equiv \lambda_3(t)$ are satisfied. For the problem to minimize I_+ we get

$$\lambda_1(t) - \lambda_3(t) = \nu_1 - \nu_3 + 4\lambda_0(t - T),$$

and this vanishes if and only if $\nu_1 = \nu_3$ and $\lambda_0 = 0$. Thus in this case Γ is a corank 1 abnormal extremal whose unique multiplier is given by $\lambda_0 = 0$ and $\lambda(t) \equiv (\nu, 0, \nu)$ for some nonzero constant ν. If, however, we minimize I_-, then we have

$$\lambda_1(t) - \lambda_3(t) = \nu_1 - \nu_3,$$

and the minimum condition $\lambda_1(t) \equiv \lambda_3(t)$ is automatically satisfied if $\nu_1 = \nu_3 = \nu$. In this case the extremal Γ has two linearly independent multipliers, one normal, the other abnormal, given by $\lambda_0 = 0$ and $\lambda(t) \equiv (1, 0, 1)$ and $\lambda_0 = 1$ and $\lambda(t) \equiv (2(t - T), 0, 2(t - T))$.

In either case the generalized Legendre–Clebsch condition [6] is trivially satisfied: since the multipliers are at most linear functions in t, we have for any positive integer k that

$$\frac{d^{2k}}{dt^{2k}} H_u(\lambda_0, \lambda(t), 0, 0) \equiv 0,$$

and thus the controls are singular of infinite order. For the case when there is a unique multiplier further necessary conditions for optimality of multi-input control systems are given by Goh [13]. In its simplified version [13, section 4.3] these conditions are satisfied trivially here since the control vector fields are constant and thus commute. The stronger condition stated in the "fundamental theorem" in [13] in addition requires that the matrix $B^T H_{xx}(0, \lambda(t), 0, 0,)B$ is positive semidefinite. For $p > 2$, this holds trivially since $H_{xx}(0, \lambda(t), 0, 0,) \equiv 0$, but for $p = 2$, this excludes the optimality of Γ. We have $H(0, \lambda, x, u,) = \lambda_2 x_1^2 + \lambda_3 \alpha x_2 x_3 + Bu$, and thus using $\lambda_2(t) \equiv 0$ and $\lambda_3(t) \equiv \nu$, we get that

$$\begin{aligned} B^T H_{xx}(0, \lambda(t), 0, 0,)B &= \begin{pmatrix} 0 & -1 & 0 \\ 1 & 0 & -1 \end{pmatrix} \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & \alpha\nu \\ 0 & \alpha\nu & 0 \end{pmatrix} \begin{pmatrix} 0 & 1 \\ -1 & 0 \\ 0 & -1 \end{pmatrix} \\ &= \alpha\nu \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}, \end{aligned} \tag{9.3}$$

which is indefinite for $\alpha \neq 0$. In this case the optimality of Γ can also be excluded using the result of Agrachev and Sarychev [1, Theorem 3.4] recalled in Proposition 5.2: It is clear from above that $\operatorname{Im} F'(0, 0)$ has codimension 1. More specifically, the variational equation is given by the linear system $\dot h(t) = B\xi$, $h(0) = 0$, and thus its reachable set is simply the image of B. Hence

$$\operatorname{Im} F'(0, 0) = \left\{ (a, b) \in \overline{W}{}^3_{11}(0, T) \times \mathbb{R}^2 : b^{[1]} + b^{[3]} = a^{[1]}(T) + a^{[3]}(T) \right\}, \tag{9.4}$$

where the superscripts denote the components. Also,

$$\ker F'(0,0) = \Big\{(\eta,\xi) \in \overline{W}^3_{11}(0,T) \times L^2_\infty(0,T) : \dot\eta(t) = B\xi, \eta(0) = 0, \\ \eta^{[1]}(T) = 0 \text{ and } \eta^{[3]}(T) = 0\Big\}. \tag{9.5}$$

Note that for $(\eta,\xi) \in \ker F'(0,0)$

$$I'_\pm(0,0)(\eta,\xi) = -2\int_0^T \eta^{[1]}(t) \mp \eta^{[3]}(t)dt = -2(1 \pm 1)\int_0^T \eta^{[1]}(t)dt, \tag{9.6}$$

and thus we have $I'_-(0,0)(\eta,\xi) \equiv 0$ for all $(\eta,\xi) \in \ker F'(0,0)$ while this quantity can be made nonzero for I_+. Hence the codimension of $\operatorname{Im}(F'(0,0), I'_-(0,0))$ is 2, while the codimension of $\operatorname{Im}(F'(0,0), I'_+(0,0))$ is 1. A nontrivial annihilator λ^* of $\operatorname{Im} F'(0,0)$ is determined by the abnormal costate λ of the weak maximum principle, and since $\lambda(t) = (\nu, 0, \nu)$, it follows that

$$\begin{aligned}\lambda^* F''(0,0)((\eta,\xi);(\eta,\xi)) &= \langle y^*, F_1''(0,0)\,((\eta,\xi),(\eta,\xi))\rangle + \langle w^*, F_2''(0,0)\,((\eta,\xi),(\eta,\xi))\rangle \\ &= \int_0^T \lambda(t)A_{xx}(0)(\eta(t),\eta(t))dt + 0 \\ &= \nu\int_0^T 2\alpha\eta^{[2]}(s)\eta^{[3]}(s)ds, \end{aligned} \tag{9.7}$$

where $\eta^{[i]}$ denotes the ith component of η. This quantity can be made both positive and negative for $(\eta,\xi) \in \ker F'(0,0)$. For instance, choose $\eta^{[3]}(t) = \sin(\pi\frac{t}{T})$ and $\eta^{[2]}(t) = \pm\eta^{[3]}(t)$. Hence $\lambda^* F''(0,0)$ is indefinite on $\ker F'(0,0)$, and thus Γ is not optimal.

For $p > 2$ or for the problem to minimize I_-, however, these results do not apply. We now show how Theorem 6.1 can be used to obtain additional information. We first consider the problem to minimize I_+ and show how to eliminate the optimality of Γ for any $p \geq 2$ using Theorem 6.1. For this we first need to find an appropriate p-regular critical direction

$$H_{p-1} = ((\eta_1,\xi_1);\dots;(\eta_{p-1},\xi_{p-1})) \in \Big(\overline{W}^3_{11}(0,T) \times L^2_\infty(0,T)\Big)^{p-1}.$$

Since the components of f are homogeneous polynomials of degree p and the terminal constraints are linear, it follows that for $i = 1,\dots,p-1$, the directional derivatives $\nabla^i F(z_*)(H_i)$ are simply given by

$$\nabla^i F(z_*)(H_i) = F'(0,0)(\eta_i,\xi_i), \tag{9.8}$$

and thus these derivatives vanish provided $(\eta_i,\xi_i) \in \ker F'(0,0)$ for $i = 1,\dots,p-1$. Since the dynamics involves only p-order terms it seems reasonable to consider only first- and pth-order approximations. We therefore choose H_{p-1} of the form

$$H_{p-1} = ((\eta_1,\xi_1);(0,0);\dots;(0,0)) \tag{9.9}$$

with $(\eta_1,\xi_1) \in \ker F'(0,0)$. With this choice of directions the compatibility conditions (ii) simplify considerably and reduce to the first condition only,

$$R_p[F](\Gamma;H_{p-1}) = \frac{1}{p!}F^{(p)}(0,0)\,((\eta_1,\xi_1);\dots;(\eta_1,\xi_1)) \in \operatorname{Im} F'(0,0). \tag{9.10}$$

The conditions for $i = 2, \dots, p-1$, are satisfied since $R_{p-1+i}[F](\Gamma; H_{p-1}) = 0$. Condition (9.10) is equivalent to

$$(1,0,1) \cdot \int_0^T D_x^{(p)} A(0)(\eta_1, \dots, \eta_1) ds = 0,$$

where $D_x^{(p)} A(0)(\eta_1, \dots, \eta_1)$ denotes the 3-vector whose ith entry is given by the values of the multilinear form $D_x^{(p)} A_i(0)$ acting on the vector η_1 in every component. In particular, since $A_1(x) \equiv 0$ in the example, we must therefore have

$$\int_0^T D_x^{(p)} A_3(0)(\eta_1, \dots, \eta_1) ds = 0. \tag{9.11}$$

If, as above, we denote the components of $\eta_1$by $\eta_1^{[j]}$, $j = 1, 2, 3$, then for $A_3(x) = \alpha x_2^{p-1} x_3$ this requires that

$$\int_0^T \left(\eta_1^{[2]}\right)^{p-1} \left(\eta_1^{[3]}\right) ds = 0. \tag{9.12}$$

We satisfy this by choosing $\eta_1^{[3]} = -\eta_1^{[1]} \equiv 0$ (i.e., $\xi_1^{[2]} \equiv 0$). Then choosing a nonzero $\eta_1^{[2]}$ with zero boundary conditions defines a nontrivial vector H_{p-1} of the form (9.9) for which conditions (i) and (ii) in the definition of p-regular critical directions are satisfied. The operator F is p-regular in this direction if the operator $\mathcal{G}_p(\Gamma; H_{p-1})$ is onto. Because of the homogeneity properties of A, the pth component G_p of $\mathcal{G}_p$ is given by

$$G_p[F](\Gamma; H_{p-1})(\widetilde{\eta}, \widetilde{\xi}) = \frac{1}{(p-1)!} F^{(p)}(0,0) \left((\eta_1, \xi_1); \dots; (\eta_1, \xi_1); (\widetilde{\eta}, \widetilde{\xi}) \right),$$

while all operators G_i for $i = 2, \dots, p-1$, vanish. Thus $\mathcal{G}_p(\Gamma; H_{p-1})$ is onto if there exists a direction $(\widetilde{\eta}, \widetilde{\xi}) \in \overline{W}^3_{11}(0,T) \times L^2_\infty(0,T)$ such that $G_p[F](\Gamma; H_{p-1})(\widetilde{\eta}, \widetilde{\xi}) \notin \operatorname{Im} F'(0,0)$. Like above, this is equivalent to

$$\int_0^T D_x^{(p)} A_3(0)(\eta_1, \dots, \eta_1, \widetilde{\eta}) ds \neq 0, \tag{9.13}$$

and thus

$$\int_0^T \left(\eta_1^{[2]}\right)^{p-2} \left[\left(\eta_1^{[2]}\right) \left(\widetilde{\eta}^{[3]}\right) + (p-1) \left(\eta_1^{[3]}\right) \left(\widetilde{\eta}^{[2]}\right) \right] ds \neq 0. \tag{9.14}$$

Since we have $\eta_1^{[3]} \equiv 0$ in our chosen direction H_{p-1}, this can simply be satisfied by choosing $\widetilde{\eta}^{[3]} = \eta_1^{[2]}$ if p is even or $\widetilde{\eta}^{[3]} = (\eta_1^{[2]})^2$ if p is odd. Hence F is p-regular in direction of H_{p-1} at Γ. Finally, these directions are also critical for the objective: by (9.6) we have $I'_+(0,0)(\eta_1, \xi_1) = 0$ and furthermore

$$\begin{aligned} \nabla^2 I_+(0,0)(H_2) &= \frac{1}{2} I''_+(0,0)((\eta_1, \xi_1); (\eta_1, \xi_1)) \\ &= \int_0^T \left(\eta_1^{[1]}\right)^2 + \left(\eta_1^{[3]}\right)^2 ds = 0, \end{aligned}$$

provided $p > 2$. Since no other I_+-derivatives arise in the directional derivatives $\nabla^i I_+(0,0)(H_i)$ for $i = 3, \dots, p-1$, the direction $H_{p-1} = ((\eta_1, \xi_1); (0,0); \dots; (0,0))$ with $\eta_1^{[1]} = \eta_1^{[3]} \equiv 0$ and a nonzero $\eta_1^{[2]}$ is a nonzero p-regular critical direction for the problem to minimize I_+ subject to $F = 0$ for any $p \geq 2$.

We thus can apply Theorem 6.1. Since there are no control constraints we can normalize the multipliers so that $\nu_0 = 1$. The additional multipliers ρ_i, $i = 1, \dots, p-1$, are associated with elements in the dual spaces of the quotients Y_{i+1}/Y_i. But here $Y_i = \operatorname{Im} F'(0,0)$ for $i = 1, \dots, p-1$, and Y_p is the full space. Thus we have $\rho_i \equiv 0$ for $i = 1, \dots, p-2$, and the only nonzero multipliers are ψ and ρ_{p-1} which for simplicity of notation we just call ρ. Now (6.4) of Theorem 6.1 states that ρ is an adjoint multiplier for which the conditions of the weak local maximum principle for an abnormal extremal are satisfied. Since this multiplier is unique, we must have $\rho(t) = (\nu, 0, \nu)$, but $\nu \in \mathbb{R}$ could now be zero. To write down the extended adjoint equation and minimum condition (6.7), we need to evaluate the directional derivatives $\nabla^{p-1} f(x,u)(H_i)$, where $f(x,u) = A(x) + Bu$. Note that all partial derivatives of order at least two which contain one u-derivative vanish. Thus we need only to calculate the actual x-partials. The leading term in $\nabla^{p-1} f(x,u)(H_i)$ is therefore given by

$$D_x^{(p-1)} A(x)(\eta_1, \dots, \eta_1) = \begin{pmatrix} 0 \\ p! \left(\eta_1^{[1]}\right)^{p-1} x_1 \\ (p-1)! \left[(p-1) \left(\eta_1^{[2]}\right)^{p-2} \left(\eta_1^{[3]}\right) x_2 + \left(\eta_1^{[2]}\right)^{p-1} x_3 \right] \end{pmatrix}.$$

All other terms come from lower derivatives and contain at least quadratic terms in the x_i. After taking another derivative and evaluating at zero all these terms will vanish. Since we also have $\eta_1^{[1]} = \eta_1^{[3]} \equiv 0$, we therefore get

$$\left(\nabla^{p-1} f(0,0)(H_i)\right)_x = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & \left(\eta_1^{[2]}\right)^{p-1} \end{pmatrix} \tag{9.15}$$

and

$$\left(\nabla^{p-1} f(0,0)(H_i)\right)_u \equiv 0. \tag{9.16}$$

Thus the extended minimum condition reduces to $\psi B \equiv 0$, the minimum condition of the weak maximum principle. Hence also $\psi_2(t) \equiv 0$ and $\psi_1(t) = \psi_3(t)$. But now the extended adjoint equation is given by

$$\dot{\psi}(t) = (2, 0, -2) - \rho \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & \left(\eta_1^{[2]}\right)^{p-1} \end{pmatrix}, \tag{9.17}$$

and thus

$$4 = \dot{\psi}_1(t) - \dot{\psi}_3(t) - \nu \left(\eta_1^{[2]}(t)\right)^{p-1} = -\nu \left(\eta_1^{[2]}(t)\right)^{p-1}. \tag{9.18}$$

But we can certainly choose $\eta_1^{[2]}$ nonconstant to violate this condition. This contradiction proves that Γ cannot be optimal for the problem to minimize I_+ for any $p \geq 2$.

We now consider the problem to minimize I_- for which the codimension of $\mathrm{Im}(F'(0,0), I'_-(0,0))$ is 2. This becomes a harder problem in some aspects but an easier problem in others. Essentially, since there exist two linearly independent multipliers which satisfy the weak local maximum principle, conditions (a)–(c) of Theorem 6.1 can be satisfied in a straightforward way using these multipliers where (ν_0, ψ) is the normal multiplier and $\rho \equiv 0$. Hence these conditions will give nothing new. On the other hand, since the codimension is 2, the restrictions on the critical directions are less stringent since any direction $(\eta, \xi) \in \ker F'(0,0)$ is automatically critical for the objective as well. Hence more directions are critical and the separation condition (d) becomes stronger. The details follow.

We first assume $p > 2$ and consider the same p-regular tangent directions as above. Note that these directions remain p-order critical for the objective and thus are p-regular critical directions. The only difference to the analysis above is that the extended adjoint equation now reads

$$\dot{\psi}(t) = (2,0,2) - \rho \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & \left(\eta_1^{[2]}(t)\right)^{p-1} \end{pmatrix}, \tag{9.19}$$

and thus we get

$$0 = \dot{\psi}_1(t) - \dot{\psi}_3(t) - \nu \left(\eta_1^{[2]}(t)\right)^{p-1} = -\nu \left(\eta_1^{[2]}(t)\right)^{p-1}. \tag{9.20}$$

But this equation can be satisfied with $\nu = 0$ implying $\rho \equiv 0$. The separation condition (d) reduces to

$$0 \le \int_0^T R_p[L](H_{p-1}) + \psi(t) R_p[f](H_{p-1}) dt. \tag{9.21}$$

Since $\eta_1^{[1]} \equiv \eta_1^{[3]} \equiv 0$, we have

$$R_p[f](H_{p-1}) = \begin{pmatrix} 0 \\ \left(\eta_1^{[2]}\right)^p \\ 0 \end{pmatrix}, \tag{9.22}$$

and this term will be annihilated by $\psi_2(t) \equiv 0$; and $R_p[L](H_{p-1})$ only generates the term $(\eta_1^{[2]}(t))^p$ as nonzero term. Thus it is a necessary condition for optimality of Γ that

$$0 \le \int_0^T \left(\eta_1^{[2]}(t)\right)^p dt \tag{9.23}$$

for any p-regular critical direction which satisfies $\eta_1^{[1]} \equiv \eta_1^{[3]} \equiv 0$. But we can multiply these directions by -1, and thus Γ is not optimal if p is odd.

If p is even, (9.23) will always be satisfied and these directions do not give rise to better values of the objective. In fact, if $H_{p-1} = ((\eta_1, \xi_1), (\eta_2, \xi_2), \dots, (\eta_{p-1}, \xi_{p-1}))$ is any p-order tangent direction for which $\eta_1^{[3]} \neq 0$, then, since by (9.8) necessarily $(\eta_2, \xi_2) \in \ker F'(0,0)$, we get

$$\begin{aligned} \nabla^2 I_-(0,0)(H_2) &= I'_-(0,0)(\eta_2, \xi_2) + \frac{1}{2} I''_-(0,0)((\eta_1, \xi_1), (\eta_1, \xi_1)) \\ &= \int_0^T \left(\eta_1^{[1]}\right)^2 + \left(\eta_1^{[3]}\right)^2 ds > 0. \end{aligned} \tag{9.24}$$

Hence no p-order tangent direction can lead to an improved value of the objective. Indeed, it can be shown that no improvement of the objective is possible for any tangent direction. These calculations strongly suggest that Γ is locally optimal if p is even and $p \geq 4$, but we are not aware of sufficient conditions for optimality which would apply to this situation.

The case $p = 2$ is the most interesting one. As above, in this case 2-regular critical directions for which $\eta_1^{[1]} \equiv \eta_1^{[3]} \equiv 0$ cannot be used to exclude the optimality of Γ. However, now any direction in $\ker F'(0,0)$ is critical, and thus any nontrivial $(\eta, \xi) \in \ker F'(0,0)$ which satisfies the compatibility condition

$$\int_0^T \eta^{[2]}(s)\eta^{[3]}(s)ds = 0 \tag{9.25}$$

defines a 2-regular critical direction. It is easily seen from (9.14) that the operator F is 2-regular at Γ in direction of (η, ξ) by choosing $\widetilde{\eta}^{[3]} = \eta_1^{[2]}$ and $\widetilde{\eta}^{[2]} = \eta_1^{[3]}$. Hence any direction $(\eta, \xi) \in \ker F'(0,0)$ which satisfies (9.25) is a 2-regular critical direction. The fact that $\operatorname{co\,dim}(\operatorname{Im} I'_-(0,0), \operatorname{Im} F'(0,0)) = 2$ allows for this *large* class of critical directions H_1. The separation condition (d) of Theorem 6.1 in conjunction with Corollary 7.2 and the fact that $\rho \equiv 0$ therefore implies that the quadratic form

$$\mathcal{Q} = \frac{1}{2}\int_0^T L^{(2)}(0,0)(H_1, H_1) + \psi(t)f^{(2)}(0,0)(H_1, H_1)dt \tag{9.26}$$

is positive semidefinite on the set of all 2-regular critical directions H_1. Note that (9.26) takes the form of the accessory problem for the normal multiplier $(1, \psi)$, but that the domain is restricted to the actual tangent directions to the equality constraint. In this case we have $\psi_3(t) = \nu + 2(t-T)$, $\nu = \psi_3(T)$. Note that in accordance with our remark on the uniqueness of multipliers the one degree of freedom in the multipliers is taken by the multipliers $(\nu, 0, \nu)$ from the annihilator of $\operatorname{Im} F'(0,0)$. But, as stated in Corollary 7.2, this freedom does not enter into the value of the quadratic form:

$$\begin{aligned}
\mathcal{Q} &= \int_0^T \left(\eta^{[1]}\right)^2 + \left(\eta^{[2]}\right)^2 + \left(\eta^{[3]}\right)^2 + \psi_3(t)\alpha\eta^{[2]}\eta^{[3]}dt \\
&= \int_0^T \left(\eta^{[2]}\right)^2 + 2\left(\eta^{[3]}\right)^2 + (\nu + 2(t-T))\,\alpha\eta^{[2]}\eta^{[3]}dt \\
&= \int_0^T \left(\eta^{[2]}\right)^2 + 2\left(\eta^{[3]}\right)^2 + 2t\alpha\eta^{[2]}\eta^{[3]}dt. \qquad (9.27)
\end{aligned}$$

Regardless of the value of ν, the quadratic form $\mathcal{Q}$ takes the same value for all possible multipliers because of the compatibility condition (9.25).

Now we pick a suitable subset of 2-regular critical directions H_1. Let C denote the space of all twice continuously differentiable functions $h : [0,T] \to \mathbb{R}$ which satisfy zero boundary conditions $h(0) = h(T) = 0$. Since

$$\int_0^T \dot{h}(t)h(t)dt = \frac{1}{2}h(t)^2\Big|_0^T = 0,$$

C is isomorphic to a subspace of critical directions defined by

$$\eta^{[3]}(t) = -\eta^{[1]}(t) = h(t), \quad \eta^{[2]}(t) = \dot{h}(t) \tag{9.28}$$

and

$$\xi^{[1]}(t) = \ddot{h}(t), \quad \xi^{[2]}(t) = -\dot{h}(t). \tag{9.29}$$

Hence it is a necessary condition for optimality that the quadratic form

$$\begin{aligned} \mathcal{Q} &= \int_0^T \left(\dot{h}(t)\right)^2 + 2\left(h(t)\right)^2 + 2\alpha t\dot{h}(t)h(t)dt \\ &= \int_0^T \left(\dot{h}(t)\right)^2 + (2-\alpha)\left(h(t)\right)^2 dt \end{aligned} \tag{9.30}$$

is positive semidefinite on C. It follows from the Jacobi equation [10] that $\mathcal{Q}$ is positive definite on C for $\alpha < 2 + \left(\frac{T}{\pi}\right)^2$, positive semidefinite for $\alpha = 2 + \left(\frac{T}{\pi}\right)^2$, and indefinite for $\alpha > 2 + \left(\frac{T}{\pi}\right)^2$. Thus Γ is not optimal for $\alpha > 2 + \left(\frac{T}{\pi}\right)^2$.

10. Conclusion. These examples illustrate the hierarchy of results derived in this paper which give necessary conditions for optimality of abnormal extremals. This extension is nontrivial in the sense that the multiplier at the objective does not vanish. The results apply regardless of whether the multiplier is unique or not, whether additional multipliers are normal or not, etc. The key idea behind our results is to characterize the directions which are actually tangent to the constraint set, and this question has been answered conclusively in [21]. Hence it is possible to set up high-order constructions. Naturally, near abnormal points, the resulting necessary conditions for optimality will take a different look since the constraint typically is no longer a manifold and intersecting branches need to be analyzed separately. Our constructions provide an approach for doing this.

The local version of the extended maximum principle presented in this paper, besides its intrinsic interest for Lagrangian problems when the reference control takes values in the interior of the control set, can also be considered a first step in the derivation of a more general version analogous to the Pontryagin maximum principle. This theorem will apply to more general problems in the sense that the control set can be arbitrary and consequently no differentiability assumptions need to be made on the control. Also the terminal time will be free. For the case $p = 2$, this transition, which is accomplished by a technique of variable time transformations (introduced by Dubovitskii in this context) has already been carried out in [18]. Its generalization to the general p-order maximum principle is a logical next step. This result has already been formulated in [20] with an outline of the proof.

Acknowledgments. We would like to thank the associate editor who was handling this paper and two anonymous referees for many valuable suggestions which helped us clarify our results.

REFERENCES

[1] A.A. Agrachev and A.V. Sarychev, *On abnormal extremals for Lagrange variational problems*, J. Math. Systems Estim. Control, 5 (1995), pp. 127–130.

[2] A.A. Agrachev and A.V. Sarychev, *Abnormal sub-Riemannian geodesics: Morse index and rigidity*, Ann. Inst. H. Poincaré Anal. Non Linéaire, 13 (1996), pp. 635–690.

[3] A.V. Arutyunov, *Optimality conditions in abnormal extremal problems*, Systems Control Lett., 27 (1996), pp. 279–284.

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

[5] E.R. Avakov, *Necessary extremum conditions for smooth abnormal problems with equality and inequality-type constraints*, Math. Zametki, 45 (1989), pp. 3–11.
[6] A.E. Bryson and Y.C. Ho, *Applied Optimal Control*, Hemisphere Publishing, Washington, DC, 1975.
[7] C. Caratheodory, *Variationsrechnung und partielle Differentialgleichungen erster Ordnung*, Teubner, Germany, 1935.
[8] A.V. Dmitruk, *Quadratic order conditions of a local minimum for abnormal extremals*, in Proceedings of the Second World Congress of Nonlinear Analysts, Part 4, Athens, 1996, Nonlinear Anal., 30 (1998), pp. 2439–2448.
[9] R. Gabasov and F.M. Kirillova, *High order necessary conditions for optimality*, SIAM J. Control, 10 (1972), pp. 127–168.
[10] I.M. Gelfand and S.V. Fomin, *Calculus of Variations*, Prentice Hall, Englewood Cliffs, NJ, 1963.
[11] E.G. Gilbert and D.S. Bernstein, *Second-order necessary conditions in optimal control: accessory-problem results without normality conditions*, J. Optim. Theory Appl., 41 (1983), pp. 75–106.
[12] I.V. Girsanov, *Lectures on Mathematical Theory of Extremum Problems*, Springer-Verlag, Berlin, Heidelberg, New York, 1972.
[13] B.S. Goh, *Necessary conditions for singular extremals involving multiple control variables*, SIAM J. Control, 4 (1966), pp. 716–731.
[14] A.D. Ioffe and V.M. Tihomirov, *Theory of Extremal Problems*, North-Holland, Amsterdam, 1979.
[15] A.F. Izmailov, *Optimality conditions for degenerate extremum problems with inequality-type constraints*, Comput. Maths. Math. Phys., 34 (1994), pp. 723–736.
[16] A.J. Krener, *The high order maximal principle and its application to singular extremals*, SIAM J. Control Optim., 15 (1977), pp. 256–293.
[17] U. Ledzewicz and H. Schättler, *Second order conditions for extremum problems with non-regular equality constraints*, J. Optim. Theory Appl., 86 (1995), pp. 113–144.
[18] U. Ledzewicz and H. Schättler, *An extended maximum principle*, Nonlinear Anal., 29 (1997), pp. 59–183.
[19] U. Ledzewicz and H. Schättler, *A high-order generalization of the Lyusternik theorem and its application to optimal control problems,* in Dynamical Systems and Differential Equations II, W. Chen and S. Hu, eds., Southwest Missouri State University, Springfield, MO, 1998, pp. 45–59.
[20] U. Ledzewicz and H. Schättler, *High order extended maximum principles for optimal control problems with non-regular constraints*, in Optimal Control: Theory, Algorithms and Applications, W.W. Hager and P.M. Pardalos, eds., Kluwer Academic Publishers, Dordrecht, The Netherlands, 1998, pp. 298–325.
[21] U. Ledzewicz and H. Schättler, *A high-order generalization of the Lyusternik theorem*, Nonlinear Anal., 34 (1998), pp. 793–815.
[22] U. Ledzewicz and H. Schättler, *High-order approximations and generalized necessary conditions for optimality*, SIAM J. Control Optim., 37 (1999), pp. 33–53.
[23] W.S. Liu and H.J. Sussmann, *Abnormal Sub-Riemannian minimizers*, in Differential Equations, Dynamical Systems and Control Science, K.D. Elworthy, W.N. Everitt, and E.B. Lee, eds., Marcel Dekker, New York, 1993, pp. 705–716.
[24] W.S. Liu and H.J. Sussmann, *Shortest Paths for Sub-Riemannian Metrics on Rank-2 Distributions*, Mem. Amer. Math. Soc. 118, AMS, Providence, RI, 1995.
[25] A.A. Milyutin, *Quadratic conditions of an extremum in smooth problems with a finite-dimensional image*, in Methods of the Theory of Extremal Problems in Economics, Nauka, Moscow, 1981, pp. 138–177 (in Russian).
[26] Z. Páles and V. Zeidan, *Nonsmooth optimum problems with constraints*, SIAM J. Control Optim., 32 (1994), pp. 1476–1502.
[27] Z. Páles and V. Zeidan, *First and second order necessary conditions for control problems with constraints*, Trans. Amer. Math. Soc., 346 (1994), pp. 421–453.
[28] L.S. Pontryagin, V.G. Boltyanskii, R.V. Gamkrelidze, and E.F. Mishchenko, *The Mathematical Theory of Optimal Processes*, John Wiley, New York, 1962.
[29] G. Stefani and P.L. Zezza, *Optimality conditions for a constrained control problem*, SIAM J. Control Optim., 34 (1996), pp. 635–659.
[30] H.J. Sussmann, *Geometry and optimal control*, in Mathematical Control Theory, Festschrift in Honor of Roger Brockett, J. Baillieul and J.C. Willems, eds., Springer Verlag, New York, 1998, pp. 140–198.

SIAM J. CONTROL OPTIM.
Vol. 38, No. 3, pp. 855–874

STABILIZATION OF NONHOLONOMIC SYSTEMS USING ISOSPECTRAL FLOWS*

ANTHONY M. BLOCH†, SERGEY V. DRAKUNOV‡, AND MICHAEL K. KINYON§

Abstract. In this paper we derive and analyze a discontinuous stabilizing feedback for a Lie algebraic generalization of a class of kinematic nonholonomic systems introduced by Brockett [*New Directions in Applied Mathematics*, P. Hilton and G. Young, eds., Springer-Verlag, New York, 1982, pp. 11–27]. The algorithm involves discrete switching between isospectral and norm-decreasing flows. We include a rigorous analysis of the convergence.

Key words. nonlinear control, nonholonomic systems, isospectral flows, Lie theory

AMS subject classifications. 93E15, 93B25

PII. S0363012998335607

1. Introduction.

In this paper we present a stabilization algorithm for a Lie algebraic generalization of a class of nonholonomic systems originally introduced by Brockett [11]. These are the systems of the general form $\dot{x} = B(x)u$ in which the dimension of the control vector u is smaller than that of the state vector x, but the system is nonetheless controllable. Such systems are sometimes referred to as kinematic nonholonomic systems (as opposed to dynamic nonholonomic systems which arise from a Lagrangian—see, e.g., Bloch and Crouch [2] or Bloch, Krishnaprasad, Marsden, and Murray [9]). Following standard usage, we will simply refer to kinematic nonholonomic systems as nonholonomic.

A prototypical system in the class we study here is the Heisenberg system or nonholonomic integrator (Brockett [11], [12]):

$$\dot{x} = u, \tag{1.1}$$

$$\dot{y} = v, \tag{1.2}$$

$$\dot{z} = xv - yu, \tag{1.3}$$

where $x, u, y, v, z \in \mathbb{R}$. If we identify the variable z with the skew-symmetric matrix $Y = \left(\begin{smallmatrix} 0 & z \\ -z & 0 \end{smallmatrix}\right)$, and observe that

$$\begin{pmatrix} x \\ y \end{pmatrix}(u, v) - \begin{pmatrix} u \\ v \end{pmatrix}(x, y) = \begin{pmatrix} 0 & xv - yu \\ -xv + yu & 0 \end{pmatrix},$$

then a generalization of (1.1)–(1.3) immediately suggests itself. This is the $so(n)$ system (see Brockett [11]):

$$\dot{x} = u, \tag{1.4}$$

*Received by the editors March 16, 1998; accepted for publication (in revised form) May 18, 1999; published electronically March 8, 2000.

http://www.siam.org/journals/sicon/38-3/33560.html

†Department of Mathematics, The University of Michigan, Ann Arbor, MI 48109 (bloch@math.lsa.umich.edu). The research of this author was supported in part by NSF grants DMS-9496221 and DMS-9803181, AFOSR grant F49620-96-1-0100, a Guggenheim Fellowship and the Institute for Advanced Study.

‡Department of Electrical Engineering & Computer Science, Tulane University, New Orleans, LA 70118 (drakunov@eecs.tulane.edu). The research of this author was supported in part by NSF grant ECS-9631321 and a Lilly Endowed Fellowship.

§Department of Mathematics & Computer Science, Indiana University South Bend, South Bend, IN 46634-7111 (mkinyon@iusb.edu).

$$\dot{Y} = xu^T - ux^T, \tag{1.5}$$

where x, u are column vectors in $\mathbb{R}^n$ and $Y \in so(n)$, $n \geq 2$. Here $so(n)$ is the Lie algebra of $n \times n$ skew-symmetric matrices: $Y^T = -Y$.

The importance of the $so(n)$ system (1.4)–(1.5) is that it is a canonical form for a class of controllable systems of the form $\dot{x} = B(x)u$, $u \in \mathbb{R}^n$, $x \in \mathbb{R}^{n(n+1)/2}$. The class in question is the controllable systems of this type, where the first derived algebra of control vector fields spans the tangent space $T\mathbb{R}^{n(n+1)/2}$ at any point. (Recall that if E^0 is the subbundle of the tangent bundle spanned by the control fields, then the first derived algebra is given by $E^1 = E^0 + [E^0, E^0]$.) Brockett showed that such a system can be transformed to the form (1.4)–(1.5) up to a suitable order in the neighborhood of a given point such as the origin (see Brockett [11]). (In the terminology of Bloch, Reyhanoglu, and McClamroch [10], for example, this is a controllable nonholonomic kinematic system of nonholonomy degree 1.)

A different generalization of the Heisenberg system (1.1)–(1.3) is obtained by identifying the vectors $(x, y)^T$ and $(u, v)^T$ with the matrices $X = \frac{1}{\sqrt{2}}\left(\begin{smallmatrix} x & -y \\ -y & -x \end{smallmatrix}\right)$ and $U = \frac{1}{\sqrt{2}}\left(\begin{smallmatrix} u & -v \\ -v & -u \end{smallmatrix}\right)$, respectively. Then we have

$$[U, X] = UX - XU = \begin{pmatrix} 0 & xv - yu \\ -(xv - yu) & 0 \end{pmatrix}.$$

This suggests the following matrix system occurring in the Lie algebra $sl(n, \mathbb{R})$ of $n \times n$ matrices with trace 0:

$$\dot{X} = U, \tag{1.6}$$

$$\dot{Y} = [U, X], \tag{1.7}$$

where $X, U \in sym_0(n, \mathbb{R})$ and $Y \in so(n)$. Here $sym_0(n, \mathbb{R})$ is the space of $n \times n$ real symmetric matrices with trace zero. Note that $sl(n, \mathbb{R}) = sym_0(n, \mathbb{R}) \oplus so(n)$, a direct sum.

The system we study in this paper generalizes both the $so(n)$ system (1.4)–(1.5) and the $sl(n, \mathbb{R})$ system (1.6)–(1.7). Let $\mathfrak{g}$ be a Lie algebra. Assume $\mathfrak{g}$ has a direct sum decomposition $\mathfrak{g} = \mathfrak{m} \oplus \mathfrak{h}$ such that $\mathfrak{h}$ is a Lie subalgebra, $[\mathfrak{h}, \mathfrak{m}] \subseteq \mathfrak{m}$, and $[\mathfrak{m}, \mathfrak{m}] = \mathfrak{h}$. The exact hypotheses will be given in section 3; for now, we note that every simple Lie algebra with a Cartan decomposition is of this type. We will consider the following system in $\mathfrak{g}$:

$$\dot{x} = u, \tag{1.8}$$

$$\dot{Y} = [u, x], \tag{1.9}$$

where $x, u \in \mathfrak{m}$, $Y \in \mathfrak{h}$.

Clearly the $sl(n, \mathbb{R})$ system (1.6)–(1.7) has the form (1.8)–(1.9). In addition, the $so(n)$ system (1.4)–(1.5) can also be written in this form as we now show. Let $\mathfrak{h} = so(n)$ and let $\mathfrak{m} = \mathbb{R}^n$. For $x, u \in \mathfrak{m}$, define $[u, x] \equiv xu^T - ux^T \in \mathfrak{h}$. For $Y \in \mathfrak{h}$, $x \in \mathfrak{m}$, define $[Y, x] = -[x, Y] \equiv Yx$. Then the Lie algebra $\mathfrak{g} \equiv \mathfrak{m} + \mathfrak{h}$ is isomorphic to $so(n + 1)$. Indeed, this is clear if we make the identifications

$$\mathfrak{h} \cong \left\{ \begin{pmatrix} 0 & 0 \\ 0 & Y \end{pmatrix} : Y \in so(n) \right\} \quad \text{and} \quad \mathfrak{m} \cong \left\{ \begin{pmatrix} 0 & -x^T \\ x & 0 \end{pmatrix} : x \in \mathbb{R}^n \right\}.$$

It is easy to check that the desired commutation relations hold, and, with these particular identifications, that the adjoint action of $\mathfrak{h}$ on $\mathfrak{m}$ agrees with the standard

action of $so(n)$ on $\mathbb{R}^n$. (Incidentally, that the Heisenberg system (1.1)–(1.3) can be viewed as either an $sl(2,\mathbb{R})$ system or an $so(3)$ system is just a consequence of the fact that $sl(2,\mathbb{R})$ and $so(3)$ are isomorphic Lie algebras.)

We have already noted that the $so(n)$ system (1.4)–(1.5) is a canonical form for those controllable nonholonomic systems for which the first derived algebra $E^1 = E^0 + [E^0, E^0]$ of the control subbundle E^0 spans the tangent space at each point and for which the dimensions of E^0 and E^1 are arithmetically related in a particular way (see Brockett [11]). As partial motivation for considering the more general Lie algebra system (1.8)–(1.9), we expect that this system will turn out to be a canonical form for a wider class of controllable nonholonomic systems satisfying conditions which suitably generalize those characterizing the $so(n)$ system, including certain symmetry conditions. In particular, we suspect that (1.8)–(1.9) is a canonical form for nonholonomic systems occurring in certain homogeneous spaces, such as symmetric spaces.

The problem we consider herein is that of finding a stabilizing control for the general system (1.8)–(1.9). Since the dimension of $\mathfrak{m}$, which is where the control u takes its values, is less than the dimension of the state space $\mathfrak{g}$, the system fails Brockett's necessary condition for the existence of a continuous feedback law (see Brockett [12]). Previous work on the stabilization problem for nonholonomic systems such as the Heisenberg and $so(n)$ systems has focused mainly on the development of either smooth dynamic feedback or nonsmooth static feedback. For the former approach, see Coron [15] and also Pomet [28] and M'Closkey and Murray [25]. A discontinuous, discrete-time approach can be found in Bloch, Reyhanoglu, and McClamroch [10] and Kolmanovsky and McClamroch [23]. Brockett [14], on the other hand, used a stochastic approach. Other interesting work includes that of Liu and Sussmann [24], Sontag [30], Hespanha [21], Khennouf and Canudas de Wit [22], Morse [26], and others. The role of differential flatness in these systems is also interesting; see the work of Sira-Ramirez [29] and of Fliess et al. [19].

In this paper we present a new discontinuous feedback law and an algorithm for its implementation. We also give a rigorous analysis of the convergence of the algorithm. This completely solves the stabilization problem for (1.8)–(1.9), which is more general than the systems considered by others. We stress, however, that our results are new even for the well-known $so(n)$ system (1.4)–(1.5). Our present results are related to our earlier work on the Heisenberg system (1.1)–(1.3) (see Bloch and Drakunov [4], [5], [6]), the $so(n)$ system (1.4)–(1.5) (see Bloch and Drakunov [7]), and the general system (1.8)–(1.9) (see Bloch, Drakunov and Kinyon [8]). However, this paper is not so much an extension of our earlier work as it is a completely new approach to the stabilization problem. In addition we provide insights into the natural geometric structure of the problem.

Stabilizing (1.8)–(1.9), even locally, is a nontrivial task, since, as can be easily seen, linearization in the vicinity of the origin $0 \in \mathfrak{g}$ gives the noncontrollable system

$$\begin{aligned} \dot{x} &= u, \\ \dot{Y} &= 0. \end{aligned}$$

The main difficulty with (1.8)–(1.9) is the fact that stabilization of x leads to the right-hand side of (1.9) being 0. Therefore Y cannot be directly steered to zero when $x = 0$. This simple observation implies that to stabilize the system one needs to make Y converge "faster" than x. The feedback law and algorithm we present in sections 4 and 5 satisfy this criterion. The idea is as follows: we switch back and forth between flows which decrease one of the variables x and Y in norm, while at the same time leaving

the other variable as "unchanged" as possible. One of the most remarkable features of our feedback law is what "unchanged" turns out to mean; for not only does the nondecreasing variable remain constant in norm, it turns out to evolve *isospectrally*. For matrix Lie algebras, the meaning of this is clear; for general Lie algebras, the adjoint representation of the variable in question evolves with constant spectrum.

In the case of the $so(n)$ system, we have an additional feature. As will be seen in sections 4 and 6, in those parts of the algorithm where the skew-symmetric matrix Y is decreasing in norm (as x evolves isospectrally), all of the eigenvalues of Y^TY except for the largest one remain constant. The results in section 4 suggest that this *partial isospectrality* of Y probably generalizes to other Lie algebras.

We conclude this introduction with an outline of what follows. In section 2, we consider the Heisenberg system (1.1)–(1.3). We introduce our control law for this special case and we show that it stabilizes the system. As might be expected, the feedback is a prototype for the more general ones to follow.

In section 3 we introduce the general Lie algebraic setting. We give the specific hypotheses on the Lie algebra $\mathfrak{g}$ and we show that in a certain sense, these hypotheses cannot be weakened in order to have a stabilization problem that can be solved. We prove certain useful operator identities in $\mathfrak{g}$, and we also verify two crucial inequalities which play a role in the proof of the convergence of the stabilization algorithm. The proof of one of the inequalities, (3.12), is rather delicate and relies on the structure of the Lie algebra $\mathfrak{g}$ in an essential way. This turns out to be somewhat obscured in the setting of specific Lie algebras, and provides another justification for considering the general system (1.8)–(1.9) instead of just, say, the $so(n)$ system.

In section 4 we present our discontinuous controls and we analyze the various cases between which we will be switching in our algorithm. As mentioned above, we show that in each of the main cases, one variable decreases in norm while the other evolves isospectrally. We also verify (4.22), which, in $so(n)$, is the key to the previously mentioned partial isospectrality of Y in the case when it is the norm-decreasing variable.

In section 5 we present the stabilization algorithm and we give a rigorous proof of its convergence. Again, the estimates are rather delicate and are most easily seen in the general setting, not in the setting of specific Lie algebras. It is here, as well as in section 3, where it will be shown that the generalization to the Lie algebraic setting is not as straightforward as it seems.

Finally, in section 6 we apply our results to the $so(n)$ system. We show that stabilization will be achieved in, at most, $\lfloor \frac{n}{2} \rfloor$ iterations of our algorithm. We show some numerics in $so(3)$ to illustrate our results.

2. Stabilization of the Heisenberg system. To illustrate part of our full algorithm for stabilization of the system (1.8)–(1.9), we discuss it in the context of the Heisenberg system or nonholonomic integrator (1.1)–(1.3), which we repeat here for convenience:

$$\dot{x} = u, \tag{2.1}$$

$$\dot{y} = v, \tag{2.2}$$

$$\dot{z} = xv - yu. \tag{2.3}$$

The usefulness of stabilization for this system may be illustrated by an application to the kinematics of a knife edge in point contact with a plane surface or the motion of a rolling wheel—the simplest form of "mobile robot" (see, for example, Bloch and

Drakunov [5], Bloch, Reyhanoglu, and McClamroch [10], and Murray and Sastry [27]). One may transform these systems to one of Heisenberg type.

Our model for the control of (2.1)–(2.3) is the following (slightly incomplete) control law:

$$u = -\alpha x + \beta\, z\, y, \tag{2.4}$$

$$v = -\alpha y - \beta\, z\, x, \tag{2.5}$$

where α and β are positive constants. With this choice, the system (2.1)–(2.3) becomes

$$\dot{x} = -\alpha x + \beta z y, \tag{2.6}$$

$$\dot{y} = -\alpha y - \beta z x, \tag{2.7}$$

$$\dot{z} = -\beta z(x^2 + y^2). \tag{2.8}$$

Let $V = x^2 + y^2$. Then

$$\dot{V} = 2(x\dot{x} + y\dot{y}) = 2(\beta x z y - \beta y z x - \alpha x^2 - \alpha y^2) = -2\alpha V. \tag{2.9}$$

There are a number of strategies for choosing α and β to stabilize the system. It is clear from (2.8) that if we initially choose $\alpha = 0$ and $\beta > 0$, then for x or y not equal to zero, z will be driven asymptotically to 0, while (2.9) shows that V will remain fixed. On the other hand for $\alpha > 0$ and $\beta = 0$, V will be driven to 0. All stabilization strategies to be discussed later are generalizations of this simple observation.

Note also from (2.8) the following: for $\alpha = 0$ the rate of convergence for the variable z depends on the initial conditions for V. The greater the value of $V|_{t=0}$, the faster will be the convergence. On the other hand, if $V|_{t=0} = 0$, i.e., if the initial state belongs to the z-axis, then the variable z will not converge at all. In order to initialize motion in this case, one may apply any nonzero control for a short period of time. It may be a constant nonzero vector or any other suitable control. One possibility is to use

$$u = -\alpha(x - x_1),$$
$$v = -\alpha(y - y_1),$$

with some x_1, y_1 such that $x_1^2 + y_1^2 \neq 0$. With this control the state will leave the z-axis. Then the control (2.4), (2.5) can be used with $\alpha = 0$, $\beta > 0$. When the state reaches an ϵ-neighborhood of $z = 0$ (observe that the derivative of z in this mode is just a constant times $-z$), we can switch to $\alpha > 0$, $\beta = 0$; thus stabilizing the state to the origin.

Let us note here that because of the switchings, the above strategy assumes that the control input is a discontinuous function of the state variables. Thus the existence of the corresponding solution of the differential equations (off of the z-axis only) should be understood in the sense of the Filippov definition (see Filippov [18]). In such systems sliding mode behavior (motion along a discontinuity set) is possible, which can be used to stabilize the system (see DeCarlo, Zak, and Drakunov [16] and a generalization of the sliding mode concept in Drakunov and Utkin [17]). In Bloch and Drakunov [4], [5], [7] several methods are given for achieving stability of (2.1)–(2.3) using sliding mode theory.

Similar considerations apply in the general case discussed here, although we will not discuss explicitly sliding mode behavior; only the discrete switching pattern is needed to obtain stability.

3. The general setting. We now consider the general situation. Let $\mathfrak{g}$ be a real semisimple Lie algebra, and let $B : \mathfrak{g} \times \mathfrak{g} \to \mathbb{R}$ be the Killing form of $\mathfrak{g}$. Assume that $\mathfrak{g}$ has a direct sum decomposition $\mathfrak{g} = \mathfrak{h} \oplus \mathfrak{m}$, where $\mathfrak{h}$ is a compactly imbedded subalgebra and the subspace $\mathfrak{m}$ is the orthogonal complement of $\mathfrak{h}$ relative to B. (In the terminology of Helgason [20], $\mathfrak{g}$ is a semisimple orthogonal symmetric Lie algebra. See this reference, especially Chapter V, for subsequent assertions about Lie algebras.) Under these assumptions, the commutation relations $[\mathfrak{h}, \mathfrak{m}] \subseteq \mathfrak{m}$ and $[\mathfrak{m}, \mathfrak{m}] \subseteq \mathfrak{h}$ hold, and the restriction $B|_{\mathfrak{h}\times\mathfrak{h}}$ of the Killing form B to $\mathfrak{h} \times \mathfrak{h}$ is negative definite. In addition, assume that no ideal of $\mathfrak{g}$ is contained in $\mathfrak{h}$. The Jacobi identity together with the commutation relations for $\mathfrak{h}$ and $\mathfrak{m}$ imply that $[\mathfrak{m}, \mathfrak{m}] \oplus \mathfrak{m}$ is an ideal of $\mathfrak{g}$, and thus so is its orthogonal complement relative to B. But this complement is contained in $\mathfrak{h}$, and therefore $[\mathfrak{m}, \mathfrak{m}] = \mathfrak{h}$. In particular, the representation of $\mathfrak{h}$ on $\mathfrak{m}$ is faithful (because the kernel of this representation is B-orthogonal in $\mathfrak{h}$ to $[\mathfrak{m}, \mathfrak{m}]$).

We will consider stabilization of the system (1.8)–(1.9) in $\mathfrak{g}$, which we repeat here for convenience:

$$\dot{x} = u, \tag{3.1}$$

$$\dot{Y} = [u, x], \tag{3.2}$$

where $x, u \in \mathfrak{m}$, $Y \in \mathfrak{h}$. We first show that this system can be analyzed without loss of generality in a more specialized type of Lie algebra.

Under the given hypotheses on $\mathfrak{g}$, there exist B-orthogonal ideals $\mathfrak{g}_+$ and $\mathfrak{g}_-$ with the following properties: (i) $\mathfrak{g} = \mathfrak{g}_+ \oplus \mathfrak{g}_-$ (direct sum), (ii) $\mathfrak{h}_\pm = \mathfrak{g}_\pm \cap \mathfrak{h}$ is compactly imbedded in $\mathfrak{g}_\pm$ and contains no ideal of $\mathfrak{g}_\pm$, and (iii) $\mathfrak{g}_+ = \mathfrak{h}_+ \oplus \mathfrak{m}_+$ is of noncompact type and $\mathfrak{g}_- = \mathfrak{h}_- \oplus \mathfrak{m}_-$ is of compact type, where $\mathfrak{m}_\pm = \mathfrak{g}_\pm \cap \mathfrak{m}$. If we let $x = x_+ + x_-$, $u = u_+ + u_-$, and $Y = Y_+ + Y_-$ denote the corresponding decompositions, then the system (3.1)–(3.2) decomposes into the systems

$$\dot{x}_\pm = u_\pm, \tag{3.3}$$

$$\dot{Y}_\pm = [u_\pm, x_\pm]. \tag{3.4}$$

It follows that to stabilize the system (3.1)–(3.2) in $\mathfrak{g}$, it is enough to stabilize simultaneously the systems (3.3)–(3.4).

We thus assume from now on without loss of generality that $\mathfrak{g}$ is either of noncompact type or of compact type. This implies that the restriction $B|_{\mathfrak{m}\times\mathfrak{m}}$ of B to $\mathfrak{m}$ is positive definite if $\mathfrak{g}$ is of noncompact type and negative definite if $\mathfrak{g}$ is of compact type.

We should stress here that all simple Lie algebras satisfy our assumptions when $\mathfrak{g} = \mathfrak{h} \oplus \mathfrak{m}$ is a Cartan decomposition. Thus the results we obtain in this paper for our general class of Lie algebras apply to many important cases. It is also worth noting that we would gain no generality by weakening our assumptions that $\mathfrak{g}$ is semisimple and that $\mathfrak{h}$ contains no ideal of $\mathfrak{g}$. For example, if we were to assume merely that $\mathfrak{g}$ is effective, that is, that $\mathfrak{h}$ does not contain any elements of the center of $\mathfrak{g}$, then it would follow that $\mathfrak{g}$ would have an additional ideal $\mathfrak{g}_0 = \mathfrak{h}_0 + \mathfrak{m}_0$ of Euclidean type (see Helgason [20]). But in such an ideal, $[u_0, x_0] = 0$, which would imply that Y_0 is constant. Thus stabilization would not be possible. This would still be the case even if we were to keep the semisimplicity, but not the assumption that $\mathfrak{h}$ contains no ideal of $\mathfrak{g}$.

In order to discuss the compact and noncompact cases simultaneously, let

$$\epsilon = \begin{cases} 1 & \text{if } \mathfrak{g} \text{ is of noncompact type,} \\ -1 & \text{if } \mathfrak{g} \text{ is of compact type.} \end{cases} \tag{3.5}$$

We will use the inner product on $\mathfrak{g}$ defined by the Killing form

$$\langle x_1 + Y_1, x_2 + Y_2 \rangle \equiv \epsilon B(x_1, x_2) - B(Y_1, Y_2) \tag{3.6}$$

for $x_1, x_2 \in \mathfrak{m}$, $Y_1, Y_2 \in \mathfrak{h}$. The corresponding norm will be denoted by $\|\cdot\|$.

Now for $x \in \mathfrak{m}$, let

$$M(x) = \epsilon(\text{ad } x)^2|_{\mathfrak{h}}. \tag{3.7}$$

Then $M(x)$ is a nonnegative symmetric operator on $\mathfrak{h}$ relative to the inner product. To see this, note that ad x is B-symmetric if $\mathfrak{g}$ is noncompact and B-skew-symmetric if $\mathfrak{g}$ is compact. In either case, $(\text{ad } x)^2$ is B-symmetric, and thus so is $M(x)$. Note that $(\text{ad } x)^2$ is nonnegative if $\mathfrak{g}$ is noncompact and nonpositive if $\mathfrak{g}$ is compact, and thus introducing ϵ into the definition of $M(x)$ guarantees that it is a nonnegative operator.

Next, for $Y \in \mathfrak{h}$, let

$$N(Y) = -(\text{ad } Y)^2|_{\mathfrak{m}}. \tag{3.8}$$

Then $N(Y)$ is a nonnegative symmetric operator on $\mathfrak{m}$ relative to the inner product. This follows from the fact that ad Y is B-skew-symmetric.

We will make frequent use of two identities relating the operators $M(x)$ and $N(Y)$. First, the Jacobi identity implies

$$[Y, M(x)Y] = \epsilon[x, N(Y)x] \tag{3.9}$$

for all $x \in \mathfrak{m}$, $Y \in \mathfrak{h}$. Second, the invariance of the Killing form implies

$$\langle Y, M(x)Y \rangle = \|[Y, x]\|^2 = \langle x, N(Y)x \rangle \tag{3.10}$$

for all $x \in \mathfrak{m}$, $Y \in \mathfrak{h}$.

We will also require two estimates arising from $M(x)$ and $N(Y)$. First, since $\mathfrak{m}$ and $\mathfrak{h}$ are each invariant under $(\text{ad } x)^2$ for every $x \in \mathfrak{m}$, we have

$$\begin{aligned} \|x\|^2 &= \epsilon \text{tr}((\text{ad } x)^2) = \epsilon \text{tr}((\text{ad } x)^2|_{\mathfrak{m}}) + \epsilon \text{tr}((\text{ad } x)^2|_{\mathfrak{h}}) \\ &= \epsilon \text{tr}((\text{ad } x)^2|_{\mathfrak{m}}) + \text{tr}(M(x)). \end{aligned}$$

This implies the following inequality:

$$\text{tr}(M(x)) \leq \|x\|^2 \tag{3.11}$$

for all $x \in \mathfrak{m}$.

Our second estimate is the following: there exists a constant $0 < \eta < 1$ such that

$$\text{tr}(N(Y)) > \eta \|Y\|^2 \tag{3.12}$$

for all $Y \in \mathfrak{h}$. This fact, which will be crucial for our discussion of the convergence of our algorithm, is probably part of Lie algebra folklore, but the authors are unaware of any reference. We conclude this section with a proof of this result. Since the discussion that follows is somewhat far afield from our stabilization problem, the reader who is more interested in the control-theoretic aspects might wish to skip ahead to section 4.

In addition to the restriction of the Killing form B, the Lie algebra $\mathfrak{h}$ has two other natural invariant forms:

$$B_{\mathfrak{h}}(Y_1, Y_2) = \text{tr}((\text{ad } Y_1)(\text{ad } Y_2)|_{\mathfrak{h}}), \tag{3.13}$$

$$B_{\mathfrak{m}}(Y_1, Y_2) = \text{tr}((\text{ad } Y_1)(\text{ad } Y_2)|_{\mathfrak{m}}) \tag{3.14}$$

for $Y_1, Y_2 \in \mathfrak{h}$. The form $B_\mathfrak{h}$ is just the Killing form of $\mathfrak{h}$ itself, while the form $B_\mathfrak{m}$ is the natural trace form associated with the representation of $\mathfrak{h}$ on $\mathfrak{m}$. All three invariant forms are related by

$$B(Y_1, Y_2) = B_\mathfrak{h}(Y_1, Y_2) + B_\mathfrak{m}(Y_1, Y_2) \tag{3.15}$$

for $Y_1, Y_2 \in \mathfrak{h}$. Also notice that

$$\operatorname{tr}(N(Y)) = -B_\mathfrak{m}(Y, Y) \tag{3.16}$$

for all $Y \in \mathfrak{h}$.

We have already noted that the restriction of B to $\mathfrak{h}$ is negative definite since $\mathfrak{h}$ is compactly imbedded in $\mathfrak{g}$. In addition, $B_\mathfrak{m}$ is negative definite. Indeed, suppose $B_\mathfrak{m}(Z, Z) = 0$ for some $Z \in \mathfrak{h}$. Then the nonpositive symmetric operator $(\operatorname{ad} Z)^2|_\mathfrak{m}$ must be the zero operator (because all of its nonpositive eigenvalues must be 0). Now for all $x \in \mathfrak{m}$, we have $0 = \langle [Z, [Z, x]], x\rangle = -\langle [Z, x], [Z, x]\rangle$, and thus $[Z, x] = 0$. Since $\mathfrak{h}$ acts faithfully on $\mathfrak{m}$, $Z = 0$.

Letting $\mathfrak{z}(\mathfrak{h})$ denote the center of $\mathfrak{h}$, we have

$$\mathfrak{h} = \mathfrak{z}(\mathfrak{h}) + [\mathfrak{h}, \mathfrak{h}], \tag{3.17}$$

a direct sum, which is orthogonal relative to $\langle \cdot, \cdot \rangle$ (that is, relative to B). Indeed, if Z is orthogonal to $[\mathfrak{h}, \mathfrak{h}]$ relative to $\langle \cdot, \cdot \rangle$, then $0 = \langle Z, [Y_1, Y_2]\rangle = -\langle [Y_1, Z], Y_2\rangle$ for all $Y_1, Y_2 \in \mathfrak{h}$. Thus $[Y_1, Z] = 0$ for all $Y_1 \in \mathfrak{h}$, i.e., $Z \in \mathfrak{z}(\mathfrak{h})$. By a similar computation, the subspaces $\mathfrak{z}(\mathfrak{h})$ and $[\mathfrak{h}, \mathfrak{h}]$ are also orthogonal relative to $B_\mathfrak{h}$ and it therefore follows from (3.15) that they are orthogonal relative to $B_\mathfrak{m}$.

The restriction of $B_\mathfrak{h}$ to $[\mathfrak{h}, \mathfrak{h}]$ is negative definite. Indeed, if $B_\mathfrak{h}(Z, Z) = 0$, then the nonpositive symmetric operator $(\operatorname{ad} Z)^2|_\mathfrak{h}$ must be the zero operator. Now for all $Y \in \mathfrak{h}$, we have $0 = \langle [Z, [Z, Y]], Y\rangle = -\langle [Z, Y], [Z, Y]\rangle$, and thus $[Z, Y] = 0$. Therefore $Z \in \mathfrak{z}(\mathfrak{h})$. Since the restriction of $B_\mathfrak{h}$ to $[\mathfrak{h}, \mathfrak{h}]$ is just the Killing form of $[\mathfrak{h}, \mathfrak{h}]$ itself (because $[\mathfrak{h}, \mathfrak{h}]$ is an ideal of $\mathfrak{h}$), it follows that $[\mathfrak{h}, \mathfrak{h}]$ is compact semisimple.

Now since the restrictions of both B and $B_\mathfrak{m}$ to $[\mathfrak{h}, \mathfrak{h}]$ are negative definite and invariant, it follows that each is a positive multiple of $B_\mathfrak{h}$. Thus there exist $c_1, c_2 > 0$ such that

$$B_\mathfrak{m}(Y_1, Y_2) = c_1 B_\mathfrak{h}(Y_1, Y_2) \quad \text{and} \quad B(Y_1, Y_2) = c_2 B_\mathfrak{h}(Y_1, Y_2)$$

for all $Y_1, Y_2 \in [\mathfrak{h}, \mathfrak{h}]$. The relation (3.15) implies $c_2 = 1 + c_1$. Thus let

$$\eta = \frac{c_1}{c_2} = \frac{c_1}{1 + c_1}. \tag{3.18}$$

Then we have the relation

$$B_\mathfrak{m}(Y_1, Y_2) = \eta B(Y_1, Y_2) \tag{3.19}$$

for all $Y_1, Y_2 \in [\mathfrak{h}, \mathfrak{h}]$.

For $Y, Z \in \mathfrak{h}$, write $Y = Y_c + Y_s$ and $Z = Z_c + Z_s$ for the decompositions into $\mathfrak{z}(\mathfrak{h})$ and $[\mathfrak{h}, \mathfrak{h}]$ ("c" = center, "s" = semisimple). Using the orthogonality relative to all three forms as well as (3.15) and (3.19), we compute

$$\begin{aligned} B_\mathfrak{m}(Y, Z) &= B_\mathfrak{m}(Y_c, Z_c) + B_\mathfrak{m}(Y_s, Z_s) \\ &= B(Y_c, Z_c) - B_\mathfrak{h}(Y_c, Z_c) + \eta B(Y_s, Z_s) \\ &= B(Y_c, Z_c) + \eta B(Y_s, Z_s), \end{aligned} \tag{3.20}$$

where we have used the vanishing of $B_{\mathfrak{h}}$ on $\mathfrak{z}(\mathfrak{h})$ to obtain (3.20). Now since $0 < \eta < 1$ and B is *negative* definite, (3.20) and the orthogonality yield

$$\begin{aligned} B_{\mathfrak{m}}(Y,Z) &< \eta B(Y_c,Z_c) + \eta B(Y_s,Z_s) \\ &= \eta B(Y,Z). \end{aligned} \tag{3.21}$$

If we specialize (3.21) to the case $Y = Z$, we obtain

$$B_{\mathfrak{m}}(Y,Y) < \eta B(Y,Y).$$

Multiplying through by -1 (and noting (3.16)) gives (3.12) for all $Y \in \mathfrak{h}$.

4. Controls. We now present our controls for the system (3.1)–(3.2). These are given by

$$\begin{aligned} u &= -\alpha x + \beta[Y,x] - \gamma[Y,[Y,x]] \\ &= -\alpha x + \beta[Y,x] + \gamma N(Y)x, \end{aligned} \tag{4.1}$$

where $\alpha, \beta, \gamma : \mathfrak{g} \to \mathbb{R}$ are functions. We will assume that $\alpha, \gamma \geq 0$ and $\beta\epsilon \leq 0$.

With (4.1) as our choice of u (and using (3.9)), the system (3.1)–(3.2) becomes

$$\dot{x} = -\alpha x + \beta[Y,x] + \gamma N(Y)x, \tag{4.2}$$

$$\begin{aligned} \dot{Y} &= \beta\epsilon M(x)Y + \gamma[N(Y)x,x] \\ &= \beta\epsilon M(x)Y - \gamma\epsilon[Y,M(x)Y]. \end{aligned} \tag{4.3}$$

Using (4.2), we compute

$$\begin{aligned} \frac{d}{dt}\|x\|^2 &= 2\langle x,\dot{x}\rangle \\ &= -2\alpha\langle x,x\rangle + 2\beta\langle x,[Y,x]\rangle + 2\gamma\langle x,N(Y)x\rangle \\ &= -2\alpha\|x\|^2 + 2\gamma\langle x,N(Y)x\rangle. \end{aligned} \tag{4.4}$$

Here the last equality of (4.4) follows because ad Y is B-skew-symmetric. Now let λ_* denote the largest eigenvalue of the symmetric operator $N(Y)$. Then $\langle x,N(Y)x\rangle \leq \lambda_*\|x\|^2$ for all $x \in \mathfrak{m}$, and thus the right-hand side of (4.4) is nonpositive if $\lambda_*\gamma \geq \alpha$. In this case $\|x\|$ is nonincreasing, and we also have that $\|x\|$ is constant if $\alpha = \gamma = 0$.

Using (4.3), we compute

$$\begin{aligned} \frac{d}{dt}\|Y\|^2 &= 2\langle Y,\dot{Y}\rangle \\ &= 2\beta\epsilon\langle Y,M(x)Y\rangle - 2\gamma\epsilon\langle Y,[Y,M(x)Y]\rangle \\ &= 2\beta\epsilon\langle Y,M(x)Y\rangle. \end{aligned} \tag{4.5}$$

Since $\beta\epsilon \leq 0$ and $M(x)$ is a nonnegative operator, it follows that the right-hand side of (4.5) is nonpositive. Thus $\|Y\|$ is nonincreasing in general and is constant if $\beta = 0$.

Our stabilization algorithm will be (necessarily) discontinuous and will require switching of the control (4.1) between the following three cases: (i) $\alpha > 0$, $\beta = \gamma = 0$; (ii) $\alpha = \kappa\lambda_*$, $\gamma = \kappa$, and $\beta = 0$, where, as above, λ_* is the largest eigenvalue of $N(Y)$ and where κ is a positive function; (iii) $\alpha = \gamma = 0$, $\beta\epsilon < 0$. We now discuss the dynamical behavior of the system (4.2)–(4.3) in each of these cases.

Case I. $\alpha > 0$, $\beta = \gamma = 0$.

In this case, the system (4.2)–(4.3) is

$$\dot{x} = -\alpha x, \tag{4.6}$$

$$\dot{Y} = 0. \tag{4.7}$$

The dynamics here are quite clear: x is driven to 0 radially while Y remains fixed. The remarks made in section 2 regarding the Heisenberg system hold here as well; if Y was not already 0 in the first place, implementing the control (4.1) with these parameter values will render the system unstabilizable. Thus this case will only be used if $Y \equiv 0$.

Case II. $\alpha = \kappa\lambda_*$, $\gamma = \kappa$, $\beta = 0$.

As noted above, $\kappa > 0$. In this case, the control (4.1) has the form

$$u = -\kappa\ (\lambda_* x - N(Y)x), \tag{4.8}$$

while the system (4.2)–(4.3) is

$$\dot{x} = -\kappa(\lambda_* x - N(Y)x), \tag{4.9}$$

$$\dot{Y} = -\kappa\epsilon[Y, M(x)Y]. \tag{4.10}$$

In this case, $\|Y\|$ is constant. In addition, (4.10) is a Lax equation in Y. It follows that the spectrum of ad Y is constant. Therefore the spectrum of the operator $N(Y)$ is constant, as are the dimensions of its eigenspaces. In particular, the eigenvalue λ_*, which occurs in (4.9), is constant.

Let $0 \le \lambda_0 < \lambda_1 < \cdots < \lambda_s = \lambda_*$ denote those eigenvalues of $N(Y)$ which are distinct. (Thus $s \le \dim \mathfrak{m} - 1$.) Let $x = x_0 + x_1 + \cdots + x_s$ be the unique decomposition of x into the eigenspaces of $N(Y)$. Then the differential equation (4.9) decouples into the following system of equations in $\mathfrak{m}$:

$$\begin{aligned} \dot{x}_0 &= -\kappa(\lambda_* - \lambda_0)x_0, \\ \dot{x}_1 &= -\kappa(\lambda_* - \lambda_1)x_1, \\ &\vdots \\ \dot{x}_{s-1} &= -\kappa(\lambda_* - \lambda_{s-1})x_{s-1}, \\ \dot{x}_s &= 0. \end{aligned} \tag{4.11}$$

Since $\kappa(\lambda_* - \lambda_j) > 0$ for $j = 0, 1, \ldots, s-1$, it follows that $x_j \to 0$ asymptotically. If we let x_* denote the projection of x onto the λ_*-eigenspace of $N(Y)$, that is, $x_* = x_s$, then noting that $x_* \equiv x_*|_{t=0}$ is constant, we conclude that

$$x \to x_*$$

asymptotically.

Note that (4.9)–(4.10) and (3.9) imply the following:

$$\dot{Y} = -\kappa[x, N(Y)x] = [x, \dot{x}]. \tag{4.12}$$

Since x is converging to a λ_*-eigenvector of $N(Y)$, the right-hand side of (4.9) is converging to 0, and thus $\dot{x}$ is converging to 0. Therefore (4.12) implies that $\dot{Y}$ is converging to 0.

Summarizing this case, we have that Y evolves isospectrally with constant norm and asymptotically vanishing velocity, while x is driven to x_*, its (constant) projection onto the λ_*-eigenspace of $N(Y)$.

Case III. $\alpha = \gamma = 0$, $\beta\epsilon < 0$.
The system (4.2)–(4.3) for this case is

$$\dot{x} = \beta[Y, x], \tag{4.13}$$

$$\dot{Y} = \beta\epsilon M(x)Y. \tag{4.14}$$

In this case, $\|x\|$ is constant. In addition, (4.13) is a Lax equation in x, and thus ad x has constant spectrum. Therefore the spectrum of the operator $M(x)$ is constant, as are the dimensions of its eigenspaces. Let $0 \leq \mu_0 < \mu_1 < \cdots < \mu_r$ denote those eigenvalues of $M(x)$ which are distinct. (Thus $r \leq \dim \mathfrak{h} - 1$.) For $Y \in \mathfrak{h}$, let $Y = Y_0 + \cdots + Y_r$ denote the unique decomposition of Y into the eigenspaces of $M(x)$. Then the differential equation (4.14) decouples into the following system of equations in $\mathfrak{h}$:

$$\begin{aligned} \dot{Y}_0 &= \beta\epsilon\mu_0 Y_0, \\ \dot{Y}_1 &= \beta\epsilon\mu_1 Y_1, \\ &\vdots \\ \dot{Y}_r &= \beta\epsilon\mu_r Y_r. \end{aligned} \tag{4.15}$$

Since $\beta\epsilon\mu_j < 0$ for $j = 1, \ldots, r$, we have that $Y_j \to 0$ asymptotically. If $\mu_0 > 0$, then the same applies to Y_0. Otherwise, if $M(x)$ has $\mu_0 = 0$ as an eigenvalue, then Y_0 remains constant. Thus we have either $Y \to 0$ or $Y \to Y_0$ asymptotically, where $Y_0 \equiv Y_0|_{t=0}$ is constant. In either case, if we let $Y_\#$ denote the projection of Y onto the nullspace of $M(x)$, then noting that $Y_\# \equiv Y_\#|_{t=0}$ is constant, we conclude that

$$Y \to Y_\#$$

asymptotically.

Using system (4.13)–(4.14), we can derive the following equation:

$$\frac{d}{dt}M(x)^n Y = \beta\epsilon[Y, M(x)^n Y] + \beta\epsilon M(x)^{n+1} Y \tag{4.16}$$

for every nonnegative integer n. Indeed, the case $n = 0$ is just (4.14). Using the induction hypothesis, we have for $n > 0$,

$$\begin{aligned} \tfrac{d}{dt}M(x)^n Y = {} & \epsilon[\dot{x}, [x, M(x)^{n-1}Y]] + \epsilon[x, [\dot{x}, M(x)^{n-1}Y]] \\ & + M(x)(\beta\epsilon[Y, M(x)^{n-1}Y] + \beta\epsilon M(x)^n Y). \end{aligned} \tag{4.17}$$

Now

$$[\dot{x}, [x, M(x)^{n-1}Y]] = \beta[[Y, x], [x, M(x)^{n-1}Y]] \tag{4.18}$$

and

$$[x, [\dot{x}, M(x)^{n-1}Y]] = \beta[x, [[Y, x], M(x)^{n-1}Y]], \tag{4.19}$$

while applying the Jacobi identity repeatedly gives

$$\begin{aligned} M(x)[Y, M(x)^{n-1}Y] = {} & [Y, M(x)^n Y] + \epsilon[[x, Y], [x, M(x)^{n-1}Y]] \\ & + \epsilon[x, [[x, Y], M(x)^{n-1}Y]]. \end{aligned} \tag{4.20}$$

Substituting (4.18), (4.19), and (4.20) into (4.17) and simplifying gives (4.16).

Then from (4.16) immediately follows

$$\frac{d}{dt}f(M(x))Y = \beta\epsilon[Y, f(M(x))Y] + \beta\epsilon f(M(x))M(x)Y \tag{4.21}$$

for every real analytic function f. As an interesting special case of this, let $p(\mu)$ be the minimal polynomial of $M(x)$ and assume that $\mu_0 = 0$ is an eigenvalue of $M(x)$ (so that Y does not converge to 0). Then $p(\mu) = \mu q(\mu)$ for some polynomial q. Taking $f = q$ in (4.21) gives

$$\frac{d}{dt}q(M(x))Y = \beta\epsilon[Y, q(M(x))Y]. \tag{4.22}$$

This is a Lax equation, and it follows that the spectrum of $q(M(x))Y$ remains constant. We will see in section 6 what we suggested in the introduction, namely, that in the $so(n)$ case, the constancy of the spectrum of $q(M(x))Y$ implies that Y itself evolves partially isospectrally.

Summarizing this case, we have that x evolves isospectrally with constant norm, Y is driven to $Y_\#$, its (constant) projection onto the nullspace of $M(x)$, and $q(M(x))Y$ evolves with constant spectrum.

As an aside, it is interesting to compare the system of equations in Case III with the double bracket equations discussed in Brockett [13] and Bloch, Brockett, and Ratiu [1], for example. In these papers the isospectral flow $\dot{L} = [L, [L, N]]$ was considered, where L, N live in a compact Lie algebra and N is fixed. This flow is the gradient flow of $\langle L, N\rangle$ on an adjoint orbit of the corresponding Lie group with respect to the so-called normal metric. The second equation of system (4.13)–(4.14) is, on the other hand, of the form $\dot{Y} = \beta\epsilon[X, [X, Y]]$. This equation is *not* isospectral (although it is coupled to the isospectral equation (4.13)), and we have a different function, $\langle Y, Y\rangle$, decreasing along its flow. This, of course, is precisely what we want in the present context. We note also that in this system there is an interesting coupling between two equations involving brackets. This is reminiscent of Bloch and Crouch [3], although there the coupling is between two Lax equations and the overall flow is Hamiltonian.

5. The stabilization algorithm. We now describe our feedback strategy. As before, λ_* denotes the largest eigenvalue of the operator $N(Y)$, x_* denotes the projection of x onto the λ_*-eigenspace of $N(Y)$, and $Y_\#$ denotes the projection of Y onto the nullspace of $M(x)$. Let $\delta > 0$ be a prescribed error tolerance. In an informal pseudocode, our algorithm can be described as follows.

begin

while $\|Y\| \geq \delta$.

1. Let $r = \|x\|$. Implement the control (4.1) with $\alpha = \lambda_*\kappa$, $\gamma = \kappa$, and $\beta = 0$. Then Y evolves isospectrally with constant norm, while x converges to the constant x_*. If $x_* \neq 0$, then go to Step 3.
2. Let z_* denote a fixed λ_*-eigenvector of $N(Y)$ with $\|z_*\| = r\,(1 - 1/\dim\,\mathfrak{m})^{1/2}$. Let $u = -\alpha(x - z_*)$, where $\alpha > 0$. Then x converges to z_*, while Y remains constant.
3. Implement the control (4.1) with $\alpha = \gamma = 0$, $\beta\epsilon < 0$. Then x evolves isospectrally with constant norm, while Y converges to the constant $Y_\#$.

end while

if $\|x\| \geq \delta$, **then**

4. implement the control (4.1) with $\alpha > 0$, $\beta = \gamma = 0$. Then x will converge to 0 radially, while Y remains 0.

end

In Step 1, if α is a constant, then x will converge to x_* in infinite time; if, for example, $\alpha = 1/\|x - x_*\|$, then x will converge in finite time. Similarly, in Step 3, if β is a constant, then Y will converge to $Y_\#$ in infinite time; if, for example, $\beta = 1/\|Y - Y_\#\|$, then Y will converge in finite time. To establish the convergence claim made in Step 2, we simply note that in this case $x(t)$ has the form $x(t) = f(t)z_*$, where $f(t)$ is a scalar-valued function satisfying $\dot{f} = -\alpha(f-1)$, $f(0) = 0$. (For instance, if $\alpha > 0$ is constant, we have $f(t) = 1-e^{-\alpha t}$.) It follows from (3.2) that $\dot{Y} = [u, x] = 0$, so that Y is constant as claimed.

Step 2 is implemented if x converges to 0 in Step 1. One instance where this could happen is if the initial value of x is 0, in which case the first implementation of Step 1 is trivial. More generally, the case where the projection of x onto the λ_*-eigenspace of $N(Y)$ is 0 seems to be the natural higher-dimensional analog of the situation in the Heisenberg system (2.1)–(2.3), where the initial value starts on the z-axis. As in Steps 1 and 3, Step 2 can also be implemented in finite time.

The λ_*-eigenspace of $N(Y)$ will, in general, have dimension greater than 1 (since the nonzero eigenvalues of the B-skew-symmetric operator ad Y come in complex conjugate pairs). Thus there is no unique choice of eigenvector z_* in Step 2. Any lexicographic ordering of the eigenvectors relative to a coordinate basis will suffice as a selection scheme. The rationale behind the particular normalization of z_* will be explained below. (Choices of this type occur naturally in stabilizing nonholonomic systems; see Sontag [30] for comments on this and related robustness issues.)

We will now show that our algorithm successfully stabilizes the system (3.1)–(3.2) by showing that each of $\|x\|$ and $\|Y\|$ can be brought to within the prescribed error tolerance. Note that as soon as the test condition of the while loop fails, that is, as soon as $\|Y\| < \delta$, then the system will be stabilized whether Step 4 needs to be executed or not. Thus we may assume that the initial value of Y satisfies $\|Y\| \geq \delta$ so that the while loop will be executed at least once. If Y ever converges to 0 in Step 3 because $Y_\# = 0$, then the test condition of the while loop will eventually fail. As noted, this is enough to guarantee that the system is stabilizable.

Assume that for every iteration of Step 3, we have $Y_\# \neq 0$. We will show that after finitely many iterations of the while loop, the test condition will fail. In other words, the projection of Y onto the nullspace of $M(x)$ is eventually arbitrarily small in norm. In fact, we will show a stronger result, for when this situation occurs, then it turns out that $\|x\|$ is simultaneously brought to within the error tolerance. Thus as soon as the while loop's test condition fails, the test condition of the if-then statement (Step 4) will also fail, and the system will already be stabilized.

Assume first that Step 3 is about to be executed. Since Step 1 and possibly Step 2 have already been executed, the initial values $x(0) = x_*$ and $Y(0) = Y_*$ satisfy $N(Y_*)x_* = \lambda_* x_*$. As before, let Y_j denote the projection of Y onto the μ_j-eigenspace of $M(x)$. Recall that $Y_\# = Y_0 \equiv Y_0(0)$ throughout Step 3 and that $Y(t) \to Y_\#$ asymptotically. Using the orthogonality of the eigenspaces, we compute

$$\begin{aligned} \|Y_\#\|^2 &= \|Y_*\|^2 - \textstyle\sum_{j=1}^r \|Y_j(0)\|^2 \\ &\leq \|Y_*\|^2 - \frac{1}{\sum_{j=0}^r \mu_j} \textstyle\sum_{j=0}^r \mu_j \|Y_j(0)\|^2. \end{aligned} \tag{5.1}$$

Note that we are using $\mu_0 = 0$. Now using the orthogonality once again, we compute

$$\sum_{j=0}^{r} \mu_j \|Y_j(0)\|^2 = \langle Y_*, \sum_{j=0}^{r} \mu_j Y_j(0) \rangle = \langle Y_*, M(x_*) Y_* \rangle$$

$$= \langle x_*, N(Y_*) x_* \rangle \tag{5.2}$$

$$= \lambda_* \|x_*\|^2. \tag{5.3}$$

Here we have used (3.10) to obtain (5.2). In addition, using (3.11), we have

$$\sum_{j=1}^{r} \mu_j \leq \operatorname{tr}(M(x_*)) \leq \|x_*\|^2. \tag{5.4}$$

Applying (5.3) and (5.4) to (5.1) yields

$$\|Y_\#\|^2 \leq \|Y_*\|^2 - \lambda_*. \tag{5.5}$$

Now using (3.12), we have

$$\lambda_* \geq \frac{1}{\dim \mathfrak{m}} \operatorname{tr}(N(Y_*)) > \frac{\eta}{\dim \mathfrak{m}} \|Y_*\|^2. \tag{5.6}$$

Applying (5.6) to (5.5) gives our final estimate for Step 3

$$\|Y_\#\|^2 < \left(1 - \frac{\eta}{\dim \mathfrak{m}}\right) \|Y_*\|^2. \tag{5.7}$$

Now assume that Step 3 has already been executed and that Step 1 is about to be executed again. Then the initial values $x(0) = x_\#$ and $Y(0) = Y_\#$ in Step 1 satisfy $M(x_\#)Y_\# = 0$. By (3.10), this implies $\langle x_\#, N(Y_\#)x_\# \rangle = 0$. As before, let x_j denote the projection of x into the λ_j-eigenspace of $N(Y)$. Recall that $x_* = x_s \equiv x_s(0)$ throughout Step 1 and that $x(t) \to x_*$ asymptotically. Using the orthogonality of the eigenspaces, we compute

$$\begin{aligned} \|x_*\|^2 &= \|x_\#\|^2 - \textstyle\sum_{j=0}^{s-1} \|x_j(0)\|^2 \\ &\leq \|x_\#\|^2 - \frac{1}{\sum_{j=0}^{s} (\lambda_s - \lambda_j)} \sum_{j=0}^{s} (\lambda_s - \lambda_j) \|x_j(0)\|^2. \end{aligned} \tag{5.8}$$

Using orthogonality again, we compute

$$\sum_{j=0}^{s} (\lambda_s - \lambda_j) \|x_j(0)\|^2 = \lambda_s \|x_\#\|^2 - \langle x_\#, \sum_{j=0}^{s} \lambda_s x_j(0) \rangle$$

$$= \lambda_s \|x_\#\|^2 - \langle x_\#, N(Y_\#) x_\# \rangle$$

$$= \lambda_s \|x_\#\|^2. \tag{5.9}$$

Also

$$\sum_{j=0}^{s} (\lambda_s - \lambda_j) = s\lambda_s - \sum_{j=0}^{s-1} \lambda_j. \tag{5.10}$$

Applying (5.9) and (5.10) to (5.8) gives

$$\|x_*\|^2 \leq \left(1 - \frac{\lambda_s}{s\lambda_s - \sum_{j=0}^{s-1} \lambda_j}\right) \|x_\#\|^2. \tag{5.11}$$

Finally,

$$\frac{\lambda_s}{s\lambda_s - \sum_{j=0}^{s-1}\lambda_j} \geq \frac{1}{s} \geq \frac{1}{\dim\,\mathfrak{m}}, \tag{5.12}$$

and applying (5.12) to (5.11) gives our final estimate for Step 1

$$\|x_*\|^2 \leq \left(1 - \frac{1}{\dim\,\mathfrak{m}}\right)\|x_\#\|^2. \tag{5.13}$$

Now assume Step 2 is executed because $x = 0$ (that is, $x_* = 0$ in Step 1). Rename $x_* = z_*$, where z_* is the chosen λ_*-eigenvector. Then the normalization of z_* described in Step 2 immediately implies (5.13) holds as an equality.

Define two sequences of real numbers as follows: Let a_j and b_j denote, respectively, the initial values of $\|x\|^2$ and $\|Y\|^2$ prior to the $(j+1)$st iteration of the while loop, where $j = 0, 1, \ldots$. Recall that $\|Y\|$ remains constant during Steps 1 and 2 and $\|x\|$ remains constant during Step 3. Our estimates (5.7) and (5.13) imply that the sequences $\{a_j\}$ and $\{b_j\}$ satisfy

$$a_{j+1} \leq \left(1 - \frac{1}{\dim\,\mathfrak{m}}\right)a_j, \tag{5.14}$$

$$b_{j+1} < \left(1 - \frac{\eta}{\dim\,\mathfrak{m}}\right)b_j. \tag{5.15}$$

Since

$$0 < 1 - \frac{1}{\dim\,\mathfrak{m}} < 1 - \frac{\eta}{\dim\,\mathfrak{m}} < 1, \tag{5.16}$$

it follows from (5.14)–(5.15) that the sequences $\{a_j\}$ and $\{b_j\}$ each converge to 0. In particular, it is immediate that each of $\|x\|$ and $\|Y\|$ can be brought to within the prescribed error tolerance $\delta > 0$ in finitely many iterations of the while loop.

In summary, we have proven the following result.

THEOREM 5.1. *The algorithm globally stabilizes the system* (3.1)–(3.2).

We remark while we have used the error tolerance δ above to indicate how the stabilization algorithm works in practice, the formal proof of stability follows from letting δ limit to zero.

6. Example: *so(n).* We consider the $so(n)$ systems (1.4)–(1.5). Let $\mathfrak{g} = so(n+1)$, the Lie algebra of $(n+1)\times(n+1)$ skew-symmetric matrices. As in section 1, we identify the Lie subalgebra

$$\mathfrak{h} \equiv \left\{ \begin{pmatrix} 0 & 0 \\ 0 & Y \end{pmatrix} : Y \in so(n) \right\}$$

with $so(n)$ and the subspace

$$\mathfrak{m} \equiv \left\{ \begin{pmatrix} 0 & -x^T \\ x & 0 \end{pmatrix} : x \in \mathbb{R}^n \right\}$$

with $\mathbb{R}^n$, and the adjoint action of $so(n)$ on $\mathbb{R}^n$ is the standard action. Since $\mathfrak{g}$ is of compact type, $\epsilon = -1$. For $x \in \mathbb{R}^n$, the operator $M(x) : so(n) \to so(n)$ is given by

$$M(x)Y = -[x,[x,Y]] = xx^TY + Yxx^T. \tag{6.1}$$

This satisfies the following minimal polynomial equation:

$$M(x)^2 Y = (x^T x) M(x) Y. \tag{6.2}$$

For $Y \in so(n)$, the operator $N(Y) : \mathbb{R}^n \to \mathbb{R}^n$ is given by

$$N(Y)x = Y^T Y x. \tag{6.3}$$

The control (4.1) in this setting is given by

$$u = -\alpha x + \beta Y x + \gamma Y^T Y x \tag{6.4}$$

for $x \in \mathbb{R}^n$, $Y \in so(n)$, and thus the system (4.2)–(4.3) is

$$\dot{x} = -\alpha x + \beta Y x + \gamma Y^T Y x, \tag{6.5}$$

$$\dot{Y} = -\beta(x x^T Y + Y x x^T) + \gamma(Y^T Y x x^T - x x^T Y^T Y). \tag{6.6}$$

In Case III, with $\alpha = \gamma = 0$, we know from (4.22) that the matrix $M(x)Y - (x^T x)Y$ evolves with constant spectrum. Since $x^T x$ remains constant throughout this case, this means that

$$Y - \frac{1}{x^T x} M(x) Y \tag{6.7}$$

has constant spectrum. Assume now that the system is in its initial configuration for Step 3. Then $Y^T Y x = \lambda_* x$, where λ_* is the largest eigenvalue of $N(Y) = Y^T Y$. Observe that

$$\begin{aligned} M(x)Y &= x x^T Y + Y x x^T = x x^T Y + \tfrac{1}{\lambda_*} Y x x^T Y^T Y \\ &= (x x^T + \tfrac{1}{\lambda_*} Y x x^T Y^T) Y. \end{aligned}$$

We claim the symmetric operator $\frac{1}{x^T x}(x x^T + \frac{1}{\lambda_*} Y x x^T Y^T)$ is actually an orthogonal projector onto the subspace of $\mathbb{R}^n$ spanned by x and Yx. Indeed, using the antisymmetry of Y, we compute

$$\begin{aligned} (x x^T + \tfrac{1}{\lambda_*} Y x x^T Y^T)^2 &= (x^T x) x x^T + \tfrac{1}{\lambda_*^2} Y x x^T Y^T Y x x^T Y^T \\ &= (x^T x)(x x^T + \tfrac{1}{\lambda_*} Y x x^T Y^T). \end{aligned}$$

This establishes our claim. It follows that the operator given by (6.7) is the composition of the orthogonal projector onto the orthogonal complement of the subspace spanned by x and Yx with Y.

In particular, suppose v is an eigenvector of $Y^T Y$ corresponding to a nonzero eigenvalue $\lambda \neq \lambda_*$. Then Yx is also such an eigenvector and $Yv \pm \lambda^{1/2} i v$ are complex eigenvectors of Y itself corresponding to the eigenvalues $\pm \lambda^{1/2} i$, respectively. Both v and Yv are orthogonal to the subspace spanned by x and Yx. The preceding discussion shows that

$$\begin{aligned} (Y - \tfrac{1}{x^T x} M(x) Y) v &= Yv, \\ (Y - \tfrac{1}{x^T x} M(x) Y) Y v &= Y^2 v = -\lambda v. \end{aligned}$$

It follows that

$$\left(Y - \frac{1}{x^T x} M(x)Y\right)(Yv \pm \lambda^{1/2} iv) = \pm\lambda^{1/2} i(Yv \pm \lambda^{1/2} iv),$$

and thus $\pm\lambda^{1/2}i$ are eigenvalues of $Y - \frac{1}{x^T x}M(x)Y$. (Similarly, if v is an eigenvector of Y^TY corresponding to the possible eigenvalue 0, then consideration of Yv shows that 0 is an eigenvalue of $Y - \frac{1}{x^T x}M(x)Y$.) On the other hand, we have

$$(Y - \tfrac{1}{x^T x}M(x)Y)x = 0,$$

$$(Y - \tfrac{1}{x^T x}M(x)Y)Yx = Y^2x + \lambda_* x = 0.$$

We see that the eigenvalues of Y other than $\pm{\lambda_*}^{1/2}i$ are also eigenvalues of $Y - \frac{1}{x^T x}M(x)Y$ with the same multiplicities. It follows that throughout Step 3, the other eigenvalues of Y, and hence Y^TY, will remain constant. (While the preceding discussion seems to assume implicitly that λ_* has multiplicity 1, it is easy to see that the same result applies if there are additional eigenvectors corresponding to λ_*.)

In Step 3, the eigenvalue of Y^TY, whose initial value is λ_*, is the only one that is evolving nontrivially and it will in fact converge to 0. To see this, recall that Y converges to $Y_\#$, the projection of Y onto the nullspace of $M(x)$. It follows that $\frac{1}{x^T x}M(x)Y = \frac{1}{x^T x}(\mu_0 Y_0 + \cdots + \mu_r Y_r)$ will converge to 0, for if $Y_\# = Y_0 \neq 0$, then $\mu_0 = 0$. But as we have just seen, $\pm{\lambda_*}^{1/2}i$ are the only nonzero eigenvalues of $\frac{1}{x^T x}M(x)Y$ at time $t = 0$, and thus they must converge to 0 asymptotically.

These considerations also tell us how many times we can expect the stabilization algorithm to iterate. Indeed, since Y^TY can have at most $\lfloor \frac{n}{2} \rfloor$ distinct positive eigenvalues, stabilization will be achieved in at most $\lfloor \frac{n}{2} \rfloor$ iterations.

Specializing further, let us consider the case $so(3)$. Here Y^TY has only one nonzero eigenvalue, which has multiplicity 2. It follows that after one execution of Step 3, Y will converge to 0. Thus the algorithm will stabilize the system with just one iteration of the while loop.

As a numerical example of this, consider the 6th order system

$$\dot{x} = u, \tag{6.8}$$

$$\dot{Y} = xu^T - ux^T, \tag{6.9}$$

where $x, u \in R^3$, $Y \in so(3)$ with the following initial conditions:

$$x(0) = \begin{pmatrix} 0.2 \\ 1.1 \\ 1.1 \end{pmatrix},$$

$$Y(0) = \begin{pmatrix} 0 & 0.1 & -0.2 \\ -0.1 & 0 & 3.0 \\ 0.2 & -3.0 & 0 \end{pmatrix}.$$

The spectrum of $Y(0)^TY(0)$ is $\{9.05, 9.05, 0\}$.

After we apply Step 1 (with $u = -\lambda_* x + Y^TYx$) over the interval $[0, t_1]$ with $t_1 = 2$ sec, x and Y become, respectively,

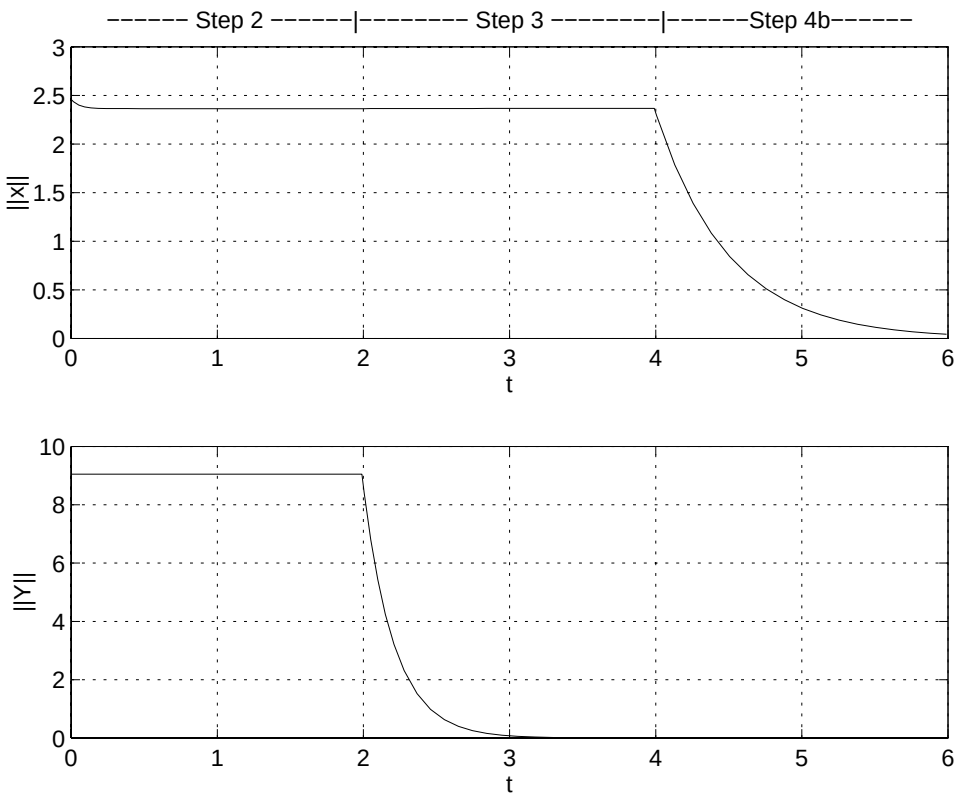


FIG. 6.1. *Numerical example.*

$$x(t_1) = \begin{pmatrix} -0.1076 \\ 1.0966 \\ 1.0726 \end{pmatrix},$$

$$Y(t_1) = \begin{pmatrix} 0 & 0.4359 & 0.1346 \\ -0.4359 & 0 & 2.9738 \\ -0.1346 & -2.9738 & 0 \end{pmatrix}.$$

The spectrum of $Y(t_1)^T Y(t_1)$ remains constant $\{9.05, 9.05, 0\}$ with good accuracy, but as expected, the vector $x(t_1)$ is now the eigenvector corresponding to the eigenvalue $\lambda = 9.05$.

Application of Step 3 (with $u = Yx$) over the interval $[t_1, t_2]$ ($t_2 = 4$ sec) results in the fast decay of $\|Y(t)\|$ to zero as $\|x(t)\|$ remains constant. At the end of this interval, the eigenvalues of $Y(t_2)^T Y(t_2)$ become very small: $\{0.0007, 0.0007, 0\}$.

Finally, Step 4 is executed (with $u = -x$). As expected, x converges to 0 as the value of Y remains unchanged.

Time plots of $\|x(t)\|^2$ and $\|Y(t)\|^2$ are shown in Figure 6.1. The decrease in these magnitudes is clearly seen.

Acknowledgments. We would like to thank Roger Brockett for useful discussions on this material. We would also like to thank the referees for their help in improving the paper.

REFERENCES

[1] A. M. Bloch, R. W. Brockett, and T. S. Ratiu, *Completely integrable gradient flows*, Comm. Math. Phys., 147 (1992), pp. 57–74.

[2] A. M. Bloch and P. E. Crouch, *Nonholonomic control systems on Riemannian manifolds*, SIAM J. Control Optim., 33 (1995), pp. 126–148.

[3] A. M. Bloch and P. E. Crouch, *Optimal control and geodesic flows*, Systems Control Lett., 28 (1996), pp. 65–72.

[4] A. M. Bloch and S. V. Drakunov, *Stabilization of nonholonomic systems via sliding modes*, in Proceedings of the 33rd IEEE Conference on Decision and Control, Orlando, FL, 1994, pp. 2961–2963.

[5] A. M. Bloch and S. V. Drakunov, *Tracking in nonholonomic dynamic systems via sliding modes*, in Proceedings of the 34th IEEE Conference on Decision and Control, New Orleans, LA, 1995, pp. 2103–2106.

[6] A. M. Bloch and S. V. Drakunov, *Stabilization and tracking in the nonholonomic integrator via sliding modes*, Systems Control Lett., 29 (1996), pp. 91–99.

[7] A. M. Bloch and S. V. Drakunov, *Discontinuous stabilization of Brockett's canonical driftless system*, in Proceedings of the IMA, Essays on Mathematical Robotics, J. Baillieul, S. S. Sastry, and H. J. Sussmann, eds., IMA Vol. Math. Appl. 104, Springer-Verlag, New York, 1998, pp. 169–184.

[8] A. M. Bloch, S. V. Drakunov, and M. K. Kinyon, *Stabilization of Brockett's generalized canonical driftless system*, in Proceedings of the 36th IEEE Conference on Decision and Control, San Diego, CA, 1997, pp. 4260–4265.

[9] A. M. Bloch, P. S. Krishnaprasad, J. E. Marsden, and R. M. Murray, *Nonholonomic mechanical systems with symmetry*, Arch. Rational Mech. Anal., 136 (1996), pp. 21–99.

[10] A. M. Bloch, M. Reyhanoglu, and N. H. McClamroch, *Control and stabilization of nonholonomic dynamic systems*, IEEE Trans. Automat. Control, 37 (1992), pp. 1746–1757.

[11] R. W. Brockett, *Control theory and singular Riemannian geometry*, in New Directions in Applied Mathematics, P. Hilton and G. Young, eds., Springer-Verlag, New York, 1982, pp. 11–27.

[12] R. W. Brockett, *Asymptotic stability and feedback stabilization*, in Differential Geometric Control Theory, R. Brockett, R. Millman, and H. Sussmann, eds., Birkhauser, Boston, 1983, pp. 181–191.

[13] R. W. Brockett, *Dynamical systems that sort lists and solve linear programming problems*, Linear Algebra Appl., 146 (1991), pp. 79–91.

[14] R. W. Brockett, *Pattern generation and feedback control of nonholonomic systems*, in Proceedings of the Workshop on Mechanics, Holonomy and Control, San Antonio, IEEE, 1993.

[15] J.-M. Coron, *Global asymptotic stabilization for controllable systems without drift*, Math. Control Signals Systems, 5 (1992), pp. 295–312.

[16] R. DeCarlo, S. Zak, and S. V. Drakunov, *Variable structure and sliding mode control*, in The Control Handbook: A Volume in the Electrical Engineering Handbook Series, CRC Press, Boca Raton, FL, 1996.

[17] S. V. Drakunov and V. I. Utkin, *Sliding mode control in dynamic systems*, Internat J. Control, 55 (1992), pp. 1029–1037.

[18] A. F. Filippov, *Differential Equations with Discontinuous Right-Hand Sides*, Kluwer Academic Publishers, Dordrecht, 1988.

[19] M. Fliess, J. Levine, P. Martin, and P. Rouchon, *Flatness and defect of non-linear systems: Introductory theory and examples*, Internat. J. Control., 61 (1995), pp. 1327–1361.

[20] S. Helgason, *Differential Geometry, Lie Groups, and Symmetric Spaces*, Academic Press, New York, 1978.

[21] J. P. Hespanha, *Stabilization of the non-holonomic integrator via logic-based switchings*, Automatica, 35 (1999), pp. 385–393.

[22] H. Khennouf and C. Canudas de Wit, *On the construction of stabilizing discontinuous controllers for nonholonomic systems*, in Proceedings of IFAC Nonlinear Control Systems Design Symposium, Tahoe City, CA, Pergamon Press, 1995, pp. 747–752

[23] I. Kolmanovsky and N. H. McClamroch, *Hybrid control for global stabilization of nonlinear systems*, in Proceedings of Block Island Workshop on Control using Logic-Based Switching, Block Island, RI, Springer-Verlag, New York, 1997, pp. 128–141.

[24] W. Liu and H. Sussmann, *Limits of highly oscillatory controls and the approximation of general paths by admissible trajectories*, in Proceedings of the 30th IEEE Conference on Decision and Control, Brighton, UK, 1991, pp. 432–437.

[25] R. M'Closkey and R. Murray, *Convergence rates for nonholonomic systems in power form*, in Proceedings of the American Control Conference, San Francisco, IEEE, 1993, pp. 2967–2972.

[26] A. S. Morse, *Control using logic-based switching*, in Trends in Control, A. Isidori, ed., European Control Conference, Springer-Verlag, New York, 1995, p. 74.

[27] R. M. Murray and S. S. Sastry, *Nonholonomic motion planning: Steering using sinusoids*, IEEE Trans. Automat. Control, 38 (1993), pp. 700–716.

[28] J.-B. Pomet, *Explicit design of time-varying stabilizing control laws for a class of controllable systems without drift*, Systems Control Lett., 18 (1992), pp. 147–158.

[29] H. Sira-Ramírez, *On the sliding mode control of differentially flat systems*, Control Theory Adv. Tech., 10 (1995), pp. 1093–1113.

[30] E. D. Sontag, *Stability and stabilization: Discontinuities and the effect of disturbances*, in Proceedings of NATO Advanced Study Institute, Montreal, Canada, Nonlinear Analysis, Differential Equations and Control, F. H. Clarke and R. J. Stern, eds., Kluwer Academic Publishers, Dordrecht, 1999, pp. 551–598.

SIAM J. CONTROL OPTIM.
Vol. 38, No. 3, pp. 875–892

A CLASS OF TEAM PROBLEMS WITH DISCRETE ACTION SPACES: OPTIMALITY CONDITIONS BASED ON MULTIMODULARITY*

PETER R. DE WAAL†‡ AND JAN H. VAN SCHUPPEN†

Abstract. In this paper we discuss a class of team problems with discrete action spaces. We introduce multimodularity into team theory as a natural alternative to convexity in continuous spaces. The main result relates coordinatewise-optimal (cw-optimal) points to the optimal team decision for a class of team problems. The method is based on a characterization of coordinatewise minima of multimodular functions.

Key words. team theory, multimodularity, person-by-person optimality

AMS subject classifications. 93A14, 93E20

PII. S0363012996301452

1. Introduction.

In 1955 Marschak introduced in [6] *team problems* as a mathematical model for cooperative decision making. In a team problem there are two or more decision makers or controllers who receive a common reward as the joint result of their decisions. The fact that the decision makers have a common objective sets it apart from the models that are usually encountered in game theory. Team problems differ from ordinary decision problems with one controller, since the controllers may have different information on which they have to base their decision. The role of information in control problems is discussed in Witsenhausen [11, 12] and Ho and Chu [3]. For some examples and a tutorial introduction to team problems see Ho [4].

The applications of team problems were at first found in the area of decision making in organizations (see Marschak [6]). Recently the attention in team theory has acquired a new impulse from the area of load balancing in distributed computer systems (see, for instance, Wang and Morris [10]). The environment of high-performance computer networking provides a typical example of a complex and highly-distributed system for which decentralized control and team theory appear to provide the right framework.

Despite a history of more than forty years, there are not that many fundamental results in team theory. The verification of the optimality of a team strategy, for instance, is equivalent to a minimization over a function space, and this is infeasible without additional assumptions. To the best of our knowledge, there are only two papers in the literature that present conditions for optimality of team strategies. Radner presents in [7] a sufficient condition that guarantees optimality: if the cost is a convex function of the decision variables and the expected cost is locally finite and stationary for a given team strategy, then this strategy is optimal. Stationarity is defined as a first-order property of the conditional expectation of the cost given the different information patterns. In the case of a convex cost function stationarity of a strategy implies that it is person-by-person optimal (pbpo). This means that the

*Received by the editors April 1, 1996; accepted for publication (in revised form) June 8, 1999; published electronically March 8, 2000.

http://www.siam.org/journals/sicon/38-3/30145.html

†CWI, P.O. Box 94079, 1090 GB Amsterdam, The Netherlands (waal@nlr.nl, schuppen@cwi.nl).

‡Current address: National Aerospace Laboratory NLR, P.O. Box 90502, 1006 BM Amsterdam, The Netherlands (waal@nlr.nl). The research of this author was supported by a grant of NWO (The Netherlands Organization for Scientific Research) through NFI.

expected cost cannot be improved by any player alone if the other team members keep using the same strategy. The importance of this result is twofold. First, it provides a way to verify the optimality of a strategy, and second, it suggests an algorithm to search for the optimal strategy. The local finiteness condition of Radner is relaxed by Krainak, Speyer, and Marcus [5] and replaced with a weaker condition. Both results, however, rely on the fact that the cost function is defined on a continuous space and that it is convex. The continuity makes it possible to compare the expected costs of any two team strategies by effectively constructing a randomization of the strategies. The expected cost of the randomized strategy is then a convex function of the randomization weight and the equivalence of local and global optimality for this one-dimensional convex function ensures the optimality of the stationary strategy.

The primary motivation for our research comes from decentralized control in distributed computer systems. These systems consist of a large number of computers that are interconnected by a network, and they allow sharing of resources and processors. Typically one computer is able to generate processes or tasks that can be performed on the computer itself, or they can be delegated to another processor. In the framework of team theory each computer (or more accurately each process scheduler) is a team player that has to decide for each process that is generated locally where the task has to be performed, locally or on another processor. The action space for such a team problem is intrinsically discrete, and it also does not allow a straightforward extension to a continuous action space. This property prohibits applying the results of [7] and [5].

The aim of this paper is to introduce a framework for team problems with a discrete action space and to present preliminary results for the existence and uniqueness of optima. For the cost function we consider a discrete space analogy for convexity, namely multimodularity. The specific results that were obtained for a class of two-person team problems are as follows:

- we present a characterization of the set of pbpo strategies;
- we give a procedure to check, for any pbpo strategy, in which direction to look for the optimal strategy. Not only does this provide us with an efficient search procedure, but it also enables us to check the optimality of a strategy.

The outline of this paper is as follows. In section 2 we introduce the team problem. A special class of team problems is described in section 3 and we present the optimality conditions for this class. For these conditions we rely on some results on the minima of multimodular functions. These results are summarized in the appendix.

2. Team problems. In this section we introduce our general formulation of the team decision problem. We restrict our attention to a nondynamical team problem.

The following definition of a team problem is based on the definitions of Radner [7] and Krainak, Speyer, and Marcus [5]. We use an underlying probability space $(\Omega, \mathcal{F}, \mathcal{P})$, with Ω the space of elementary events, $\mathcal{F}$ a sigma field of subsets of Ω, and $\mathcal{P}$ a known probability measure on $\mathcal{F}$. We use the letter ω to denote an event, but we do not really distinguish between events and states. In fact we also refer to Ω as the underlying, unobserved, state space, and we also call ω the *state* when it represents an outcome.

A *team* is a set of N decision makers or players. Each player i can choose a decision a_i from a set A_i, the action set. In this paper we assume that the action sets are subsets of $\mathbf{Z}$. Here $\mathbf{Z}$ indicates the set of integers, and $\mathbf{N}$ indicates the set of natural numbers, including 0. If the players choose the action vector $a = (a_1, \ldots, a_N)$ and the state is ω, then a cost $C(a, \omega)$ is incurred. C is a real-valued function that is

measurable with respect to the sigma field generated on the product space $(\mathbf{Z}^m \times \Omega)$ by the Borel sets $\mathcal{B}(\mathbf{Z}^m)$ of $\mathbf{Z}^m$ and by the σ-algebra $\mathcal{F}$. On discrete spaces the σ-algebras are not necessary, but they are retained to simplify the notation and to emphasize the analogy with the case of continuous spaces.

Contrary to a conventional optimal decision problem, we assume that each player has his own observation of the underlying event space. This is implemented as follows. We assume that for each player i there exists an observation space Y_i, a given sigma field $\mathcal{Y}_i$ of subsets of Y_i, and a function $h_i : \Omega \to Y_i$ that is $\mathcal{Y}_i$-measurable. If the event ω occurs, then player i will observe $h_i(\omega)$, and thus each function h_i is a random variable on $(\Omega, \mathcal{F}, \mathcal{P})$. We refer to $\mathcal{Y}_i$ as the *information subfield* of player i, and we define $\mathcal{F}^{h_i} = \{h_i^{-1}(A) \mid A \in \mathcal{Y}_i\}$ as the sigma field that is induced by h_i. The decision that player i makes can depend only on its observation, and thus the set of admissible control laws $\mathcal{U}_i$ for player i is defined by the set of $\mathbf{Z}$-valued functions that are $\mathcal{Y}_i$-measurable. We let $\mathcal{U} = \mathcal{U}_1 \times \cdots \times \mathcal{U}_N$ denote the set of admissible team strategies.

Note that under a strategy γ the team action a is by definition a function of the state ω, i.e., $a = \gamma(h(\omega))$ with $h(\omega) = (h_1(\omega), \ldots, h_N(\omega))$. Under a strategy γ the expected cost of the strategy $J(\gamma)$ is now defined as

$$J(\gamma) = E\{C(\gamma(h(\omega)), \omega)\}, \tag{2.1}$$

where E denotes the expectation with respect to $\mathcal{P}$.

DEFINITION 2.1. *A strategy $\gamma^* \in \mathcal{U}$ is* optimal *if*

$$J(\gamma^*) \leq J(\gamma), \qquad \gamma \in \mathcal{U}. \tag{2.2}$$

The next definition is a variation on the concept of cw-optimality as was introduced in Radner [7]. In that paper a strategy γ is called pbpo if $J(\gamma)$ cannot be improved by changing the strategy for one player alone. The idea that lies behind this definition is that under some extra conditions on the cost function C there exists only one pbpo strategy and this strategy is by the conditions on the cost function then also optimal. As an extra bonus the computation of a pbpo strategy is much easier than for the globally optimal strategy, since the optimization problem in a sense becomes separable. In our model with discrete action spaces we introduce the same concept of pbpo. This is done in a way that is different from [7] and [5], where stationarity is defined by means of the differential of a conditional expectation with respect to the individual decisions. An example of the use of pbpo strategies to determine an optimal solution for a detection problem can be found in [9].

DEFINITION 2.2. *A team strategy $\overline{\gamma} \in \mathcal{U}$ is pbpo if $J(\overline{\gamma}) < \infty$ and for each player i, $i = 1, \ldots, N$,*

$$E\Big[C(\overline{\gamma}(h(\omega)), \omega) \mid \mathcal{F}^{h_i}\Big] \leq E\Big[C(\gamma(h(\omega)), \omega) \mid \mathcal{F}^{h_i}\Big], \qquad (\mathcal{P} - a.s.) \tag{2.3}$$

for all team strategies $\gamma \in \mathcal{U}$ with

$$\gamma_j \equiv \overline{\gamma}_j \text{ for all } j = 1, \ldots, N, \text{ with } j \neq i.$$

A team strategy is called strictly pbpo if the $\leq$ sign in (2.3) *is replaced by a strict inequality ($<$).*

Note that the inequality (2.3) is well defined, since both conditional expectations are random variables on the same probability space. In fact, this implies that the

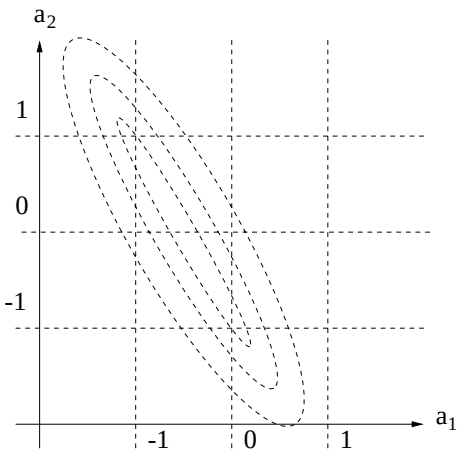


FIG. 2.1. *An example of a continuous convex function restricted to a discrete space.*

inequalities can be replaced by the usual stochastic order (cf. Shaked and Shanthikumar [8, pp. 3, 5]).

If the formulation of the team problem is such that there exists a natural continuation $\overline{C}$ of the cost function from $\mathbf{Z}^N \times \Omega$ to $\mathbf{R}^N \times \Omega$, and $\overline{C}$ is a convex, differentiable and locally finite function, then this continuous version of the problem has a unique optimal solution and this solution is also the only pbpo solution (see Radner [7]). There is no guarantee that restriction to a discrete action space leads to an optimal solution that is in the "neighborhood" of the continuous solution. By neighborhood we mean that the discrete solution is close to the continuous solution, in the usual metric of $\mathbf{R}^N$. Consider for example the cost function

$$C(a_1, a_2, \omega) = (a_1 + 0.5 - \sqrt{3}a_2)^2 + 2(\sqrt{3}(a_1 + 0.5) + a_2)^2 \tag{2.4}$$

for a team problem with only one possible outcome ω, i.e., a deterministic team or an ordinary minimization problem. From the picture of the contour lines as in Figure 2.1, we can see that the continuous minimum is in $(a_1, a_2) = (-0.5, 0)$, while the two integer valued minima are in $(a_1, a_2) = (-1, 1)$ and $(0, -1)$. This example can be modified, however, such that the "discrete" solution is arbitrarily far from the "continuous" solution. Note also that the discrete nature of the problem in this case allows *two* solutions.

In many problems the continuation of C to $\mathbf{R}^N$ may not be as straightforward as in the example. If that is the case, then we might try to construct one. This is where the idea of multimodularity comes in, and it is shown in detail in Hajek [2]. Hajek constructs atoms that span the space $\mathbf{R}^N$. Each atom contains exactly $m+1$ extreme points, and these points lie in $\mathbf{Z}^N$. The continuation $\overline{C}$ of C is piecewise affine on all the atoms. If the function C is multimodular on $\mathbf{Z}^N$, then the continuation of the function is convex in $\mathbf{R}^N$. Unfortunately this continuation is not differentiable, so the results of Radner [7] and Krainak, Speyer, and Marcus [5] cannot be applied here (see also the example in [7, p. 802]). For a justification of the use of multimodular cost functions see the remarks in [2, p. 546] and Bartroli and Stidham [1]. The discussion of multimodularity and its relation to convexity is presented in the appendix.

3. Solution of a class of team problems. In this section we investigate a special class of team problems. It is intended as an example for team problems with

discrete action spaces. It will also serve to indicate the possibilities of using multimodularity properties in solving this kind of team problem. We shall discuss various properties of optimal and cw-optimal strategies. These properties can be used in a procedure to search for the optimal team strategy.

We first need to introduce multimodular functions. For this we consider functions defined on $\mathbf{Z}^m$. We define the vectors $v_0, v_1, \ldots, v_m$ in $\mathbf{Z}^m$ as

$$\begin{aligned} v_0 &= (-1, 0, \ldots, 0), \\ v_1 &= (1, -1, 0, \ldots, 0), \\ v_2 &= (0, 1, -1, 0, \ldots, 0), \\ &\vdots \\ v_{m-1} &= (0, \ldots, 1, -1), \\ v_m &= (0, \ldots, 0, 1), \end{aligned}$$

and we let $\mathcal{V} = \{v_0, v_1, \ldots, v_m\}$. Note that any subset of m vectors of $\mathcal{V}$ is a basis for $\mathbf{Z}^m$, and furthermore we remark that

$$v_0 + v_1 + \cdots + v_m = (0, \ldots, 0). \tag{3.1}$$

DEFINITION 3.1. *A function f on $\mathbf{Z}^m$ for $m \geq 2$ is said to be* multimodular *if for all $z \in \mathbf{Z}^m$,*

$$g(z + v_i) + g(z + v_j) \geq g(z) + g(z + v_i + v_j) \tag{3.2}$$

for any $v_i, v_j \in \mathcal{V}$ and $v_i \neq v_j$.

For a function f on $\mathbf{Z}^m$, $n \in \{1, \ldots, m\}$ and $z \in \mathbf{Z}^m$, we denote the first-order n-difference of f at z as

$$\Delta_n f(z) := f(z + e_n) - f(z), \tag{3.3}$$

where e_n denotes the nth unit vector.

DEFINITION 3.2. *Let f be a real-valued function defined on $\mathbf{Z}^m$. A point $z \in \mathbf{Z}^m$ is called* minimal *for f if $f(z) \leq f(y)$ for all $y \in \mathbf{Z}^m$, $y \neq z$, and it is called* coordinatewise minimal (cw-minimal) *if $f(z) \leq f(z + \lambda e_i)$ for any $i \in \{1, \ldots, m\}$ and any $\lambda \in \mathbf{Z}$, $\lambda \neq 0$. We define a point $z \in \mathbf{Z}^m$ to be* strictly minimal *or* strictly cw-minimal *if these inequalities are replaced by strict inequalities.*

With these definitions we can now introduce the class of team problems that we want to describe. We consider a problem with two players. The underlying event space Ω has three elements, numbered as $\Omega = \{1, 2, 3\}$, and each element occurs with the same probability. We assume that $\mathcal{F}$ is the sigma algebra generated by $\{\{1\}, \{2\}, \{3\}\}$. The action sets for both players are $\mathbf{Z}$.

We assume that the two players have distinctly different information patterns. Player 1 cannot distinguish between events 1 and 2, so $\mathcal{F}^{h_1} = \sigma(\{\{1, 2\}, \{3\}\})$, while player 2 cannot distinguish between events 2 and 3, so $\mathcal{F}^{h_2} = \sigma(\{\{1\}, \{2, 3\}\})$.

For the cost structure we assume the following.

ASSUMPTION 3.3. *For each possible outcome ω, the cost function $C(u_1, u_2, \omega)$ is multimodular as a function of the decision variables $(u_1, u_2) \in \mathbf{Z}^2$.*

In this section we shall discuss the properties of pbpo strategies. For this we shall make use of the classification of cw-minimal points of multimodular functions on $\mathbf{Z}^2$. The following lemma is a direct consequence of the results of Appendix A, but we specifically state it here for easy reference.

LEMMA 3.4. *If $g : \mathbf{Z}^2 \to \mathbf{R}$ is a multimodular function and $y = (y_1, y_2)$ and $z = (z_1, z_2)$ are two distinct strictly cw-minimal points of g, where $z_1 \geq y_1$, then there is some $B > 0$ such that*

1. *$(z_1, z_2) = (y_1 + B, y_2 - B)$;*
2. *for all b, $0 < b < B$, $(y_1 + b, y_2 - b)$ is also cw-minimal;*
3. *if $g(z) > g(y)$, then the minimum of g cannot be in the set $\{(y_1+b, y_2-b)|b \geq B\}$;*
4. *if both y and z, for some $B \geq 0$, are minimal for g, then $(y_1 + b, y_2 - b)$ is minimal for all $0 \leq b \leq B$;*
5. *if both $g(y_1 + 1, y_1 - 1) \geq g(y_1, y_2)$ and $g(y_1 - 1, y_1 + 1) \geq g(y_1, y_2)$, then (y_1, y_2) is minimal; if the inequalities are strict, then the minimum is also unique.*

Proof. See Appendix A, Lemma A.10. □

Note that (strict) multimodularity of a function does not guarantee that the minimum is unique. It can take the form of a line segment $\{(y_1+z, y_2-z) \mid 0 \leq z \leq B\}$ for some $(y_1, y_2) \in \mathbf{Z}^2$ and $B \geq 0$. There can be only one such segment, however.

We now introduce the notation for a team decision strategy. We already saw in section 2 that the observations of the players can be modelled by functions that are defined on the event space. The observations of player 1 are given by a function $h_1 : \Omega \to \{1, 3\}$ of the state, where

$$h_1(\omega) = \begin{cases} 1 & \text{if } \omega = 1, 2, \\ 3 & \text{if } \omega = 3. \end{cases} \tag{3.4}$$

Similarly, we represent the information pattern of player 2 by a function h_2, which is defined as

$$h_2(\omega) = \begin{cases} 1 & \text{if } \omega = 1, \\ 3 & \text{if } \omega = 2, 3. \end{cases} \tag{3.5}$$

With this definition we can now represent a team decision function as $\gamma = (\gamma_{11}, \gamma_{13}, \gamma_{21}, \gamma_{23}) \in \mathbf{Z}^4$, where γ_{ij} represent the action that γ prescribes when player i gets observation j. Finally, we can now write the expected cost $J(\gamma)$ as a function of the team decision rule γ as

$$J(\gamma_{11}, \gamma_{13}, \gamma_{21}, \gamma_{23}) = \frac{1}{3}C(\gamma_{11}, \gamma_{21}, 1) + \frac{1}{3}C(\gamma_{11}, \gamma_{23}, 2) + \frac{1}{3}C(\gamma_{13}, \gamma_{23}, 3). \tag{3.6}$$

In principle, this makes finding the optimal strategy an optimization problem on $\mathbf{Z}^4$.

The properties that γ has to satisfy for optimality and cw-optimality are summarized in the following lemma.

LEMMA 3.5. *A team decision strategy $\gamma^* = ({\gamma_{11}}^*, {\gamma_{13}}^*, {\gamma_{21}}^*, {\gamma_{23}}^*)$ is minimal if*

$$({\gamma_{11}}^*, {\gamma_{13}}^*, {\gamma_{21}}^*, {\gamma_{23}}^*) = \underset{(\gamma_{11}, \gamma_{13}, \gamma_{21}, \gamma_{23}) \in \mathbf{Z}^4}{\arg\min} J(\gamma_{11}, \gamma_{13}, \gamma_{21}, \gamma_{23}), \tag{3.7}$$

or, in other words, if γ^* is a minimum for J. A strategy $\gamma = (\overline{\gamma}_{11}, \overline{\gamma}_{13}, \overline{\gamma}_{21}, \overline{\gamma}_{23})$ is strictly pbpo if J is strictly cw-minimal in γ, or

$$J(\overline{\gamma}_{11}, \overline{\gamma}_{13}, \overline{\gamma}_{21}, \overline{\gamma}_{23}) < J(u_{11}, \overline{\gamma}_{13}, \overline{\gamma}_{21}, \overline{\gamma}_{23}) \text{ for all } u_{11} \in \mathbf{Z}, u_{11} \neq \overline{\gamma}_{11}, \tag{3.8}$$

$$J(\overline{\gamma}_{11}, \overline{\gamma}_{13}, \overline{\gamma}_{21}, \overline{\gamma}_{23}) < J(\overline{\gamma}_{11}, u_{13}, \overline{\gamma}_{21}, \overline{\gamma}_{23}) \text{ for all } u_{13} \in \mathbf{Z}, u_{13} \neq \overline{\gamma}_{13}, \tag{3.9}$$

$$J(\overline{\gamma}_{11}, \overline{\gamma}_{13}, \overline{\gamma}_{21}, \overline{\gamma}_{23}) < J(\overline{\gamma}_{11}, \overline{\gamma}_{13}, u_{21}, \overline{\gamma}_{23}) \text{ for all } u_{21} \in \mathbf{Z}, u_{21} \neq \overline{\gamma}_{21}, \tag{3.10}$$

$$J(\overline{\gamma}_{11}, \overline{\gamma}_{13}, \overline{\gamma}_{21}, \overline{\gamma}_{23}) < J(\overline{\gamma}_{11}, \overline{\gamma}_{13}, \overline{\gamma}_{21}, u_{23}) \text{ for all } u_{23} \in \mathbf{Z}, u_{23} \neq \overline{\gamma}_{23}. \tag{3.11}$$

Proof. The first statement is immediate from the definition of optimality. The second statement follows from the fact that, by definition, a team strategy $\overline{\gamma} = (\overline{\gamma}_{11}, \overline{\gamma}_{13}, \overline{\gamma}_{21}, \overline{\gamma}_{23})$ is strictly pbpo if it satisfies the following set of equations:

$$C(\overline{\gamma}_{11}, \overline{\gamma}_{21}, 1) + C(\overline{\gamma}_{11}, \overline{\gamma}_{23}, 2) < C(\gamma_{11}, \overline{\gamma}_{21}, 1) + C(\gamma_{11}, \overline{\gamma}_{23}, 2), \tag{3.12}$$

$$C(\overline{\gamma}_{13}, \overline{\gamma}_{23}, 3) < C(\gamma_{13}, \overline{\gamma}_{23}, 3), \tag{3.13}$$

$$C(\overline{\gamma}_{11}, \overline{\gamma}_{21}, 1) < C(\overline{\gamma}_{11}, \gamma_{21}, 1), \tag{3.14}$$

$$C(\overline{\gamma}_{11}, \overline{\gamma}_{23}, 2) + C(\overline{\gamma}_{13}, \overline{\gamma}_{23}, 3) < C(\overline{\gamma}_{11}, \gamma_{23}, 2) + C(\overline{\gamma}_{13}, \gamma_{23}, 3) \tag{3.15}$$

for all $\gamma_{11}, \gamma_{13}, \gamma_{21}, \gamma_{23} \in \mathbf{Z}$. For instance, inequality (3.12) is immediate from the definition of pbpo and the fact that $\mathcal{F}^{h_1} = \sigma(\{\{1,2\},\{3\}\})$. Since $C(\overline{\gamma}_{13}, \overline{\gamma}_{23}, 3)$ is independent of $\overline{\gamma}_{11}$, (3.12) thus implies that $J(\overline{\gamma}_{11}, \overline{\gamma}_{13}, \overline{\gamma}_{21}, \overline{\gamma}_{23}) < J(\gamma_{11}, \overline{\gamma}_{13}, \overline{\gamma}_{21}, \overline{\gamma}_{23})$ for all $\gamma_{11} \in \mathbf{Z}$. In a similar way, one can prove the cw-minimality of J for the other components of $\overline{\gamma}$. □

In the remainder of this section we shall exploit the special nature of multimodular functions to derive properties of optima and cw-optima. We show how one can search for other (coordinatewise) minima starting from a cw-minimum. The main result is as follows.

THEOREM 3.6. *If $\alpha = (\alpha_{11}, \alpha_{13}, \alpha_{21}, \alpha_{23})$ and $\beta = (\beta_{11}, \beta_{13}, \beta_{21}, \beta_{23})$ are both strictly pbpo strategies, then they have to satisfy either $M\beta^T \leq M\alpha^T$ or $M\beta^T \geq M\alpha^T$ for*

$$M = \begin{pmatrix} 0 & 1 & 1 & 1 \\ 1 & 0 & 1 & 1 \\ -1 & -1 & 0 & -1 \\ -1 & -1 & -1 & 0 \end{pmatrix}. \tag{3.16}$$

The vector inequality $\leq$ is to be interpreted componentwise, i.e., $z \leq y$ if and only if $z_i \leq y_i$ for all i.

Proof. For any strategy $z = (z_{11}, z_{13}, z_{21}, z_{23})$, the expected cost $J(z)$ for this strategy is

- multimodular in (z_{11}, z_{21}) if (z_{13}, z_{23}) is fixed,
- multimodular in (z_{11}, z_{23}) if (z_{13}, z_{21}) is fixed,
- multimodular in (z_{13}, z_{21}) if (z_{11}, z_{23}) is fixed,
- multimodular in (z_{13}, z_{23}) if (z_{11}, z_{21}) is fixed.

Now assume that α is strictly pbpo. This implies that $\Delta_{11}J(\alpha) > 0$. Using the fact that J is multimodular in (z_{11}, z_{21}) for (z_{13}, z_{23}) fixed, we get that $\Delta_{11}J(\alpha_{11}+m, z_{13}, \alpha_{21}-m, \alpha_{23}) > 0$ for all $m \geq 0$ and all $z_{13} \in \mathbf{Z}$. Note that $\Delta_{11}J(z)$ is independent of z_{13}. Next use the multimodularity of J in (z_{11}, z_{23}) for (z_{13}, z_{21}) fixed to get $\Delta_{11}J(\alpha_{11} + m+n, z_{13}, \alpha_{21}-m, \alpha_{23}-n) > 0$ for all $m, n \geq 0$ and all z_{13}. Finally, use the fact that for any multimodular function f defined on $\mathbf{Z}^2$, $\Delta_i f(x) \leq \Delta_i f(x+e_j)$ for $x \in \mathbf{Z}^2$ and $i, j = 1, 2$, to prove that $\Delta_{11}J(z) > 0$ for all z in

$$\mathcal{G}_{11+} := \{(\alpha_{11} + m + n + i, z_{13}, \alpha_{21} - m + j, \alpha_{23} - n + k) \mid i, j, k, m, n \geq 0, z_{13} \in \mathbf{Z}\}.$$

Specifically, this means that the points in the interior of $\mathcal{G}_{11+}$ cannot be pbpo. With the interior of $\mathcal{G}_{11+}$ we mean the points z with $z_{11} + z_{21} + z_{23} > \alpha_{11} + \alpha_{21} + \alpha_{23}$. Analogously, we can exploit the fact that $\Delta_{11}J(\alpha_{11} - 1, \alpha_{13}, \alpha_{21}, \alpha_{23}) < 0$ to show that $\Delta_{11}J(z) < 0$ for z in

$$\mathcal{G}_{11-} := \{(\alpha_{11} - 1 - m - n - i, z_{13}, \alpha_{21} + m - j, \alpha_{23} + n - k) \mid i, j, k, m, n \geq 0, z_{13} \in \mathbf{Z}\}.$$

In the same manner, we also find that the fact that α is strictly pbpo implies $\Delta_{13}J(z) > 0$ for all z in

$$\mathcal{G}_{13+} := \{(z_{11}, \alpha_{13} + m + n + i, \alpha_{21} - m + j, \alpha_{23} - n + k) \mid i, j, k, m, n \geq 0, z_{13} \in \mathbf{Z}\},$$

and $\Delta_{13}J(z) < 0$ for z in

$$\mathcal{G}_{13-} := \{(z_{11}, \alpha_{13} - 1 - m - n - i, \alpha_{21} + m - j, \alpha_{23} + n - k) \mid i, j, k, m, n \geq 0, z_{11} \in \mathbf{Z}\}.$$

For the decisions of player 2, we get, in a similar manner, $\Delta_{21}J(z) > 0$ for z in

$$\mathcal{G}_{21+} := \{(\alpha_{11} - m + j, \alpha_{13} - n + k, \alpha_{21} + m + n + i, z_{23},) \mid i, j, k, m, n \geq 0, z_{23} \in \mathbf{Z}\},$$

and $\Delta_{21}J(z) < 0$ for all z in

$$\mathcal{G}_{21-} := \{(\alpha_{11} + m - j, \alpha_{13} + n - k, \alpha_{21} - 1 - m - n - i, z_{23},) \mid i, j, k, m, n \geq 0, z_{23} \in \mathbf{Z}\},$$

and finally $\Delta_{23}J(z) > 0$ for z in

$$\mathcal{G}_{23+} := \{(\alpha_{11} - m + j, \alpha_{13} - n + k, z_{21}, \alpha_{23} + m + n + i,) \mid i, j, k, m, n \geq 0, z_{21} \in \mathbf{Z}\},$$

and $\Delta_{23}J(z) < 0$ for z in

$$\mathcal{G}_{23-} := \{(\alpha_{11} + m - j, \alpha_{13} + n - k, z_{21}, \alpha_{23} - 1 - m - n - i,) \mid i, j, k, m, n \geq 0, z_{21} \in \mathbf{Z}\}.$$

Observe that, by the same reasoning as for $\mathcal{G}_{11+}$, there cannot be other pbpo strategies in the interior of any of the $\mathcal{G}$'s.

Now assume that β is also pbpo and $\beta_{11} > \alpha_{11}$. Since β cannot lie in $\mathcal{G}_{11+}$, this implies that

$$\beta_{11} + \beta_{21} + \beta_{23} \leq \alpha_{11} + \alpha_{21} + \alpha_{23}. \tag{3.17}$$

This proves the second inequality of $M\beta^T \leq M\alpha^T$. Since $\beta_{11} > \alpha_{11}$, it also implies that $\beta_{21} + \beta_{23} \leq \alpha_{21} + \alpha_{23}$. This in itself implies that at least one of the inequalities $\beta_{21} \leq \alpha_{21}$ and $\beta_{23} \leq \alpha_{23}$ hold. We shall prove that, in fact, both inequalities hold. Assume that $\beta_{21} > \alpha_{21}$. Since β is pbpo, it cannot lie in the interior of $\mathcal{G}_{21+}$, and thus $\beta_{11} + \beta_{13} + \beta_{21} < \alpha_{11} + \alpha_{13} + \alpha_{21}$. Because of the assumptions on β_{11} and β_{21}, we conclude that $\beta_{13} < \alpha_{13}$. Now β is also outside the interior of $\mathcal{G}_{13-}$, so $\beta_{13} + \beta_{21} + \beta_{23} \geq \alpha_{13} + \alpha_{21} + \alpha_{23}$. This contradicts $\beta_{13} < \alpha_{13}$ and $\beta_{21} + \beta_{23} \leq \alpha_{21} + \alpha_{23}$. We can thus conclude that both $\beta_{21} \leq \alpha_{21}$ and $\beta_{23} \leq \alpha_{23}$.

Since $\beta_{21} \leq \alpha_{21}$ and β is outside of the interior of $\mathcal{G}_{21-}$, this implies

$$\beta_{11} + \beta_{13} + \beta_{21} \geq \alpha_{11} + \alpha_{13} + \alpha_{21}, \tag{3.18}$$

and this proves the fourth component of the matrix inequality. Analogously to the reasoning above, we conclude from this that $\beta_{13} \geq \alpha_{13}$. This, together with the fact that β is outside of the interior of $\mathcal{G}_{13+}$, implies

$$\beta_{13} + \beta_{21} + \beta_{23} \leq \alpha_{13} + \alpha_{21} + \alpha_{23}, \tag{3.19}$$

the first inequality, and finally $\beta_{23} \leq \alpha_{23}$ leads us to

$$\beta_{11} + \beta_{13} + \beta_{23} \geq \alpha_{11} + \alpha_{13} + \alpha_{23}, \tag{3.20}$$

the third inequality.

If we start with $\beta_{11} < \alpha_{11}$, then this same reasoning gives us $M\beta^T \geq M\alpha^T$. The case where $\beta_{11} = \alpha_{11}$ is treated in the following lemma. □

Note that the proof of Theorem 3.6 does not depend on the particular information patterns of this problem. It is based solely on the additive structure of the expected cost and on the multimodularity of the cost function.

Theorem 3.6 provides us with a characterization of the areas around a known pbpo strategy, where we might find other pbpo strategies. If we want to design a search procedure, then we may want to know in which direction we have to search in the immediate neighborhood of α. By immediate neighborhood of α we refer to those strategies β, with $|\beta_{ij} - \alpha_{ij}| \leq 1$ for all coefficients β_{ij}. The following lemmas provide us with the necessary results.

LEMMA 3.7. *If $\alpha = (\alpha_{11}, \alpha_{13}, \alpha_{21}, \alpha_{23})$ and $\beta = (\beta_{11}, \beta_{13}, \beta_{21}, \beta_{23})$ are both strictly pbpo strategies and $\alpha \neq \beta$, then*

- *if $\alpha_{11} = \beta_{11}$, then $\alpha_{21} = \beta_{21}$,*
- *if $\alpha_{23} = \beta_{23}$, then $\alpha_{13} = \beta_{13}$.*

Proof. Assume that $\alpha_{11} = \beta_{11}$. Define $g(z_{13}, z_{21}, z_{23}) := J(\alpha_{11}, z_{13}, z_{21}, z_{23})$; then

$$g(z_{13}, z_{21}, z_{23}) = \frac{1}{3}\Big[C(\alpha_{11}, z_{21}, 1)\Big] + \frac{1}{3}\Big[C(\alpha_{11}, z_{23}, 2) + C(z_{13}, z_{23}, 3)\Big].$$

It is clear that g is a convex function of z_{21} on $\mathbf{Z}$, and that $\alpha_{21} = \arg\min_{z_{21} \in \mathbf{Z}} g(z_{13}, z_{21}, z_{23})$ is independent of z_{13} and z_{23}. Since both α and β are pbpo, they are cw-minimal for g and thus $\alpha_{21} = \beta_{21}$. The proof for $\alpha_{23} = \beta_{23}$ proceeds analogously. □

Lemma 3.7 does not tell us what happens if $\alpha_{13} = \beta_{13}$ or $\alpha_{21} = \beta_{21}$. It appears that we can construct two strictly pbpo strategies α and β that have some components in common. All these possibilities are summarized in the next lemma.

LEMMA 3.8. *If $\alpha = (\alpha_{11}, \alpha_{13}, \alpha_{21}, \alpha_{23})$ and $\beta = (\beta_{11}, \beta_{13}, \beta_{21}, \beta_{23})$ are two distinct strictly pbpo strategies, then one of the following possibilities holds:*

1. *α and β differ in at least two components,*
2. $(\alpha_{11}, \alpha_{13}) \neq (\beta_{11}, \beta_{13})$,
3. $(\alpha_{21}, \alpha_{23}) \neq (\beta_{21}, \beta_{23})$,
4. $(\alpha_{11}, \alpha_{23}) \neq (\beta_{11}, \beta_{23})$.

Proof. Part 1. Assume that the statement is false and that, for instance, α_{11} and β_{11} are the only two coefficients that are different. This implies $C(\alpha_{11}, \alpha_{21}, 1) + C(\alpha_{11}, \alpha_{23}, 2) < C(\beta_{11}, \alpha_{21}, 1) + C(\beta_{11}, \alpha_{23}, 2) = C(\beta_{11}, \beta_{21}, 1) + C(\beta_{11}, \beta_{23}, 2) < C(\alpha_{11}, \beta_{21}, 1) + C(\alpha_{11}, \beta_{23}, 2)$. This gives a contradiction, since $\alpha_{21} = \beta_{21}$ and $\alpha_{23} = \beta_{23}$.

Parts 2 and 3. Assume $(\alpha_{11}, \alpha_{13}) = (\beta_{11}, \beta_{13})$. Consider the function $g(z_{21}, z_{23})$ defined as $J(\alpha_{11}, \alpha_{13}, z_{21}, z_{23})$, so

$$g(z_{21}, z_{23}) = \frac{1}{3}\Big[C(\alpha_{11}, z_{21}, 1) + C(\alpha_{11}, z_{23}, 2) + C(\alpha_{13}, z_{23}, 3)\Big] =: g_1(z_{21}) + g_2(z_{23}).$$

Since α is a strictly pbpo strategy, $(\alpha_{21}, \alpha_{23})$ must be cw-minimal for g. The function g is multimodular in (z_{21}, z_{23}), and thus g_1 and g_2 are convex functions of z_{21} and z_{23}, respectively. This implies that g has a unique minimum, so $\alpha = \beta$.

Part 4. From Lemma 3.7, we see that $\alpha_{11} = \beta_{11}$ implies $\alpha_{21} = \beta_{21}$ and $\alpha_{23} = \beta_{23}$ implies $\alpha_{13} = \beta_{13}$. □

Now assume that we have found a pbpo strategy α, and we want to check the strategies in the neighborhood of α to see whether they are pbpo. Among the pbpo

strategies we can then check the value of the expected cost J for optimality. The neighborhood of a strategy α is the set

$$\{(\beta_{11}, \beta_{13}, \beta_{21}, \beta_{23}) \mid |\beta_{ij} - \alpha_{ij}| \leq 1, \quad i = 1, 2, \ j = 1, 3\}.$$

Note that there are 80 strategies (excluding α) in this set. If we combine the results of Theorem 3.6 and Lemmas 3.7 and 3.8, the following corollary shows how 68 of these strategies can be eliminated.

COROLLARY 3.9. *If $\alpha = (\alpha_{11}, \alpha_{13}, \alpha_{21}, \alpha_{23})$ is a strictly pbpo strategy, then the set of possible pbpo strategies in the neighborhood of α are the strategies of the form $\beta = \alpha \pm \epsilon$ with $\epsilon \in \{(1, 0, -1, 0), (0, 1, 0, -1), (1, 0, 0, -1), (1, 0, -1, -1), (1, 1, 0, -1), (1, 1, -1, -1)\}$. Outside of this set there cannot be pbpo strategies in the neighborhood of α.*

Proof. We sketch the proof in three steps.

1. Assume that a pbpo strategy $\beta = \alpha + \epsilon$ is of the form $\epsilon = (1, -1, \epsilon_{21}, \epsilon_{23})$ for any $\epsilon_{21}, \epsilon_{23} \in \{-1, 0, 1\}$. We show that β cannot satisfy $M\beta^T \leq M\alpha^T$, since then the third and fourth matrix inequalities imply that $\alpha_{21} \leq \beta_{21}$ and $\alpha_{23} \leq \beta_{23}$. These again imply via the first and last inequalities that $\alpha_{11} \geq \beta_{11}$ and $\alpha_{13} \geq \beta_{13}$, and this gives a contradiction with the assumption on the signs of the first two coefficients of ϵ. The inequality $M\beta^T \geq M\alpha^T$ gives a similar contradiction. Analogously, we can show that ϵ cannot be of the form $\epsilon = (-1, 1, \epsilon_{21}, \epsilon_{23})$ or $\pm(\epsilon_{11}, \epsilon_{13}, 1, -1)$.

2. Assume that a pbpo strategy $\beta = \alpha + \epsilon$ is of the form $\epsilon = (1, \epsilon_{13}, \epsilon_{21}, 1)$ for any ϵ_{13}, $\epsilon_{21} \in \{-1, 0, -1\}$. If β were to satisfy $M\beta^T \leq M\alpha^T$, then this would imply $\epsilon_{21} \leq 2$, while $M\beta^T \geq M\alpha^T$ would imply $\epsilon_{21} \geq 2$. Both contradictions show that the ϵ cannot be of the proposed form. Similarly, we can show that ϵ cannot be of the forms $\pm(1, \epsilon_{13}, 1, \epsilon_{23})$, $\pm(\epsilon_{11}, 1, 1, \epsilon_{23})$, or $\pm(\epsilon_{11}, 1, \epsilon_{21}, 2)$.

3. Combine 1 and 2 to get the possible candidates of ϵ. ∎

Note that within this set there are pairs of strategies that differ in exactly one coefficient, and thus of these pairs only one can be strictly pbpo. Furthermore, for any ϵ in the set of Corollary 3.9, both $\alpha + \epsilon$ and $\alpha - \epsilon$ can be pbpo, but at most one of these two strategies can have an expected cost smaller than $J(\alpha)$ (see Lemma A.6). Finally, if it turns out that $J(\alpha + \epsilon) \geq J(\alpha)$ for some ϵ, then the same lemma ensures that $J(\alpha + k\epsilon) \geq J(\alpha)$ for all $k \in \mathbf{N}$, and thus these points cannot be minimal. In immediate consequence of this is the following, which can be proven analogously to part 5 of Lemma 3.4.

COROLLARY 3.10. *If α is a strictly pbpo strategy and for all ϵ as in Corollary 3.9 we have $J(\alpha+\epsilon) \geq J(\alpha)$ and $J(\alpha-\epsilon) \geq J(\alpha)$, then α is minimal. If all the inequalities are strict, then α is the unique optimal strategy.*

This concludes our exploration of this class of team problems. We have developed a check for the optimality of a team strategy, and we have given the description of a procedure to search for the optimal strategy.

4. Conclusions. In this paper we have discussed team problems with discrete action spaces. Inspired by known results for problems on continuous spaces with convex cost functions, we have introduced multimodularity as a natural abstraction of convexity onto discrete spaces.

In the class of team problems of section 3, we have seen that multimodularity of the cost function translates to properties for the expected cost as a function of the strategy. These properties allow us to check for optimality of a strategy, and they indicate how the complexity of a search for the optimum can be reduced. The example, however, indicates that the complexity is still rather high, and we feel that it must be

possible to reduce it even more by exploiting the multimodularity even further. This is a topic for future research.

If we extend the results of section 3 to a model with a larger observation space for both players, then most of the results of the section remain valid. The proof of Theorem 3.6 relies only on the multimodularity of the cost function and not on the structure of the information patterns. This means that it is a straightforward exercise to extend the results of Theorem 3.6 to a larger observation space. In Lemmas 3.7 and 3.8, the particular structure of the information patterns *is* used, and any extension in this direction has to be done on an ad hoc basis.

Extending the team problem to more than two players is not a trivial task. If we try to mimic the proof of Theorem 3.6 for an example with three players, then even for small observation spaces it is not clear if a matrix inequality of the form $M\alpha^T \leq M\beta^T$ will hold and what the form of M will be.

Appendix A. Multimodular functions and optimality.

In this appendix, we introduce the concept of multimodular functions. Furthermore, we define cw-optimality for this class of functions, and we show its relation to ordinary optimality. We present a classification of cw-optimal points, and we specify a procedure to search for the optimum. For a more elaborate introduction to multimodular functions, we refer to Hajek [2].

We consider functions defined on $\mathbf{Z}^m$. We define the vectors $v_0, v_1, \ldots, v_m$ in $\mathbf{Z}^m$ as

$$\begin{aligned} v_0 &= (-1, 0, \ldots, 0), \\ v_1 &= (1, -1, 0, \ldots, 0), \\ v_2 &= (0, 1, -1, 0, \ldots, 0), \\ &\vdots \\ v_{m-1} &= (0, \ldots, 1, -1), \\ v_m &= (0, \ldots, 0, 1), \end{aligned}$$

and we let $\mathcal{V} = \{v_0, v_1, \ldots, v_m\}$. Note that any subset of m vectors of $\mathcal{V}$ is a basis for $\mathbf{Z}^m$, and furthermore we remark that

$$v_0 + v_1 + \cdots + v_m = (0, \ldots, 0). \tag{A.1}$$

Definition A.1. *A function f on $\mathbf{Z}^m$ for $m \geq 2$ is said to be* multimodular *if for all $z \in \mathbf{Z}^m$,*

$$g(z + v_i) + g(z + v_j) \geq g(z) + g(z + v_i + v_j) \tag{A.2}$$

for any $v_i, v_j \in \mathcal{V}$, and $v_i \neq v_j$.

For a function f on $\mathbf{Z}^m$, $n \in \{1, \ldots, m\}$, and $z \in \mathbf{Z}^m$ we denote the first-order n-difference of f at z as

$$\Delta_n f(z) := f(z + e_n) - f(z), \tag{A.3}$$

where e_n denotes the nth unit vector.

Definition A.2. *Let f be a real-valued function defined on $\mathbf{Z}^m$. A point $z \in \mathbf{Z}^m$ is called* minimal *for f if $f(z) \leq f(y)$ for all $y \in \mathbf{Z}^m$, $y \neq z$, and it is called* coordinatewise minimal (cw-minimal) *if $f(z) \leq f(z + \lambda e_i)$ for any $i \in \{1, \ldots, m\}$ and any $\lambda \in \mathbf{Z}$, $\lambda \neq 0$. We define a point $z \in \mathbf{Z}^m$ to be* strictly minimal *or* strictly cw-minimal *if these inequalities are replaced by strict inequalities.*

Note that of course a minimal point is also cw-minimal. The following lemma gives an indication of the properties of cw-optimal points of a multimodular function.

LEMMA A.3. *Let z^* be a strictly cw-minimal point of a multimodular function f, let z be any point in $\mathbf{Z}^m$, and let the coordinates of $z - z^*$ with respect to the bases $\{v_1, \ldots, v_m\}$ and $\{v_0, \ldots, v_{m-1}\}$ be*

$$z - z^* = k_1 v_1 + k_2 v_2 + \cdots + k_m v_m \tag{A.4}$$

and

$$z - z^* = l_0 v_0 + l_1 v_1 + \cdots + l_{m-1} v_{m-1}, \tag{A.5}$$

respectively.

A. *If $k_i > 0$ for all $i = 1, \ldots, m$, then $0 < \Delta_1 f(z - e_1)$, and thus z is not cw-minimal.*
B. *If $k_i < 0$ for all $i = 1, \ldots, m$, then $\Delta_1 f(z) < 0$, and z is not cw-minimal.*
C. *If $l_i > 0$ for all $i = 0, \ldots, m-1$, then $\Delta_m f(z) < 0$, and z is not cw-minimal.*
D. *If $l_i < 0$ for all $i = 0, \ldots, m-1$, then $0 < \Delta_m f(z - e_1)$, and z is not cw-minimal.*

Proof. For statement A we assume, without loss of generality, that $z^* = 0$, and let $z \in \mathbf{Z}^m$ be $z = k_1 v_1 + \cdots + k_m v_m$. From the definition of multimodularity, we get, by taking $v_i = v_0 = (-1, 0, \ldots, 0)$:

$$f(u - e_1) - f(u) \geq f(u - e_1 + v_j) - f(u + v_j)$$

for all $u \in \mathbf{Z}^m$ and all $v_j \in \mathcal{V}, v_j \neq v_0$. Since u is arbitrary, we can rewrite this as

$$\Delta_1 f(u) \leq \Delta_1 f(u + v_j), \quad u \in \mathbf{Z}^m,\ v_j \in \mathcal{V},\ v_j \neq v_0. \tag{A.6}$$

Note that $-e_1 = v_0 = -v_1 - v_2 - \cdots - v_m$, so

$$z - e_1 = (k_1 - 1)v_1 + (k_2 - 1)v_2 + \cdots + (k_m - 1)v_m,$$

where, by assumption, $k_i - 1 \geq 0$ for all i. By repeated application of (A.6), we thus get

$$\Delta_1 f(z - e_1) \geq \Delta_1 f(0) > 0.$$

From $\Delta_1 f(z - e_1) > 0$ follows that z cannot be a cw-minimal point, and this proves A.

For statement B we note that if $z^* = 0$ is a strictly cw-minimal point, then $\Delta_1 f(-e_1) < 0$, and

$$\begin{aligned} \Delta_1 f(z) &= \Delta_1 f(k_1 v_1 + \cdots + k_m v_m) \\ &= \Delta_1 f(-e_1 + (k_1 + 1)v_1 + \cdots + (k_m + 1)v_m) \\ &\leq \Delta_1 f(-e_1) \\ &< 0, \end{aligned}$$

so z is not cw-minimal.

For statements C and D, note that $v_m = e_m$, so the proof is analogous to cases A and B by showing that (A.6) now becomes

$$\Delta_m f(u) \geq \Delta_m f(u + v_j), \quad u \in \mathbf{Z}^m,\ v_j \in \mathcal{V},\ v_j \neq v_m.$$

This concludes the proof. □

To continue with a classification of cw-optimal points, we introduce the following definition of cones and atoms.

DEFINITION A.4. *For $z \in \mathbf{Z}^m$, define the following polyhedral cones:*

$$C_{0+}(z) = \{u \in \mathbf{Z} \mid u = z + k_1 v_1 + \cdots + k_m v_m,\ k_i \in \mathbf{Z},\ k_i > 0\}, \tag{A.7}$$

$$C_{0-}(z) = \{u \in \mathbf{Z} \mid u = z + k_1 v_1 + \cdots + k_m v_m,\ k_i \in \mathbf{Z},\ k_i < 0\}, \tag{A.8}$$

$$C_{m+}(z) = \{u \in \mathbf{Z} \mid u = z + k_0 v_0 + \cdots + k_{m-1} v_{m-1},\ k_i \in \mathbf{Z},\ k_i > 0\}, \tag{A.9}$$

$$C_{m-}(z) = \{u \in \mathbf{Z} \mid u = z + k_0 v_0 + \cdots + k_{m-1} v_{m-1},\ k_i \in \mathbf{Z},\ k_i < 0\}. \tag{A.10}$$

We let $C(z)$ denote the union

$$C(z) = C_{0+}(z) \cup C_{0-}(z) \cup C_{m+}(z) \cup C_{m-}(z). \tag{A.11}$$

From Lemma A.3 we know that if z is a strictly cw-minimal point, then there are no other cw-minimal points in $C(z)$. This means that if we start from a known cw-minimal point z, then we have to search only the complement of $C(z)$ for other possible cw-minimal points. This complement can be characterized by means of a simplicial decomposition of $\mathbf{R}^m$. We now continue with a brief introduction to this decomposition. For a detailed discussion, we refer to Hajek [2].

DEFINITION A.5. *We let Σ denote the set of permutations of $\{0, \ldots, m\}$. Let $\sigma \in \Sigma$ and $z \in \mathbf{Z}^m$. The set $\{u_0, \ldots, u_m\}$ of extreme points of the atom $S(z, \sigma)$ is defined as follows:*

$$\begin{aligned} u_0 &= z, \\ u_1 &= u_0 + v_{\sigma(1)}, \\ u_2 &= u_1 + v_{\sigma(2)}, \\ &\vdots \\ u_m &= u_{m-1} + v_{\sigma(m)}; \end{aligned}$$

hence $u_0 = u_m + v_{\sigma(0)}$. The atom $S(z, \sigma) \subset \mathbf{R}^m$ is thus the set of convex combinations of $\{u_0, \ldots, u_m\}$: $S(z, \sigma) = \{\sum_{i=0}^m a_i u_i \in \mathbf{Z}^m \mid a_i \in \mathbf{R}_+, \sum_{i=0}^m a_i = 1\}$. We denote $S(z, \sigma) = \langle u_0, \ldots, u_m \rangle$.

Each atom is in fact a simplex, since it contains exactly $m + 1$ extreme points in $\mathbf{Z}^m$. Examples of atoms in two and three dimensions are depicted in Figure A.1. In $\mathbf{R}^2$ the atoms are triangles. In $\mathbf{R}^3$ each atom is bounded by four triangles, and each of the triangles that belong to the same atom share exactly one side.

The atoms allow the following alternative characterization of multimodularity. Every atom S contains exactly $m + 1$ points $\{u_0, \ldots, u_m\}$, so for any function f defined on $\mathbf{Z}^m$, there is a unique affine function $L_S(z)$ that agrees with f on the $m + 1$ extreme points. If f is multimodular, then $L_S(z) \leq f(z)$ for $z \in \mathbf{Z}^m$ (see Hajek [2, Lemma 4.2]). The entire $\mathbf{R}^m$ can be decomposed uniquely into atoms of the form $S(z, \sigma)$, and for a function f defined on $\mathbf{Z}^m$ we can thus uniquely construct a continuous function $\underline{f}$ on $\mathbf{R}^m$ that is piecewise affine on all the atoms $S(z, \sigma)$, $z \in \mathbf{Z}^m$. If f is a multimodular on $\mathbf{Z}^m$, then this $\underline{f}$ is a convex function on $\mathbf{R}^m$.

The next property of multimodular functions will be used a couple of times in this paper, so for this reason we state it here explicitly.

LEMMA A.6. *If f is a multimodular function, then $f(z + ke_i)$ is a convex function of $k \in \mathbf{Z}$ for any $z \in \mathbf{Z}$ and unit vector e_i, $i = 1, \ldots, m$.*

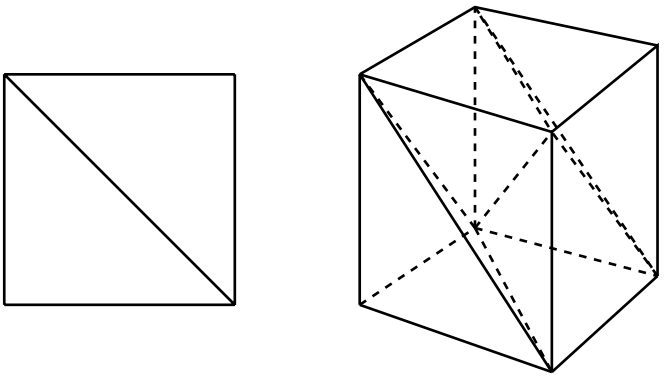


FIG. A.1. *Atoms in two and three dimensions.*

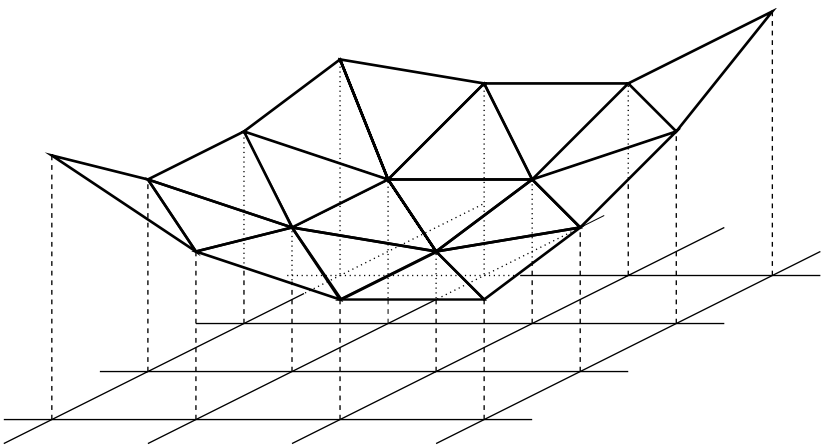


FIG. A.2. *Example of a multimodular function in* $\mathbf{Z}^2$.

Proof. Let $z \in \mathbf{Z}^m$ and e_i be some unit vector. Let S be an atom that contains both z and $z + e_i$. Such an atom exists, since $e_i = v_i + \cdots + v_m$. Using $L_S(z) \leq f(z)$ for this atom, we get

$$\Delta_i f(z) = f(z + e_i) - f(z) = L_S(z + e_i) - L_S(z) = L_S(z + 2e_i) - L_S(z + e_i)$$

$$\leq f(z + 2e_i) - f(z + e_i) = \Delta_i f(z + e_i).$$

The third equality is due to the fact that L_S is affine. □

In a similar manner, we may conclude that in fact for any vector $v \in \mathcal{V}$ and for any $z \in \mathbf{Z}$, the function $f(z + kv)$ is convex in $k \in \mathbf{Z}$.

For an example of a multimodular function on $\mathbf{Z}^2$, see Figure A.2. The atoms that decompose $\mathbf{R}^2$ were depicted in Figure A.1.

DEFINITION A.7. *For an atom* $S(z, \sigma)$, $z \in \mathbf{Z}^m$, $\sigma \in \Sigma$, *we define* $C_\sigma(z)$ *as the polyhedral cone in* $\mathbf{Z}^m$:

$$C_\sigma(z) = \Big\{ u \in \mathbf{Z}^m \mid u = z + k_1(u_1 - u_0) + \cdots + k_m(u_m - u_0), k_i \in \mathbf{N}, S(z, \sigma) = \langle u_0, \ldots, u_m \rangle \Big\}.$$

Define Σ^* *to be the following subset of permutations:*

$$\Sigma^* = \{\sigma \in \Sigma \mid \sigma(0) \neq 0,\ \sigma(0) \neq m,\ \sigma(1) \neq 0,\ \sigma(1) \neq m\},$$

and

$$C_{\Sigma^*}(z) := \bigcup_{\sigma \in \Sigma^*} C_\sigma(z),$$

$$\underline{C}_{\Sigma^*}(z) := \{u \in C_{\Sigma^*}(z) \mid u \neq z + ke_i \text{ for all } k \in \mathbf{Z}, i = 1, \ldots, m\}.$$

Finally, we define $P(z)$ *as the plane through* z *that has normal vector* $(1, \ldots, 1)$, *i.e.,*

$$P(z) := \{u \in \mathbf{Z}^m | u_1 + u_2 + \cdots + u_m = z_1 + z_2 + \cdots + z_m\}.$$

Note in this definition that for the case of $m = 2$, the set Σ^* is empty, and thus $C_{\Sigma^*}(z)$ and $\underline{C}_{\Sigma^*}(z)$ are also empty.

In the remainder of this section, we shall use these definitions to build a characterization of the set of cw-minimal points. The following lemma gives us the necessary preliminary results.

LEMMA A.8. *For any* $z \in \mathbf{Z}$,

$$C(z) \cap C_{\Sigma^*}(z) = \emptyset, \tag{A.12}$$

$$\mathbf{Z}^m = C(z) \cup C_{\Sigma^*}(z) \cup P(z). \tag{A.13}$$

Proof. First consider (A.12). If $m = 2$, then $C_{\Sigma^*}(z) = \emptyset$ and the result is immediate. For $m > 2$, assume that we have $z \in \mathbf{Z}^m$ and $y \in C_\sigma(z)$ for some $\sigma \in \Sigma^*$. This means that we can write $y - z$ as

$$y - z = k_1(u_1 - u_0) + \cdots + k_m(u_m - u_0)$$

or

$$y - z = k_1 v_{\sigma(1)} + k_2(v_{\sigma(1)} + v_{\sigma(2)}) + \cdots + k_m(v_{\sigma(1)} + v_{\sigma(2)} + \cdots + v_{\sigma(m)}), \tag{A.14}$$

and thus

$$\begin{array}{rcl} y - z & = & (k_1 + k_2 + \cdots + k_m)v_{\sigma(1)} \\ & + & (k_2 + k_3 + \cdots + k_m)v_{\sigma(2)} \\ & \vdots & \\ & + & k_m v_{\sigma(m)}. \end{array} \tag{A.15}$$

Note that all $k_i \geq 0$. Since $\sigma \in \Sigma^*$, we know that $\sigma(j) = 0$ for some $j \neq 0,\ 1$. Using (A.1) and subtracting $(k_j + k_{j+1} + \cdots + k_m)(v_0 + \cdots + v_m)$ from (A.15), we get

$$\begin{array}{rcl} y - z & = & (k_1 + \cdots + k_{j-1})v_{\sigma(1)} \\ & + & (k_2 + \cdots + k_{j-1})v_{\sigma(2)} \\ & \vdots & \\ & + & k_{j-1}v_{\sigma(j-1)} \\ & - & k_j v_{\sigma(j+1)} \\ & - & (k_j + k_{j+1})v_{\sigma(j+2)} \\ & \vdots & \\ & - & (k_j + k_{j+1} + \cdots + k_{m-1})v_{\sigma(m)}. \end{array}$$

Recall that $\{v_{\sigma(1)}, \ldots, v_{\sigma(j-1)}, v_{\sigma(j+1)}, \ldots, v_{\sigma(m)}\}$ is a basis for $\mathbf{Z}^m$, so this representation is unique. Since $k_i \geq 0$, we see that y is neither in $C_{0+}(z)$ nor in $C_{0-}(z)$. Analogously, one can prove that $z \notin C_{m+}(z)$ and $z \notin C_{m-}(z)$, and thus $z \notin C(z)$. This proves that $C(z)$ and $C_{\Sigma^*}(z)$ are disjoint.

To prove (A.13), assume that $y \in \mathbf{Z}^m$ and $y \notin C_{\Sigma^*}(z)$. We have to prove that y is in $C(z)$ or $P(z)$. The set $\{S(z,\sigma) \mid z \in \mathbf{Z}^m, \sigma \in \Sigma\}$ forms a partition of $\mathbf{Z}^m$, so there must be a permutation σ of $\{0, \ldots, m\}$, such that $y \in C_\sigma(z)$. Since $y \notin C_{\Sigma^*}(z)$, σ is not in Σ^*, and we must have $\sigma(0) = 0$ or m, or $\sigma(1) = 0$ or m. We shall deal with the case of $\sigma(1) = 0$ first. According to the definition of $C_\sigma(z)$, we can write $y - z$ as

$$y - z = (k_1 + k_2 + \cdots + k_m)v_0 + (k_2 + \cdots + k_m)v_{\sigma(2)} + \cdots + k_m v_{\sigma(m)} \tag{A.16}$$

for some $k_i \geq 0$. We now have to distinguish between the three following cases.

1: $k_1 > 0$. Use (A.1) to show that

$$\begin{aligned} y - z &= -k_1 v_{\sigma(2)} \\ &- (k_1 + k_2)v_{\sigma(3)} \\ &\vdots \\ &- (k_1 + \cdots + k_{m-1})v_{\sigma(m)} \\ &- (k_1 + \cdots + k_m)v_{\sigma(0)}. \end{aligned}$$

Since k_1 is strictly positive, this means that $y \in C_{0-}(z)$.

2: $k_1 = 0$, $\sigma(0) = m$. We first show that $k_m > 0$. Assume that $k_m = 0$. Construct a permutation τ as follows: $\tau(1) = \sigma(2)$, $\tau(2) = \sigma(1) = 0$, $\tau(m) = \sigma(0) = m$, $\tau(0) = \sigma(m)$, and $\tau(i) = \sigma(i)$ for $i = 3, \ldots, m-1$. This makes $\tau \in \Sigma^*$. From (A.16), using $k_1 = k_m = 0$, we get

$$\begin{aligned} y - z &= (k_2 + \cdots + k_m)v_{\tau(1)} \\ &+ (k_2 + \cdots + k_m)v_{\tau(2)} \\ &\vdots \\ &+ k_{m-1}v_{\tau(m-1)}, \end{aligned}$$

which implies that $y \in C_{\Sigma^*}(z)$, and this contradicts $y \notin C_{\Sigma^*}(z)$. We may conclude that the assumption $k_m = 0$ is incorrect, and then it is immediate from (A.16) that $y \in C_{m+}(z)$.

3: $k_1 = 0$ and $\sigma(0) \neq m$. Define the permutation τ by $\tau(1) = \sigma(2)$, $\tau(2) = \sigma(1)$ and $\tau(i) = \sigma(i)$, $i = 3, \ldots, m$. From (A.16), we have

$$\begin{aligned} y - z &= (k_2 + \cdots + k_m)v_{\tau(1)} \\ &+ (k_2 + \cdots + k_m)v_{\tau(2)} \\ &\vdots \\ &+ k_m v_{\tau(m)}, \end{aligned} \tag{A.17}$$

and thus $y \in C_\tau(z)$. Now assume that $\tau(1) \neq m$, then $\tau \in \Sigma^*$, and thus $y \in C_{\Sigma^*}(z)$, which contradicts the assumption that $y \notin C_{\Sigma^*}(z)$. We conclude that $\tau(1) = m$. Since $v_{\tau(1)} = v_m$ and $v_{\tau(2)} = v_0$, we may conclude from (A.17) that the coefficients of $y - z$ sum up to zero, and thus $y \in P(z)$.

The case where $\sigma(1) = m$ is proven analogously, with the roles of v_0 and v_m interchanged. The case where $\sigma(0) = 0$ (or $\sigma(0) = m$) proceeds as follows. Again, use (A.1) by substituting

$$v_{\sigma(1)} + v_{\sigma(2)} + \cdots + v_{\sigma(j)} = -v_{\sigma(j+1)} - \cdots - v_{\sigma(m)} - v_{\sigma(0)} \tag{A.18}$$

into each line of (A.14) to get

$$\begin{aligned} y - z \quad = \quad & -(k_1 + \cdots + k_m)v_{\sigma(0)} \\ & -(k_1 + \cdots + k_{m-1})v_{\sigma(m)} \\ & -(k_1 + \cdots + k_{m-2})v_{\sigma(m-1)} \\ & \vdots \\ & -k_1 v_{\sigma(2)}. \end{aligned} \tag{A.19}$$

This brings $y - z$ in a form similar to (A.15), with the exception that now all the coefficients of the v_is become negative. The proof concludes analogously to the case where $\sigma(1) = 0$. □

Note that the three sets $C(z)$, $C_{\Sigma^*}(z)$, and $P(z)$ are not mutually disjoint. $C(z)$ and $P(z)$ are disjoint, but $C_{\Sigma^*}(z)$ and $P(z)$ have a nonempty intersection. The immediate consequence of Lemma A.8 is summarized in the following theorem, which is the main result of this section.

THEOREM A.9. *If $z \in \mathbf{Z}^m$ is a strictly cw-minimal point of a multimodular function f, then there can be other cw-minimal points only in $\underline{C}_{\Sigma^*}(z)$ or in the plane $P(z)$.*

Proof. It is immediate from Lemma A.8 that the only other cw-minimal points must lie in $C_{\Sigma^*}(z)$ or in $P(z)$. Actually, we do not need to include the entire set $C_{\Sigma^*}(z)$ as a possibility for other coordinatewise minima. It may contain an axis of the form $\{z + ke_i \mid k \in \mathbf{Z}\}$ for some unit vector e_i, $i = 1, \ldots, m$. Since z is strictly cw-minimal and f is convex along this axis (see Lemma A.6), it is immediately obvious that there cannot be other coordinatewise minima on this axis. □

Theorem A.9 not only gives a characterization of the set of cw-minimal points, but it also enables us to search for the minimum in an efficient manner. In the two-dimensional case, for instance, the theorem means that if $z \in \mathbf{Z}^2$ is strictly cw-minimal, then the possible other coordinatewise minima must lie on the line $\{(z_1 + k, z_2 - k) \mid k \in \mathbf{Z}\}$. For an indication of the implications of Theorem A.9, take a look at Figure A.3. We assume that 0 (the center of the cube) is strictly cw-minimal. The points indicated with a bold filled circle are the points on the unit cube that are in both $\underline{C}_{\Sigma^*}(0)$ and in $P(0)$. The bold open circles are the points of $\underline{C}_{\Sigma^*}(0)$ that are not in $P(0)$. The bold open diamonds are the points $P(0)$ that are not in $\underline{C}_{\Sigma^*}(0)$.

We conclude this appendix with the proof of Lemma 3.4. It summarizes the results of Theorem A.9 for multimodular functions defined on $\mathbf{Z}^2$.

LEMMA A.10. *If $g : \mathbf{Z}^2 \to \mathbf{R}$ is a multimodular function and $y = (y_1, y_2)$ and $z = (z_1, z_2)$ are two distinct strictly cw-minimal points of g where $z_1 \geq y_1$, then*

1. *$(z_1, z_2) = (y_1 + B, y_2 - B)$ for some $B > 0$;*
2. *for all b, $0 < b < B$, $(y_1 + b, y_2 - b)$ is also cw-minimal;*
3. *if $g(z) > g(y)$, then the minimum of g cannot be in the set $\{(y_1 + b, y_2 - b) | b \geq B\}$;*
4. *if both y and $z = (y_1 + B, y_2 - B)$ for some $B \geq 0$ are minimal, then $(y_1 + b, y_2 - b)$ is minimal for all $0 \leq b \leq B$;*
5. *if both $g(y_1 + 1, y_1 - 1) \geq g(y_1, y_2)$ and $g(y_1 - 1, y_1 + 1) \geq g(y_1, y_2)$, then (y_1, y_2) is minimal; if the inequalities are strict, then the minimum is also unique.*

Proof. The proof follows immediately from Theorem A.9. Note that if $m = 2$, then there exist no permutations σ of $\{0, 1, 2\}$ with both $\sigma(0) \neq 0, 1$ and $\sigma(1) \neq 0, 1$, so $C_{\Sigma^*}(z) = \emptyset$ for all $z \in \mathbf{Z}^2$. This means that if z is a strictly cw-minimal point, then

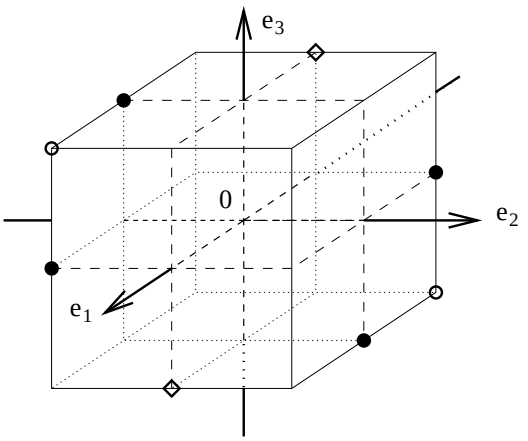


FIG. A.3. *The cone* $C_{\Sigma^*}(0)$ *and the plane* $P(0)$.

the only other cw-minimal points must lie in $P(z) = \{(z_1 + b, z_2 - b) \mid b \in \mathbf{Z}\}$. This proves 1.

To prove 2, note that y strictly cw-minimal implies that $\Delta_1 g(y) > 0$. Since g is multimodular, this implies that $\Delta_1 g(y_1 + b, y_2 - b) > 0$ for $b \geq 0$. In the same manner, $\Delta_1 g(z_1 - 1, z_2) < 0$, and by multimodularity $\Delta_1 g(z_1 - 1 - b, z_2 + b) < 0$ for all $b > 0$. Analogously, one can show that $\Delta_2 g(z_1 - b, z_2 + b) > 0$ and $\Delta_2 g(y_1 + b, y_2 - 1 - b) < 0$ for all $b \geq 0$, and these equalities combined prove 2.

For 3, 4, and 5, note that $f(z) := g(y_1 + z, y_2 - z)$ is a convex function of $z \in \mathbf{Z}$ by the remark below Lemma A.6. □

REFERENCES

[1] M. BARTROLI AND S. STIDHAM, JR., *Multimodular Triangulations*, Tech. report 88-07, Dept. of Oper. Res., University of North Carolina, Chapel Hill, NC, 1988.
[2] B. HAJEK, *Extremal splittings of point processes*, Math. Oper. Res., 10 (1985), pp. 543–556.
[3] Y. HO AND K. CHU, *Information structure in dynamic multi-person control problems*, Automatica J. IFAC, 10 (1974), pp. 341–351.
[4] Y.-C. HO, *Team decision theory and information structures*, Proc. IEEE, 68 (1980), pp. 644–654.
[5] J. KRAINAK, J. SPEYER, AND S. MARCUS, *Static team problems—Part* I: *Sufficient conditions and the exponential cost criterion*, IEEE Trans. Automat. Control, AC-27 (1982), pp. 839–848.
[6] J. MARSCHAK, *Elements for a theory of teams*, Management Sci., 1 (1955), pp. 127–137.
[7] R. RADNER, *Team decision problems*, Ann. Math. Statist., 33 (1962), pp. 857–881.
[8] M. SHAKED AND J. SHANTHIKUMAR, *Stochastic Orders and their Applications*, Academic Press, San Diego, 1994.
[9] D. TENEKETZIS AND P. VARAIYA, *The decentralized quickest detection problem*, IEEE Trans. Automat. Control, AC-29 (1984), pp. 641–644.
[10] Y.-T. WANG AND R. MORRIS, *Load sharing in distributed systems*, IEEE Trans. Comput., C-34 (1985), pp. 204–217.
[11] H.S. WITSENHAUSEN, *On information structures, feedback and causality*, SIAM J. Control, 9 (1971), pp. 149–160.
[12] H. WITSENHAUSEN, *Separation of estimation and control for discrete time systems*, Proc. IEEE, 59 (1971), pp. 1557–1566.

SIAM J. CONTROL OPTIM.
Vol. 38, No. 3, pp. 893–915

APPROXIMATION OF THE KUSHNER EQUATION FOR NONLINEAR FILTERING*

KAZUFUMI ITO† AND BORIS ROZOVSKII‡

Abstract. In this paper we discuss the well-posedness and approximation of solutions to the Kushner equation in nonlinear filtering problems. We develop and analyze a time integration method based on the operator splitting. Also, we discuss its relation to the operator-splitting method for the Zakai equation.

Key words. Kushner equation, operator-splitting method

AMS subject classifications. 93E11, 60H15, 65M10

PII. S0363012998344270

1. Introduction.

In this paper we consider time discretization schemes for the Kushner equation

$$\begin{aligned} d\pi(t) + A\pi(t)\,dt &= (h - \pi_t[h])\pi(t)\,(dy(t) - \pi_t[h]\,dt)\,, \\ \pi(0) &= p_0, \end{aligned} \tag{1.1}$$

where $\pi_t[h] = \int_{R^d} h(x)\pi(t,x)\,dx$ and

$$-A\phi(x) = \frac{\partial}{\partial x_i}\left(a_{i,j}(x)\frac{\partial}{\partial x_j}\phi(x)\right) - \frac{\partial}{\partial x_i}\left(a_i(x)\phi(x)\right). \tag{1.2}$$

The Kushner equation (1.1) arises in nonlinear filtering of diffusion-type processes (see, e.g., [13], [17], [20]). More specifically, assume that the (unobserved) system process $x(t) \in R^d$ is a solution to the Ito equation

$$dx(t) = g(x(t))\,dt + \sigma(x(t))\,dw_1(t), \quad x(0) = x_0, \tag{1.3}$$

and the observation process $y(t) \in R^p$ is of the form

$$y(t) = \int_0^t h(x(s))\,ds + w_2(t), \tag{1.4}$$

where w_1 and w_2 are Brownian motions. In addition, it is assumed that w_2 is independent of $x(t)$ and x_0 is a random variable with density function $p_0 \in L^2(R^d)$. Throughout what follows $g : R^d \to R^d$, $\sigma : R^d \to R^{d\times d}$, and $h : R^d \to R^p$ are bounded continuous functions and g and σ are also Lipschitz continuous.

The diffusion processes $x(t), y(t)$ are considered on a complete probability space $(\Omega, \mathcal{F}, P)$. Let us denote by $\mathcal{F}_t^y$ the P-completed σ-field generated by the observations $\{y(s), 0 \le s \le t\}$.

*Received by the editors September 4, 1998; accepted for publication (in revised form) August 23, 1999; published electronically March 8, 2000.
http://www.siam.org/journals/sicon/38-3/34427.html

†Center for Research in Scientific Computation, North Carolina State University, Raleigh, NC 27695-8205 (kito@eos.ncsu.edu). The research of this author was supported by Office of Naval Research grant N00014-96-1-0265.

‡Center for Applied Mathematical Sciences, University of Southern California, Los Angeles, CA 90089-1113 (rozovski@cams.usc.edu). The research of this author was partially supported by ARO grant DAAG55-98-1-0418 and ONR grant N00014-95-1-0229.

It is a standard fact that for a bounded function ϕ, the best mean square estimator of $\phi(x(t))$ based on the observations $\{y(s), 0 \le s \le t\}$ is given by $\pi[\phi] = E[\phi(x(t))|\mathcal{F}_t^y]$. Moreover, a fundamental result of filtering theory (see, e.g., [7], [17]) says that if $\phi \in C_b^2(R^d)$, then $\pi_t[\phi]$ admits the stochastic differential

$$d\pi_t[\phi] = -\pi_t[A^*\phi]\,dt + (\pi_t[h\phi] - \pi_t[\phi]\pi_t[h])(dy(t) - \pi_t[h]\,dt), \tag{1.5}$$

where A^* is the operator formally adjoint to A with $a_{i,j} = (\frac{1}{2}\sigma\sigma^*)_{i,j}$ and $a_i = g_i - \frac{\partial}{\partial x_j}a_{i,j}$. If the measure $\pi_t(dx) = E[I(x(t) \in dx)|\mathcal{F}_t^y]$ admits a smooth density $\pi(t,x)$ with respect to the Lebesgue measure, it is readily checked that $\pi(t,x)$ verifies the Kushner equation (1.1).

It is also a standard fact (e.g., see [2], [19], [20]) that

$$E\,[\phi(x(t))|\mathcal{F}_t^y] = \frac{\int \phi(x)p(t,x)\,dx}{\int p(t,x)\,dx}, \tag{1.6}$$

where the function $p(t,x)$, usually referred to as the unnormalized filtering density, satisfies the Zakai equation

$$dp(t) + Ap(t)\,dt = hp(t)\,dy(t), \quad p(0) = p_0. \tag{1.7}$$

There is a substantial body of results on numerical approximations of the equations of nonlinear filtering (see, e.g., [1], [2], [3], [4], [5], [6], [8], [9], [11], [14], [16], and the references therein). The overwhelming majority of these numerical schemes deals with the Zakai equation. The rationale for this choice is usually twofold.

1. The Kushner and Zakai equations are equivalent in that their solutions are related by the simple formulas

$$\begin{aligned} &\pi(t,x) = p(t,x)/\textstyle\int p(t,x)\,dx, \\ &p(t,x) = \pi(t,x)\exp\{\textstyle\int_0^t \pi_s[h]\,dy(s) - \frac{1}{2}\int_0^t |\pi_s[h]|^2\,ds\}. \end{aligned} \tag{1.8}$$

2. In contrast to the Kushner equation, the Zakai equation is linear, and this of course greatly simplifies its analysis.

Unfortunately, in spite of its popularity, the Zakai equation has serious deficiencies as a computational tool. These include the following: (a) fast dissipation of the solution as the number of time steps grows, and (b) the effect of intermittency which manifests itself in the appearance of rare but very large peaks. On the contrary, it appears that Kushner's equation of nonlinear filtering is not subject to the aforementioned problems.

In this paper we develop time discretization schemes that apply directly to the Kushner equation (1.1). These schemes belong to the class of splitting-up approximations. The idea of splitting an operator into simpler parts according to different physical properties has been widely used in applied analysis. For example, different alternating directions schemes were developed for the heat equation (see [18] and [10] for a survey of the various operator-splitting schemes). In the case of discrete time observation, Kushner's and Zakai's equations automatically split into two parts: the prediction term that corresponds to the (deterministic) operator A and the correction term related to the (stochastic) operator $B = hp(t)\,\dot{y}(t)$. In [14] (see also [15] and the references therein) Kushner proposed a Markov chain approximation for continuous time equations of nonlinear filtering and proved its weak convergence in the space of measures. Owing to the discrete time nature of the Markov chain approximation, it

is, naturally, of the splitting-up type. Splitting-up time discretization of the Zakai equation in L^2 spaces was pioneered by Bensoussan, Glowinski, and Rascanu [3] and further developed in [8], [11]. In particular, the first estimates of the rate of convergence of the splitting-up approximations for the Zakai equation were obtained in these papers.

In this paper we prove weak and strong convergence in L^2 spaces of the approximations to the solution of (1.1) and estimate the approximation errors.

To avoid the effects of fast dissipation and intermittency, numerical approximations to Zakai's equation are often normalized on every step. It was established experimentally that this procedure is quite efficient. However, the effect of the normalization on the convergence of the approximation has never been studied rigorously. In this paper we derive the rate of convergence of normalized splitting-up approximations to Zakai's equation from the rate of convergence of the splitting-up approximation of Kushner's equation.

2. Well-posedness and difference approximations.

2.1. Zakai and Kushner equations in weighted Sobolev spaces. Weighted Sobolev spaces are of central importance for the approximation results we will be developing below. The analytical theory of Zakai and Kushner equations in standard Sobolev spaces is well understood (see, e.g., [20] and the references therein). However, much less is known about the extensions of this theory to weighted spaces. In this section we will recall some important results in nonlinear filtering and extend them to the case of weighted Sobolev spaces.

The Zakai theory is based on the change of probability measure [2], [20]. Let η_t be a stochastic process defined by

$$\eta_t = \exp\left(-\int_0^t h(x(s))\,dy(s) + \frac{1}{2}\int_0^t |h(x(s))|^2\,ds\right). \tag{2.1}$$

Define a probability measure $\tilde{P}$ on $(\Omega, \mathcal{F})$ by

$$\frac{d\tilde{P}}{dP} = \eta_t. \tag{2.2}$$

In what follows $\tilde{E}$ stands for the expectation with respect to measure $\tilde{P}$. It is a standard fact that under the assumptions we made above the measures P and $\tilde{P}$ are mutually absolutely continuous.

By Girsanov's theorem, the observation process $y(t)$ is an $\mathcal{F}_t^y$ -adapted Wiener process with covariance I on $(\Omega, \mathcal{F}, \tilde{P})$.

Denote $S = (\Omega \times [0,T], \mathcal{F} \times \mathcal{B}, dP \times dt)$ and $\tilde{S} = (\Omega \times [0,T], \mathcal{F} \times \mathcal{B}, d\tilde{P} \times dt)$, where $\mathcal{B}$ is the Borel σ-algebra in $[0,T]$. Set $H = L^2(R^d)$ and $V = H^1(R^d)$ and denote by V^* the strong dual to V. H^* is identified with H so that $V \subset H = H^* \subset V^*$. The inner product $(\cdot,\cdot)$ in H is defined by

$$(\phi, \psi) = \int_{R_d} \phi(x)\psi(x)\,dx.$$

The dual product in $V^* \times V$ is denoted by $\langle \cdot, \cdot \rangle_H$.

Let us denote by X the Hilbert space $L^2(R^d; \varphi(|x|)\,dx)$ equipped with the inner product

$$(\phi, \psi)_X = \int_{R^d} \phi(x)\psi(x)\,\varphi(|x|)\,dx,$$

where the positive weight $\varphi \in H^1_{\ell oc}(R^+)$, which is bounded on every compact in R^+, and so that $\varphi(s) \geq 1$,

$$\varphi'(s) \leq \gamma\, \varphi(s) \ \text{ almost everywhere (a.e.)}, \quad s \in R,$$

and $\varphi^{-1} \in L^1(R^+)$. For example, one can choose $\varphi(s) = e^{\gamma\, s},\ s \geq 0$.

Note that X is continuously embedded into $L^1(R^d)$, i.e.,

$$|\phi|_{L^1} \leq |\varphi|^{\frac{1}{2}}_{L^1} |\phi|_X \quad \text{for } \ \phi \in X.$$

Let $\widetilde{V}$ be the completion of $C^1_0(R^d)$ with respect to the norm

$$|\phi|^2_{\widetilde{V}} = (\nabla\phi, \nabla\phi)_X + |\phi|^2_X.$$

Obviously $\widetilde{V}$ is a Hilbert space. Denote by $\widetilde{V}^*$ its dual with respect to the scalar product in H. It is readily checked that $\widetilde{V} \subset V \subset H$ and that the imbeddings are continuous. The duality between $\widetilde{V}$ and $\widetilde{V}^*$ is denoted by $\langle , \rangle_X$.

Throughout what follows we assume that the initial condition $p_0 \in X$.

DEFINITION 2.1. *A function $p(t)$ is a (generalized) solution of the Zakai equation* (1.7) *if $p(t)$ is a predictable process in V such that $\int_0^T |p(t)|^2_V\, dt < \infty$ (almost surely (a.s.)) and for all $\phi \in C^\infty_0(R^d)$, the equality*

$$\begin{aligned}(p(t), \phi) = (p_0, \phi) &+ \int_0^t \left[\left(-\frac{\partial}{\partial x_i} p(s) a_{i,j}, \frac{\partial}{\partial x_j}\phi\right) + \left(p(s) a_i, \frac{\partial}{\partial x_i}\phi\right)\right] ds \\ &+ \int_0^t (hp(s), \phi)\, dy(s)\end{aligned} \tag{2.3}$$

holds dtdP a.e.

DEFINITION 2.2. *A function $\pi(t)$ is a (generalized) solution of the Kushner equation* (1.1) *if $\pi(t)$ is a predictable process in V such that $\int_0^T |p(t)|^2_V\, dt < \infty$ (a.s.) and for all $\phi \in C^\infty_0(R^d)$, the equality*

$$\begin{aligned}(\pi(t), \phi) = (p_0, \phi) &+ \int_0^t \left[\left(-\frac{\partial}{\partial x_i} \pi(s) a_{i,j}, \frac{\partial}{\partial x_j}\phi\right) + \left(\pi(s) a_i(x), \frac{\partial}{\partial x_i}\phi\right)\right] ds \\ &+ \int_0^t [(\pi(s), h\phi) - (\pi(s), \phi)(\pi(s), h)](dy(s) - (\pi(s), h)\, ds)\end{aligned} \tag{2.4}$$

holds dtdP a.e.

It is well known (see, e.g., [20], [21]) that both the Kushner and Zakai equations have unique generalized solutions; these solutions are nonnegative, $p(t) \in L^1(R^d)$, and formulas (1.6), (1.8) hold. If, as assumed in this paper, the initial condition $p_0 \in X$, the solutions to the Kushner and Zakai equations belong to $\widetilde{V}$. More specifically, the following result holds.

THEOREM 2.3. (i) *The generalized solution to the Zakai equation* (1.7) $p \in L^2(\tilde{S}; \widetilde{V}) \cap L^2(\Omega, \tilde{P}, C(0, T; X))$.

(ii) *The generalized solution to the Kushner equation* (1.1) $\pi \in L^1(S; \widetilde{V}) \cap L^1(\Omega, P, C(0, T; X))$.

Proof. To begin with we will extend the operator A to a linear continuous operator from V to V^* by the formula

$$\langle A\phi, \psi\rangle = (A_0\nabla\phi - a\phi, \nabla\psi) \quad \text{for all} \quad \phi, \psi \in V,$$

where A_0 denotes the symmetric matrix $\{a_{i,j}\}$. We will use the same notation, A, for the extension. Due to the assumptions on the coefficients g, h, and σ, it can be shown that $A \in \mathcal{L}(V, V^*)$, the space of linear continuous operators from V into V^*, and there exist constants $\beta > 0$ and $0 \le \rho < \infty$ such that

$$\langle A\phi, \phi\rangle + \rho|\phi|_H^2 \ge \frac{1}{2}\beta|\phi|_V^2 \quad \text{for all } \phi \in V, \tag{2.5}$$

where $|\phi|_V^2 = (\nabla\phi, \nabla\phi) + |\phi|_H^2$.

The operator A can also be extended to the linear continuous operator from $\widetilde{V}$ to $\widetilde{V}^*$ given by

$$\begin{aligned}\langle A\phi, \psi\rangle_X &= (A\phi, \varphi\psi)\\ &= (A_0\nabla\phi - \phi\, a, \varphi\,\nabla\psi + \psi\,\nabla\varphi) \quad \text{for all} \quad \phi \in \widetilde{V},\ \psi \in C_0^\infty(R^d),\end{aligned} \tag{2.6}$$

where $\varphi(x) = \varphi(|x|)$ and

$$\nabla\varphi = \varphi'(|x|)\,\frac{x}{|x|}.$$

This extension still will be denoted by A.

It is readily checked that there exist constants $\tilde{\rho}, C < \infty$ and $\beta > 0$ so that

$$|Ax|_{\widetilde{V}^*} \le C\,|x|_{\widetilde{V}}^2 \quad \text{and} \quad \langle A\phi, \phi\rangle_X + \tilde{\rho}\,|\phi|_X^2 \ge \frac{\beta}{2}\,|\phi|_{\widetilde{V}}^2. \tag{2.7}$$

Let us consider the operator equation

$$p(t) = p_0 - \int_0^t Ap(s)\,ds + \int_0^t Bp(s)\,dy(s), \tag{2.8}$$

where A is the operator defined by (2.6) and B maps $\widetilde{V}$ into the space of Hilbert–Schmidt operators $\mathcal{L}_2(R^p, X)$ by the formula $Bf = h(x)\,f$. We will consider (2.8) on $(\Omega, \mathcal{F}, \tilde{P})$ and will interpret it as an equation in $\widetilde{V}$. The latter means (see [20]) that a function $p(t)$ is a solution of (2.8) if $p(t)$ is a predictable process in $\widetilde{V}$ such that $\int_0^T |p(t)|_{\widetilde{V}}^2\,dt < \infty$ (a.s.) and for all $v \in C_0^\infty(R^d)$,

$$(p(t), v)_X = (p_0, v)_X - \int_0^t \langle Ap(s), v\rangle_X\,ds + \int_0^t (Bp(s), v)_X\,dy(s). \tag{2.9}$$

Owing to (2.7) it is easy to show that there exist constants $K,\ C \in R_+$, and $\beta > 0$ so that for all $x \in V$,

$$-2\,\langle Ax, x\rangle_X + |Bx|_X^2 \le K\,|x|_X^2 - \frac{\beta}{2}\,|x|_{\widetilde{V}}^2 \ \text{ and } \ |Ax|_{\widetilde{V}^*} \le C|x|_{\widetilde{V}}.$$

Thus (2.8) satisfies the assumptions of Theorem 3.1.4 in [20] and so there exists a unique solution of this equation, $p(t, x)$, and

$$p \in L^2(\tilde{S}, \widetilde{V}) \cap L^2(\Omega, \tilde{P}, C(O, T; X)). \tag{2.10}$$

Let ψ be an arbitrary function from $C_0^\infty(R^d)$. Then $\varphi^{-1}\psi \in \widetilde{V}$. Now taking $v = \varphi^{-1}\psi$, we can rewrite (2.9) in the form

$$(p(t),\psi) = (p_0,\psi) - \int_0^t (A_0\nabla\phi - a\phi, \nabla\psi)\,ds + \int_0^t (hp(s),\psi)\,dy(s). \tag{2.11}$$

Thus we proved that $p(t)$ is a generalized solution of the Zakai equation, and we know that this solution is unique. From this and Proposition 3.1 and Theorem 3.1 in [21] it follows that under our assumptions for any $f \in L^\infty(R^d)$,

$$E\left[f(x(t))|\mathcal{F}_t^y\right] = \int_{R^d} f(x)\pi(t,x)\,dx \ \text{ a.s.}$$

and $\pi(t,x)$ is a unique generalized solution of the Kushner equation (1.1). Moreover, $\pi(t,x) = p(t,x)/\rho_t$, where

$$\rho_t = \exp\left(\int_0^t \pi_s[h]\ dy(s) - \frac{1}{2}\int_0^t |\pi_s[h]|^2\ ds\right) = \langle p(t),1\rangle \ \text{ a.s.} \tag{2.12}$$

Since for any $q \geq 1$, $\rho(t)^q$ is a nonnegative supermartingale with respect to the filtration $(\Omega,\mathcal{F},\{\mathcal{F}_t^y\}_{t\leq T},\tilde{P})$, it follows from (2.10) and (2.12) that $\pi \in L^1(S,\widetilde{V}) \cap L^1(\Omega,P,C(O,T;X))$. Indeed, by the Schwarz inequality,

$$E\int_0^T |\pi(s)|_{\widetilde{V}}\,dt = \tilde{E}\int_0^T \rho_t^{-1}\eta_t^{-1}|p(t)|_{\widetilde{V}}\,dt$$

$$\leq \left(\tilde{E}\int_0^T |p(t)|_{\widetilde{V}}^2\,dt\right)^{\frac{1}{2}}\left(\tilde{E}\int_0^T \rho_t^{-2}\eta_t^{-2}\,dt\right)^{\frac{1}{2}} < \infty.$$

The inclusion $\pi \in L^1(\Omega,P,C(O,T;X))$ can be proved similarly. □

We conclude this section by deriving some auxiliary results needed below to justify the splitting up approximations. Let us consider the equation

$$\pi + \lambda A\pi = \phi, \tag{2.13}$$

where $\phi \in H$ and $\lambda > 0$ is a scalar.

DEFINITION 2.4. *We say that π is a solution of* (2.13) *if it belongs to V and the equality*

$$(\pi,\psi) + \lambda(A_0\nabla\pi - a\,\pi,\nabla\psi) = (\phi,\psi)$$

holds for all $\psi \in V$.

LEMMA 2.5. *Assume that $\lambda > 0$ is sufficiently small and $\phi \in H$ and is nonnegative a.e. in R^d. Then* (2.13) *has a unique solution and it is nonnegative (a.e. in R^d).*

Proof.[1] By standard arguments (see [22]) one can derive from (2.5) that if $\rho\lambda \leq 1$, $J_\lambda = (I + \lambda A)^{-1} \in \mathcal{L}(V^*,V)$ and (2.13) has a unique solution $\pi = J_\lambda\phi$, where ϕ is an H-valued random variable. By definition, $\pi \in V$ and satisfies

$$(\pi,\psi) + \lambda\left((A_0\nabla\pi,\nabla\psi) - (a\,\pi,\nabla\psi)\right) = (\phi,\psi)$$

[1]This result is well known; however, for the sake of completeness we give a short proof.

for all $\psi \in V$.

Let $\psi = \min(\pi, 0)$. Then $\psi \in V$ and

$$|\psi|_H^2 + \lambda\,((A_0\nabla\psi, \nabla\psi) - (a\,\psi, \nabla\psi)) = (\phi, \psi) \le 0.$$

It follows from (2.5) that

$$(1 - \lambda\rho)\,|\psi|_H^2 + \frac{\lambda\beta}{2}\,|\psi|_V^2 \le 0,$$

which implies that $|\psi|_H = 0$, and thus $\pi \ge 0$ a.e. in R^d. □

Of course it follows from Lemma 2.5 that for sufficiently small $\lambda > 0$, the system (2.17) has a unique nonnegative solution.

LEMMA 2.6. *If in addition to the assumptions of Lemma* 2.5 $\phi \in X$, *then a solution to* (2.13) *belongs to* $\widetilde{V}$.

Proof. The same arguments as in Lemma 2.5 yield that if $\tilde{\rho}\lambda \le 1$, then the equation

$$(\pi, v)_X + \lambda\,(A_0\nabla\pi - \pi\,a, \varphi\,\nabla v + v\nabla\varphi) = (\pi, v)_X \text{ for all } v \in \widetilde{V} \tag{2.14}$$

has a unique solution in $\widetilde{V}$. Now we will prove that a solution to (2.14) is also a solution to (2.13). This will give us the desired result since according to Lemma 2.5, a solution to the latter equation is unique.

Let ψ be an arbitrary function from $C_0^\infty(R^d)$. Then $\varphi\psi \in \widetilde{V}$. Now taking $v = \varphi^{-1}\psi$, we can rewrite (2.14) as follows:

$$(\pi, \psi) + \lambda\,((A_0\nabla\pi, \nabla\psi) - (a\,\pi, \nabla\psi)) = (\phi, \psi) \text{ for all } \psi \in C_0^\infty(R^d).$$

Since $C_0^\infty(R^d)$ is dense in V, we conclude that π is indeed a solution to (2.13). □

2.2. Operator-splitting approximations for the Kushner and Zakai equations. The operator splitting is an effective and natural approach to time discretization of the Zakai equation. In fact, it mimics the exact recursive formula which holds in the case of discrete time observations. In this subsection we define explicit and implicit operator-splitting approximations for the Kushner equation and discuss their relation to the splitting-up approximation of the Zakai equation.

For $T \in R_+$ and $m, k \in \mathbf{Z}$, set $\lambda = T/m$, $t_k = k\lambda$, $y_k = y(t_k)$, and $\Delta y_k = y_k - y_{k-1}$. The splitting-up discretization for the Zakai equation can be defined by the formulas[2]

$$\begin{aligned} &\alpha^k = \exp(h(x)\Delta y_k - \tfrac{\lambda}{2}|h(x)|^2)p^{k-1}, \\ &p^k - \alpha^k + \lambda A p^k = 0, \quad k = 1, 2, \ldots, \quad p^0 = p_0. \end{aligned} \tag{2.15}$$

This splitting-up scheme for the Zakai equation was introduced in [2]. The rate of convergence of this scheme and its higher-order modifications were studied in [8] and [11]; see also the references therein.

In practice, to avoid possible complications related to the fast dissipation of p^k and the effect of intermittency as the number of time steps grows, this scheme is often

[2]It is shown in Lemma 2.5 that for sufficiently small $\lambda > 0$ the inverse operator $(I + \lambda A)^{-1}$ is well defined.

modified as follows:

$$\begin{aligned} &\alpha^k = \exp(h(x)\Delta y_k - \tfrac{\lambda}{2}|h(x)|^2)\tilde{p}^{k-1}, \\ &\delta^k = \alpha^k / \textstyle\int \alpha^k \, dx, \text{ and} \\ &\tilde{p}^k - \delta^k + \lambda A \tilde{p}^k = 0,\ k = 1, 2, \dots,\ \tilde{p}^0 = p_0. \end{aligned} \tag{2.16}$$

As we mentioned before, in contrast to the former splitting-up scheme, the convergence of the latter one has never been studied rigorously.

Now let us consider two operator splitting-up approximations to the Kushner equation. We begin with the explicit one:

$$\begin{aligned} &\xi^k = \exp\left((h - \pi^{k-1}[h])(\Delta y_k - \pi^{k-1}[h]\lambda) - \tfrac{1}{2}|h(x) - \pi^{k-1}[h]|^2\lambda\right)\pi^{k-1}, \\ &\pi^k - \xi^k + \lambda A \pi^k = 0,\ k = 1, 2, \dots, \quad \pi^0 = p_0, \end{aligned} \tag{2.17}$$

where

$$\pi^{k-1}[h] = \frac{\int h(x)\pi^{k-1}\, dx}{\int \pi^{k-1}\, dx}. \tag{2.18}$$

Note that $\xi^k = \xi^k(t_k)$, where $\xi^k(\cdot)$ satisfies

$$\xi^k(t) = \pi^{k-1} + \int_{t_{k-1}}^{t} (h - \pi^{k-1}[h])\xi^k(s)(dy(s) - \pi^{k-1}[h]\, ds),\ t \in [t_{k-1}, t_k]. \tag{2.19}$$

Obviously, π^k is not necessarily a probability density and $\pi^{k-1}[h]$ defines the mean of h with respect to the normalizations of π^{k-1}.

Similarly, one can consider the implicit version of (2.17):

$$\begin{aligned} &\tilde{\xi}^k = \tilde{\pi}^{k-1} + \textstyle\int_{t_{k-1}}^{t_k} (h - \langle \tilde{\xi}_t, h\rangle)\tilde{\xi}(t)(dy(t) - \langle \tilde{\xi}_t, h\rangle\, dt), \\ &\tilde{\pi}^k - \tilde{\xi}^k + \lambda\, A\tilde{\pi}^k = 0. \end{aligned} \tag{2.20}$$

That is, $\tilde{\xi}^k = \tilde{\xi}(t_k)$, where $\tilde{\xi}(\cdot)$ satisfies the stochastic integro-differential equation

$$d\tilde{\xi}_t = (h - \langle \tilde{\xi}_t, h\rangle)\tilde{\xi}_t(dy(t) - \langle \tilde{\xi}_t, h\rangle\, dt), \quad \tilde{\xi}(t_{k-1}) = \tilde{\pi}^{k-1}, \tag{2.21}$$

on the interval $(t_{k-1}, t_k]$. By Ito's lemma a solution to this equation is given by the formula[3]

$$\tilde{\xi}_t = \beta_t / \langle \beta_t, 1\rangle,$$

where $\beta_t = \beta_t(x)$ is a solution to the stochastic equation

$$d\beta_t(x) = h(x)\beta_t(x)\, dy(t),\ \beta_{t_{k-1}} = \tilde{\pi}^{k-1}.$$

[3]Everywhere below, $\langle , \rangle$ denotes the duality between $L^1(R^d)$ and $L^\infty(R^d)$. In particular, $\langle \beta_t, 1\rangle = \int_{R^d} \beta_t(x)\, dx$.

The implicit approximation to the Kushner equation given by (2.20) is equivalent to the splitting-up scheme (2.16) for the Zakai equation with normalization at each step. Indeed, setting up $\beta^k = \beta(t_k)$, we can rewrite (2.21) as follows:

$$\begin{aligned} &\beta^k = \exp\left(h(x)\Delta y_k - \frac{\lambda}{2}|h(x)|^2\right)\tilde{\pi}^{k-1}\,,\ \gamma^k = \frac{\beta^k}{\langle \beta^k, 1\rangle}, \\ &\tilde{\pi}^k - \gamma^k + \lambda\, A\tilde{\pi}^k = 0,\ k = 1, 2, \ldots,\ \tilde{\pi}^0 = p_0. \end{aligned} \tag{2.22}$$

Obviously Lemmas 2.5 and 2.6 yield that the sequences $\{\pi^k\}$ and $\{\tilde{\pi}^k\}$ are well defined. The discretized solution map (2.15) for the Zakai equation is given by

$$p^k = (I + \lambda\, A)^{-1}\alpha^k, \quad \text{where } \alpha^k = \gamma^k_{k-1}p^{k-1}, \tag{2.23}$$

where

$$\gamma^k_{k-1} = \exp\left(h\,\Delta y_k - \frac{\lambda}{2}|h(x)|^2\right) \text{ and } \Delta y_k = y(t_k) - y(t_{k-1}).$$

Set

$$q^k_{k-1} = \exp\left(\pi^{k-1}[h]\,\Delta y_k - \frac{\lambda}{2}|\pi^{k-1}[h]|^2\right).$$

Since q^k_{k-1} does not depend on x, a solution to the approximation (2.17) can be written as

$$\pi^k = (q^k_{k-1})^{-1}(I + \lambda\, A)^{-1}\gamma^k_{k-1}\pi^{k-1}.$$

By induction we have the following relation between π^n and p^n:

$$\pi^n = (q^n)^{-1}\, p^n, \quad \text{where} \quad q^n = \Pi^n_{k=1}\, q^k_{k-1.} \tag{2.24}$$

Of course, (2.24) implies that $\langle \pi^n, 1\rangle = (q^n)^{-1}\ \langle p^n, 1\rangle$, and so for any $f \in X^*$,

$$\pi^n\,[f] = p^n\,[f]\,. \tag{2.25}$$

For $s \in (t_{k-1}, t_k]$, write

$$\gamma_k(s) = \exp\{h(x)\,(y(s) - y(t_{k-1})) - (s - t_{k-1})\,|h(x)|^2/2\}$$

and set

$$\begin{aligned} &p^k(s,x) = (I + \lambda A)^{-1}\,\alpha^k(s,x), \\ &\alpha^k(s,x) = \gamma_k(s)p^{k-1}(x),\ \text{and } p^0(s) = p_0. \end{aligned} \tag{2.26}$$

Obviously, $\gamma_k(t_k) = \gamma^k_{k-1}$ and $\alpha^k(t_k, x) = \alpha^k(x)$. Define the function $\alpha^\lambda_t(x) = \alpha^k(t,x)$ for $t \in [t_{k-1}, t_k)$.

LEMMA 2.7. *For every k,*

$$\begin{aligned} \langle p^k, 1\rangle = \langle p^{k-1}, 1\rangle \exp\left\{\int_{t_{k-1}}^{t_k} \alpha^k_s\,[h]\ dy(s) - \frac{1}{2}\int_{t_{k-1}}^{t_k} |\alpha^k_s\,[h]\,|^2\, ds\right\} \\ = \exp\left\{\int_0^{t_k} a^\lambda_s\,[h]\, dy(s) - \frac{1}{2}\int_0^{t_k} \left|\alpha^\lambda_s\,[h]\right|^2\, ds\right\} \end{aligned}$$

P—a.e., where $\alpha_s^k[h] = \langle \alpha^k(s), h\rangle / \langle \alpha^k(s), 1\rangle$ *and* $\alpha_t^\lambda[h] = \alpha_t^k[h]$ *for* $t \in (t_{k-1}, t_k]$.

Proof. Owing to (2.26) we have $(I + \lambda A)\, p^k(s,x) = \alpha^k(s,x)$. Note also that for every $f \in H^1(R^d)$,

$$\int_{R^d} \nabla(A_0 \nabla f - af)\, dx = 0.$$

Therefore,

$$\langle p^k(s), 1\rangle = \langle \alpha^k(s), 1\rangle = \int_{R^d} \gamma_k(s,x) p^{k-1}(x)\, dx. \tag{2.27}$$

Since for $t \in (t_{k-1}, t_k]$, $d\gamma_k(t) = h(x)\gamma_k(t)\, dy(t)$, $\gamma_k(t_{k-1}) = 1$, it follows from (2.27) that

$$\begin{aligned}
\langle p^k(t), 1\rangle &= \langle p^{k-1}, 1\rangle + \textstyle\int_{t_{k-1}}^t \int_{R^d} h(x)\gamma_k(s,x) p^{k-1}(x)\, dx\, dy(s) \\
&= \langle p^{k-1}, 1\rangle + \textstyle\int_{t_{k-1}}^t \langle p^k(s), 1\rangle \int_{R^d} h(x)\gamma_k(s,x) p^{k-1}(x)\, dx \langle \alpha^k(s), 1\rangle^{-1}\, dy(s) \\
&= \langle p^{k-1}, 1\rangle + \textstyle\int_{t_{k-1}}^t \langle p^k(s), 1\rangle \langle \alpha^k(s), h\rangle \langle \alpha^k(s), 1\rangle^{-1}\, dy(s).
\end{aligned}$$

Hence

$$\langle p^k(t), 1\rangle = \langle p^{k-1}, 1\rangle \exp\left\{ \int_{t_{k-1}}^t \alpha_s^k[h]\, dy(s) - \frac{1}{2}\int_{t_{k-1}}^t \left|\alpha_s^k[h]\right|^2 ds \right\},$$

which completes the proof. □

3. Convergence of splitting-up approximations.

3.1. Convergence in $L^1(\Omega, \mathcal{F}, P)$. In this section we will establish weak and strong convergence of the splitting-up approximation (2.17) to the solution of the Kushner equation. Since the differences in the proof of the main results for the sequences $\{\pi^k\}$ and $\{\tilde\pi^k\}$ are minimal, note that

$$\tilde\pi^n = (\tilde q^n)^{-1} p^n \ \text{ with } \ \tilde q^n = \exp\left(\sum_{k=1}^n \int_{t_{k-1}}^{t_k} \left(\tilde\xi_t[h]\, dy(t) - \frac{1}{2} |\tilde\xi_t[h]|^2\, dt \right) \right);$$

we will consider here only the former one and leave the latter to the interested reader.

Define the functions $\pi_\lambda(t)$ and p_t^λ by

$$\pi_\lambda(t) = \pi^k \ \text{ and } \ p_t^\lambda = p^k \quad \text{for} \quad t \in [k\lambda, (k+1)\lambda)\ k \ge 0.$$

Also, for $t \in [k\lambda, (k+1)\lambda)$, set $\rho_t^\lambda = \exp\{\int_0^{t_k} \alpha_s^k[h]\, dy(s) - \frac{1}{2}\int_0^{t_k} \left|\alpha_s^k[h]\right|^2 ds\}$. Owing to Lemma 2.7, we have

$$\rho_t^\lambda = \langle p_t^\lambda, 1\rangle. \tag{3.1}$$

First we will establish two auxiliary results, Lemma 3.1 and Theorem 3.2.

LEMMA 3.1. *Let* $\mu_1(t)$ *and* $\mu_2(t)$ *be* $\mathcal{F}_t^y$*-adapted processes so that*

$$\sup_t (|\mu_1(t)|^2 + |\mu_2(t)|) \le C|h|_\infty^2.$$

Write $\theta_t = \exp\{\int_0^t \mu_1(s)dy(s) + \int_0^t \mu_2(s)ds\}$. *For all* $f \in L^\infty(R^d)$ *and any* $q \geq 1$, *there exist constants* $M(q)$ *and* $c(q)$ *such that*

$$E\theta_t |\langle p_t^\lambda - p(t), f\rangle|^q \leq M(q) \exp\left(c(q)|h|_\infty^2 t\right) |f|_\infty^q \left|\varphi^{-1}\right|_1^{\frac{1}{2}} (\tilde{E}\,|p_t^\lambda - p(t)|_X^2)^{\frac{1}{2}}.$$

Proof. Assume that $t \in [k\lambda, (k+1)\lambda)$. Changing the probability law according to (2.2), applying the Schwarz inequality, and using (2.12) and (3.1), we get

$$\begin{aligned} E\theta_t |\langle p_t^\lambda - p(t), f\rangle|^q &= \tilde{E}\eta_t^{-1}\theta_t |\langle p_t^\lambda - p(t), f\rangle|^q \\ &\leq |f|_\infty^{q-1} \tilde{E}\left[\eta_t^{-1}\theta_t |\rho_t^\lambda + \rho_t|^{q-1} |\langle p_t^\lambda - p(t), f\rangle|\right] \\ &\leq |f|_\infty^q \left|\varphi^{-1}\right|_1^{\frac{1}{2}} \tilde{E}\left[\eta_t^{-1}\theta_t |\rho_t^\lambda + \rho_t|^{q-1} |p_t^\lambda - p(t)|_X\right] \\ &\leq |f|_\infty^q \left|\varphi^{-1}\right|_1^{\frac{1}{2}} (\tilde{E}|p_\lambda(t) - p(t)|_X^2)^{\frac{1}{2}} (\tilde{E}(\eta_t^{-2}\theta_t^2 |\rho_t^\lambda + \rho_t|^{2q-2}))^{\frac{1}{2}}. \end{aligned} \tag{3.2}$$

It follows from (2.1), (2.12), and (3.1) that

$$(\eta_t^{-2}\theta_t^2 |\rho_t^\lambda + \rho_t|^{2q-2}) \leq M(q)\left(I_t^\lambda + I_t\right),$$

where

$$I_t^\lambda = \exp(\int_0^t 2((q-1)\alpha_s^\lambda[h]1_{\{s<(k+1)\lambda\}} + h(x(s)) + \mu_1(s))\, dy(s)$$

$$-\tfrac{1}{2}\int_0^t 4|(q-1)\alpha_s^\lambda[h]1_{\{s<(k+1)\lambda\}} + h(x(s)) + \mu_1(s)|^2\, ds)$$

and

$$I_t = \exp(\int_0^t 2((q-1)\pi_s[h] + h(x(s)) + \mu_1(s))\, dy(s)$$

$$-\tfrac{1}{2}\int_0^t 4|(q-1)\pi_s[h] + h(x(s)) + \mu_1(s)|^2\, ds).$$

Since $\pi_s[h] \leq |h|_\infty$ and $\alpha_s^\lambda[h] \leq |h|_\infty$, I_t and I_t^λ are martingales with respect to the filtration $(\Omega, \mathcal{F}, \{\mathcal{F}_t^y\}_{t\leq T}, \tilde{P})$. Therefore,

$$\tilde{E}\eta_t^{-2}\theta_t |\rho_t^\lambda + \rho_t|^{2q-2} \leq M(q) \exp(c(q)|h|_\infty^2 t), \tag{3.3}$$

which implies the desired estimate. □

The following result follows in exactly the same manner as in [11].

THEOREM 3.2. *The following estimate holds uniformly in* λ :

$$\sup_{t\leq T} \tilde{E}|p_t^\lambda|_X^2 + \tilde{E}\int_0^T |p_t^\lambda|_{\widetilde{V}}^2\, dt < \infty.$$

Moreover,

$$\sup_{t\leq T}\ \tilde{E}|p_t^\lambda - p(t)|_X^2 + \tilde{E}\int_0^T |p_t^\lambda - p(t)|_{\widetilde{V}}^2 \to 0 \text{ as } \lambda \to 0.$$

The main results of this section are given in the following two theorems.

THEOREM 3.3. *For every $f \in L^\infty(R^d)$ and $q \geq 1$,*

$$E\,|\pi_\lambda(t)[f] - \pi_t[f]|^q \to 0 \text{ as } \lambda \to 0.$$

Moreover, there exist constants M and c (which do not depend on q) so that

$$\sup_{t\leq T} E\,|\pi_\lambda(t)[f] - \pi_t[f]|^q \leq M \exp\{cT\,|h|^2_\infty\}\,\left|\varphi^{-1}\right|_1^{\frac{1}{2}}\,|f|^q_\infty \sup_{t\leq T}(\tilde{E}\,|p^\lambda_t - p(t)|^2_X)^{\frac{1}{2}}.$$

Proof. Since $|\pi_\lambda(t)[f] - \pi(t)[f]| \leq |f|_\infty$, it suffices to prove the theorem only for $q = 1$. It follows from (2.25) that for $f \in L^\infty(R^d)$,

$$\begin{aligned}\pi_\lambda(t)[f] - \pi(t)[f] &= \frac{\langle p^\lambda_t, f\rangle}{\langle p^\lambda_t, 1\rangle} - \frac{\langle p(t), f\rangle}{\langle p(t), 1\rangle} \\ &= -\rho_t^{-1}(\langle p(t) - p^\lambda_t, f\rangle + p^\lambda_t[f]\langle p^\lambda_t - p(t), 1\rangle).\end{aligned} \tag{3.4}$$

By Lemma 3.1, we have

$$\begin{aligned}&E\,|\rho_t^{-1}\langle p(t) - p^\lambda_t, f\rangle| \\ &\quad\leq \tfrac{M}{2}\,\exp\{c\,|h|^2_\infty\,t\}\,\left|\varphi^{-1}\right|_1^{\frac{1}{2}}\,|f|_\infty\,(\tilde{E}\,|p^\lambda_t - p(t)|^2_X)^{\frac{1}{2}}.\end{aligned} \tag{3.5}$$

Since $|p^\lambda_t[f]| \leq |f|_\infty$, Lemma 3.1 yields

$$\begin{aligned}&E|\rho_t^{-1}p^\lambda_t[f]\langle\, p^\lambda_t - p(t), 1\rangle| \\ &\quad\leq \tfrac{M}{2}\,\exp\{c\,|h|^2_\infty\,t\}\,\left|\varphi^{-1}\right|_1^{\frac{1}{2}}\,|f|_\infty(\tilde{E}\,|p^\lambda_t - p(t)|^2_X)^{\frac{1}{2}}.\end{aligned} \tag{3.6}$$

The result follows now from Theorem 3.2. □

THEOREM 3.4. *The function π^λ converges to π, the solution of the Kushner equation* (1.1)*, strongly in $L^1((0,T)\times\Omega, \widetilde{V})$ as $\lambda \to 0$. Moreover, the following estimate holds:*

$$E\int_0^T |\pi_\lambda(t) - \pi(t)|_{\widetilde{V}}\,dt \leq C\left[\left(\tilde{E}\int_0^T |p^\lambda_t - p(t)|^2_{\widetilde{V}}dt\right)^{\frac{1}{2}} + \left(\int_0^T (\tilde{E}|p^\lambda_t - p(t)|^2_X)^{\frac{1}{2}}dt\right)^{\frac{1}{2}}\right].$$

Proof. Note that

$$\pi_\lambda(t) - \pi(t) = \rho_t^{-1}(p^\lambda_t - p(t)) + p^\lambda_t(\rho_t\rho^\lambda_t)^{-1}\,\langle p^\lambda_t - p(t), 1\rangle, \tag{3.7}$$

and thus

$$\begin{aligned}E\int_0^T |\pi_\lambda(t) - \pi(t)|_{\widetilde{V}}\,dt &\leq E\int_0^T \rho_t^{-1}|p_t - p^\lambda_t|_{\widetilde{V}}\,dt \\ &\quad + E\int_0^T |p^\lambda_t|_{\tilde{V}}(\rho_t\rho^\lambda_t)^{-1}|\,\langle p^\lambda_t - p(t), 1\rangle\,|\,dt.\end{aligned}$$

Now changing the probability law according to (2.2) and applying the Schwarz inequality, we get

$$E\int_0^T \rho_t^{-1}|p^\lambda_t - p(t)|_{\widetilde{V}}\,dt = \left(\tilde{E}\int_0^T |\eta_t^{-1}\rho_t{}^{-1}|^2\,dt\right)^{\frac{1}{2}}\left(\tilde{E}\int_0^T |p^\lambda_t - p(t)|^2_{\widetilde{V}}\,dt\right)^{\frac{1}{2}}. \tag{3.8}$$

Write $\zeta_t^\lambda = (\eta_t \rho_t \rho_t^\lambda)^{-1}$. Now by the Schwarz inequality and Lemma 3.1 we have

$$
\begin{aligned}
&E \int_0^T |p_t^\lambda|_{\widetilde{V}} (\rho_t \rho_t^\lambda)^{-1} |\, \langle p_t^\lambda - p(t), 1 \rangle |\, dt \\
&\leq \left(\tilde{E} \int_0^T |\zeta_t^\lambda \langle p_t^\lambda - p(t), 1 \rangle|^2 \, dt \right)^{\frac{1}{2}} \left(\tilde{E} \int_0^T |p_t^\lambda|_{\widetilde{V}}^2 \, dt \right)^{\frac{1}{2}} \qquad (3.9) \\
&\leq C \left(\int_0^T (\tilde{E} |p_t^\lambda - p(t)|_X^2)^{\frac{1}{2}} \, dt \right)^{\frac{1}{2}} \left(\tilde{E} \int_0^T |p_t^\lambda|_{\widetilde{V}}^2 \, dt \right)^{\frac{1}{4}}.
\end{aligned}
$$

Obviously,

$$
\tilde{E} \int_0^T |\eta_t^{-1} \rho_t^{-1}|^2 \, dt < \infty. \qquad (3.10)
$$

The statement follows now from Theorem 3.2 and (3.8)–(3.10). □

COROLLARY 3.5. *The following estimate holds:*

$$
E|\pi^n - \pi(t_n)|_X \leq C \left[1 + \left(\tilde{E} |p^n|_X^4 \right)^{\frac{1}{4}} \right] \left(\tilde{E} |p^n - p(t_n)|_X^2 \right)^{\frac{1}{2}}. \qquad (3.11)
$$

Proof. From (3.7) and Hölder's inequality,

$$
\begin{aligned}
E|\pi^n - \pi(t_n)|_X \leq |\varphi|_2^{\frac{1}{2}} \Big[&\left(\tilde{E} |(\eta(t_n) \rho(t_n)^{-1}|^2 \right)^{\frac{1}{2}} \left(\tilde{E} |p^n - p(t_n)|_X^2 \right)^{\frac{1}{2}} \\
&+ \left(\tilde{E} |p^n|_X^4 \right)^{\frac{1}{4}} \left(\tilde{E} |\eta(t_n) (\rho(t_n) \rho^n)^{-1}|^4 \right)^{\frac{1}{4}} \left(\tilde{E} |p^n - p(t_n)|_X^2 \right)^{\frac{1}{2}} \Big],
\end{aligned}
$$

which shows the estimate. □

In section 4 we establish a bound on $(\tilde{E}|p^n|_X^4)^{\frac{1}{4}}$ and the convergence estimate for $(\tilde{E}|p^n - p(t_n)|_X^2)^{\frac{1}{2}}$. It will be shown that

$$
\sup_n E|\pi^n - \pi(t_n)|_X \leq C \lambda \qquad (3.12)
$$

provided that $p(t)$ satisfies an appropriate regularity condition.

3.2. Convergence in $L^2(\Omega, F, \tilde{P})$. In this section we show that $\pi_\lambda(\cdot)$ converges strongly in $L^2(\tilde{S}; \widetilde{V})$. First we derive some preliminary estimates for π^k. Multiplying (2.17) by π^k, we obtain

$$
(\pi^k - \xi^k, \pi^k)_X + \lambda \langle A\pi^k, \pi^k \rangle_X = 0.
$$

Now completing the square, we obtain

$$
\frac{1}{2} (|\pi^k|_X^2 - |\xi^k|_X^2 + |\pi^k - \xi^k|_X^2) + \lambda \langle A\pi^k, \pi^k \rangle_X = 0,
$$

and thus from (2.7)

$$
(1 - 2\lambda\tilde{\rho}) \tilde{E} |\pi^k|_X^2 + \tilde{E} |\pi^k - \xi^k|_X^2 + \lambda\beta \tilde{E} |\pi^k|_{\widetilde{V}}^2 \leq \tilde{E} |\xi^k|_X^2. \qquad (3.13)
$$

Since $p_0 \geq 0$ a.e., it follows from Lemma 2.5 that the sequence $\{\pi^k\}$ generated by (2.17) satisfies $\pi^k \geq 0$ a.e. for each $k \geq 0$. Thus

$$|h - \pi^{k-1}[h]|_\infty \leq |h|_\infty \quad \text{and} \quad |\pi^{k-1}[h]|_\infty \leq |h|_\infty.$$

It follows from (2.19) and Ito's lemma that for $t \in [t_{k-1}, t_k]$,

$$\tilde{E}\,|\xi^k(t)|_X^2 - \tilde{E}\,|\pi^{k-1}|_X^2 = \tilde{E}\int_{t_{k-1}}^t \left(2(h - \pi^{k-1}[h])\pi^{k-1}[h] + (h - \pi^{k-1}[h])^2\right)\,|\xi^k(s)|_X^2\,ds$$

$$\leq \int_{t_{k-1}}^t 3|h|_\infty^2\,\tilde{E}\,|\xi(s)|_X^2\,ds.$$

From Gronwall's lemma,

$$\tilde{E}\,|\xi^k|_X^2 \leq e^{3|h|_\infty^2 \lambda}\,\tilde{E}\,|\pi^{k-1}|_X^2.$$

Thus from (3.13), we obtain

$$e^{-\tilde{\omega}\lambda}\,\tilde{E}\,|\pi^k|_X^2 + e^{-3\lambda\,|h|_\infty^2}\,(\tilde{E}\,|\pi^k - \xi^k|_X^2 + \lambda\beta\,\tilde{E}\,|\pi^k|_{\widetilde{V}}^2) \leq \tilde{E}\,|\pi^{k-1}|_X^2,$$

where $\tilde{\omega} = 1 + 2\,\tilde{\rho} + 3\,|h|_\infty^2$. (To obtain the latter inequality we used the fact that $e^{-\lambda(1+2\tilde{\rho})} \leq 1 - 2\lambda\tilde{\rho}$ for sufficiently small $\lambda > 0$.) Multiplying this by $e^{-\tilde{\omega}\,t_{k-1}}$, we have for $k \geq 1$

$$e^{-\tilde{\omega}t_k}\,(\tilde{E}\,|\pi^k|_X^2 + \tilde{E}\,|\pi^k - \xi^k|_X^2 + \lambda\beta\,\tilde{E}\,|\pi^k|_{\widetilde{V}}^2) \leq e^{-\tilde{\omega}t_{k-1}}\,\tilde{E}\,|\pi^{k-1}|_X^2.$$

Summing up both sides of the inequality in k, we obtain

$$e^{-\tilde{\omega}t_m}\,\tilde{E}\,|\pi^m|_X^2 + \tilde{E}\sum_{k=1}^m e^{-\tilde{\omega}t_k}\,(|\pi^k - \xi^k|_X^2 + \lambda\beta\,|\pi^k|_{\widetilde{V}}^2) \leq \tilde{E}\,|\pi^0|_X^2. \tag{3.14}$$

Assume that $p_0 \in X$. Then it follows from (3.14) that

$$\sup_{t\in[0,T]} \tilde{E}\,|\pi_\lambda(t)|_X^2 + \beta\int_0^T \tilde{E}\,|\pi_\lambda(t+\lambda)|_{\widetilde{V}}^2\,dt \leq e^{\tilde{\omega}\,T}\,\tilde{E}\,|p_0|_X^2. \tag{3.15}$$

In addition, since from Lemma 2.5 $\pi^k \geq 0$ a.e., we have $\pi_\lambda(t) \geq 0$ a.e. in $\tilde{S}$.

Next we show that $\pi_\lambda(t)$ converges weakly to $\pi(t)$ in $L^2(\tilde{S};\widetilde{V})$.

THEOREM 3.6. *The sequence $\{\pi^\lambda(t)\}$ converges weakly in $L^2(\tilde{S};\widetilde{V})$ and weakly star in $L^\infty(0,T,L^2(\Omega,\tilde{P};X))$ to the unique solution $\pi(t)$ of the Kushner equation*

$$\pi(t) = p_0 - \int_0^t A\pi(s)\,ds + \int_0^t (h - \pi[h](s))\pi(s)(dy(s) - \pi[h](s)\,ds) \quad \text{in } \widetilde{V}^*. \tag{3.16}$$

Proof. From (3.15) there exists a subsequence of $\{\pi_\lambda\}$, which will be denoted by the same symbol, such that $\pi_\lambda \to \pi$ weakly in $L^2(\tilde{S};\widetilde{V})$ and weakly star in $L^\infty(0,T;L^2(\Omega;\tilde{P},X))$, where $\pi(t) \in L^2(\tilde{S};\widetilde{V}) \cap sL^\infty(0,T;L^2(\Omega,\tilde{P},X))$ can be chosen

to be predictable. Moreover, $\pi \geq 0$ a.e. in $\tilde{S}$. If $[s]$, $s \in R$, denotes the largest integer j such that $j \leq s$, then from (2.17) we have

$$\begin{aligned}\pi_\lambda(t) = p_0 &- \int_0^{[t/\lambda]\lambda} A\pi_\lambda(s+\lambda)\,ds \\ &+ \int_0^{[t/\lambda]\lambda} (F(\pi_\lambda(s))\pi_\lambda(s) + r_\lambda(s))\,(dy(s) - \pi_\lambda[h](s)\,ds),\end{aligned} \tag{3.17}$$

where the operator F on X is defined by

$$F(\pi) = (h - \pi[h]) \quad \text{for } \pi \in X.$$

The residual function $r_\lambda(\cdot)$ is defined by

$$r_\lambda(t) = F(\pi_\lambda(t))(\xi_\lambda(t) - \pi_\lambda(t)),$$

where

$$\xi_\lambda = \xi^k(t) \quad \text{if} \quad t \in ((k-1)\lambda, k\lambda], \quad k \geq 1,$$

and

$$\xi^k(t) = \pi^{k-1} + \int_{t_{k-1}}^t (h - \pi^{k-1}[h])\xi^k(s)(dy(s) - \pi^{k-1}[h]\,ds).$$

Note that

$$(\phi, \psi)_H = (\phi, \varphi^{-1}\psi)_X \quad \text{and} \quad |\varphi^{-1}\psi|_X \leq |\varphi^{-1}|_1^{\frac{1}{2}} |\psi|_\infty$$

for $\psi \in L^\infty(R^d)$. Thus $(\pi_\lambda, \psi) \in L^2(\tilde{S}; R)$. Moreover, as proved in section 3.1,

$$\pi_\lambda[h] \to \pi[h] \quad \text{in } L^q(\tilde{S}; R) \tag{3.18}$$

for $h \in L^\infty(R^d)$ and $2 \leq q < \infty$. Thus it follows from the Lebesgue-dominated convergence theorem that

$$F(\pi_\lambda)\pi_\lambda \to F(\pi)\pi \quad \text{weakly in } L^2(\tilde{S}, X). \tag{3.19}$$

By Ito's lemma we have for $t \in [t_{k-1}, t_k]$

$$\begin{aligned}&\tilde{E}\,|\xi_\lambda(t) - \pi_\lambda(t)|_X^2 \\ &= \tilde{E} \int_{t_{k-1}}^t 2((h - \pi_\lambda[h])\,\xi_\lambda(s), \xi_\lambda(s) - \pi_\lambda(s) + |(h - \pi_\lambda[h])\,\xi_\lambda(s)|_X^2)\,ds \\ &\leq \int_{t_{k-1}}^t |h|_\infty^2\,\tilde{E}\,|\xi_\lambda(s) - \pi_\lambda(s)|_X^2 + 2\tilde{E}\,|(h - \pi_\lambda[h])\,\xi_\lambda(s)|_X^2\,ds.\end{aligned} \tag{3.20}$$

It follows from Gronwall's lemma that

$$\tilde{E}\,|\xi_\lambda(t) - \pi_\lambda(t)|_X^2 \leq 2|h|_\infty^2 e^{|h|_\infty^2\,(t - t_{k-1})} \int_{t_{k-1}}^t \tilde{E}\,|\xi_\lambda(s)|_X^2\,ds \leq M\,(t - t_{k-1}) \tag{3.21}$$

for $t \in [t_{k-1}, t_k]$ and some $M > 0$ independent of $\lambda > 0$, where we used the fact that

$$\tilde{E}\,|\xi_\lambda(t)|_X^2 \le e^{\tilde{\omega} t_k}\,\tilde{E}\,|p_0|_X^2.$$

Hence we have

$$\tilde{E}\int_0^T |r_\lambda(t)|_X^2\,dt \le |h|_\infty^2\,\tilde{E}\int_0^T |\xi_\lambda(t)-\pi_\lambda(t)|_X^2\,dt \le M\,\lambda. \tag{3.22}$$

It thus follows from (3.15)–(3.22) that for almost all (t,ω) a continuous modification of $\pi(t) \in L^2(\tilde{S};\widetilde{V})$ in X (see [12, Theorem 3.2]) satisfies (3.16) and $\pi \ge 0$. Note that $1 \in X^*$ and

$$\begin{aligned}(\pi(t),1)_H &= (p_0,1)_H + \int_0^t ((\pi(s),h)_H - \pi[h](s)(\pi(s),1))(dy(s) - \pi[h](s)\,ds)_H \\ &= (p_0,1)_H = 1\end{aligned} \tag{3.23}$$

a.s. Thus the random probability density $\pi(\cdot) \in L^2(\tilde{S};\widetilde{V}) \cap L^\infty(\tilde{S};X)$ satisfies the Kushner equation (1.1) in the sense that (3.16) is satisfied a.e. in $\tilde{S}$. □

Finally we show that $\pi_\lambda(t)$ converges strongly to $\pi(t)$ in $L^2(\tilde{S};\widetilde{V})$.

THEOREM 3.7. *The sequence $\{\pi^\lambda(t)\}$ converges strongly in $L^2(\tilde{S};\widetilde{V})$ to the unique solution $\pi(t)$ of the Kushner equation* (3.16).

Proof. Note that (3.17) is equivalently written as

$$\pi^k + \lambda\,A\pi^k - \lambda\,G(\pi^{k-1})\pi^{k-1} = \pi^{k-1} + F(\pi^{k-1})\pi^{k-1}\,\Delta y_k + e^k,$$

where

$$G(\pi) = \pi[h]\,(h - \pi[h]) \quad \text{for } \pi \in X$$

and

$$e^k = \int_{t_{k-1}}^{t_k} r_\lambda(s)(dy(s) - \pi^{k-1}[h]\,ds).$$

Multiplying this by π^k we have

$$\begin{aligned}&\frac{1}{2}\,(|\pi^k|_X^2 + |\pi^k - \pi^{k-1}|_X^2 - |\pi^{k-1}|_X^2) + \lambda\,(\langle A\pi^k,\pi^k\rangle_X - G(\pi^{k-1})\pi^k,\pi^k) \\ &\quad = (e^k - \lambda\,G(\pi^{k-1})(\pi^k - \pi^{k-1}),\pi^k) + (F(\pi^{k-1})\pi^{k-1}\,\Delta y_k,\pi^k).\end{aligned}$$

Since

$$\tilde{E}\,(F(\pi^{k-1})\pi^{k-1}\,\Delta y_k,\pi^{k-1}) = 0,$$

we have

$$\tilde{E}\,(F(\pi^{k-1})\pi^{k-1}\,\Delta y_k,\pi^k) \le \frac{1}{2}\,(\tilde{E}\,|\pi^k - \pi^{k-1}|_X^2 + \tilde{E}\,|F(\pi^{k-1})\pi^{k-1}\,\Delta y_k|_X^2).$$

Thus we obtain

$$\begin{aligned}&\tilde{E}\,|\pi^k|_X^2 + 2\lambda\,\tilde{E}\,(\langle A\pi^k,\pi^k\rangle_X - G(\pi^{k-1})\pi^k,\pi^k) \\ &\le \tilde{E}\,|\pi^{k-1}|_X^2 + \lambda\,\tilde{E}\,|F(\pi^{k-1})\pi^{k-1}|_X^2 + 2\,\tilde{E}\,(e^k - \lambda\,G(\pi^{k-1})(\pi^k - \pi^{k-1}),\pi^k).\end{aligned} \tag{3.24}$$

First, we note that

$$2\langle A\pi,\pi\rangle_X - 2(G(\pi^{k-1})\pi,\pi) - |F(\pi)\pi|^2 + \tilde{\omega}\,|\pi|_X^2 \geq \beta\,|\pi|_{\widetilde{V}}^2 \quad \text{for } \pi \in \widetilde{V}. \tag{3.25}$$

Second, we note that since

$$\tilde{E}\,|\pi^k - \pi^{k-1}|_X^2 \leq 2(\tilde{E}\,|\pi^k - \xi^k|_X^2 + \tilde{E}\,|\xi^k - \pi^{k-1}|_X^2),$$

it follows from (3.14)–(3.21) that

$$\sum_{k=1}^{m} \tilde{E}\,|\pi^k - \pi^{k-1}|_X^2 \leq \text{const.} \tag{3.26}$$

It follows from (3.22) that

$$\sum_{k=1}^{m} \tilde{E}\,|e^k|_X^2 \leq \text{const}\,\lambda. \tag{3.27}$$

Multiplying (3.24) by c^{k-1} with $c = 1 - \lambda\tilde{\omega}$, we obtain for $k \geq 1$

$$\begin{aligned} &c^k \tilde{E}\,|\pi^k|_X^2 - c^{k-1}\tilde{E}\,|\pi^{k-1}|_H^2 \\ &\qquad + \lambda c^{k-1}\,\tilde{E}\,(2\langle A\pi^k,\pi^k\rangle - 2(G(\pi^{k-1})\pi^k,\pi^k) - |F(\pi^{k-1})\pi^{k-1}|^2) \\ &\quad \leq 2c^{k-1}\tilde{E}\,(e^k - \lambda\,(G(\pi^{k-1})(\pi^k - \pi^{k-1}),\pi^k)). \end{aligned}$$

Summing up this in k, we obtain

$$\begin{aligned} &c^m \tilde{E}\,(|\pi_\lambda(T)|_X^2 + \lambda\,|F(\pi_\lambda(T))\pi_\lambda(T)|_X^2) - \tilde{E}\,(|p_0|_X^2 + \lambda\,|F(p_0)p_0|_X^2) + T_\lambda \\ &\qquad \leq 2\sum_{k=1}^{m} c^{k-1}\tilde{E}\,(e^k - \lambda\,(G(\pi^{k-1})(\pi^k - \pi^{k-1}),\pi^k)), \end{aligned} \tag{3.28}$$

where

$$\begin{aligned} T_\lambda = \int_0^T & e_\lambda(t)\,\tilde{E}\,(2\langle A\pi_\lambda(t+\lambda),\,\pi_\lambda(t+\lambda)\rangle + \tilde{\omega}\,|\pi_\lambda(t+\lambda)|_X^2 \\ & - 2(G(\pi_\lambda(t))\pi_\lambda(t+\lambda),\pi_\lambda(t+\lambda)) - |F(\pi_\lambda(t+\lambda))\pi_\lambda(t+\lambda)|^2)\,dt, \end{aligned} \tag{3.29}$$

where

$$e_\lambda(t) = c^k \quad \text{for } t \in (t_{k-1}, t_k].$$

It follows from (3.22) and (3.26) that

$$\sum_{k=1}^{m} c^{k-1}\,\tilde{E}\,(e^k - \lambda\,G(\pi^{k-1})(\pi^k - \pi^{k-1}),\pi^k) \to 0$$

as $\lambda \to 0$. Define

$$\begin{aligned} S_\lambda = \int_0^T & e_\lambda(t)\,\tilde{E}(2\langle A\pi_\lambda(t+\lambda) - A\pi(t), \pi(t+\lambda) - \pi(t)\rangle + \tilde{\omega}\,|\pi(t+\lambda) - \pi(t)|_X^2 \\ & - 2(G(\pi_\lambda(t))(\pi_\lambda(t+\lambda) - \pi(t)), \pi_\lambda(t+\lambda) - \pi(t)) \\ & - |F(\pi_\lambda(t+\lambda)(\pi_\lambda(t+\lambda) - \pi(t))|_X^2)\,dt. \end{aligned} \tag{3.30}$$

It then follows from (3.24) that

$$S_\lambda \geq \beta \int_0^T e_\lambda(t)\, \tilde{E}\, |\pi_\lambda(t+\lambda) - \pi(t)|^2_{\widetilde{V}}\, dt. \tag{3.31}$$

On the other hand, it follows from (3.16) and Ito's lemma that

$$\begin{aligned} & e^{-\tilde{\omega}T}\, \tilde{E}\, |\pi(T)|^2_X - \tilde{E}\, |p_0|^2_X \\ & + \int_0^T e^{-\tilde{\omega}t} \tilde{E}\, (2\langle A\pi(t), \pi(t)\rangle + \tilde{\omega}\, |\pi(t)|^2_X - 2(G(\pi(t))\pi(t), \pi(t)) \\ & - |F(\pi(t))\pi(t)|^2_X)\, dt = 0. \end{aligned} \tag{3.32}$$

Note that $|e_\lambda(t) - e^{-\tilde{\omega}t}| \to 0$ as $\lambda \to 0$ uniformly on $[0, T]$. Hence, since from (3.29)–(3.30),

$$\begin{aligned} & \int_0^T e^{-\tilde{\omega}t} \tilde{E}\, (-2\, \langle A\pi(t), \pi(t)\rangle - \tilde{\omega}\, |\pi(t)|^2_X + 2(G(\pi(t))\pi(t), \pi(t)) + |F(\pi(t))\pi(t)|^2_X)\, dt \\ & \qquad + \liminf\ T_\lambda = \liminf\ S_\lambda \geq 0, \end{aligned}$$

it follows from (3.28)–(3.32) that as $\lambda \to 0$,

$$\begin{aligned} & \int_0^T e^{-\tilde{\omega}t}\, \tilde{E}\, (-2\, \langle A\pi(t),\, \pi(t)\rangle - \tilde{\omega}\, |\pi(t)|^2_X + 2(G(\pi(t))\pi(t), \pi(t)) + |F(\pi(t))\pi(t)|^2_X)\, dt \\ & \qquad + \tilde{E}|p_0|^2_X - e^{-\tilde{\omega}T}\ \limsup\ \tilde{E}\, |\pi_\lambda(T)|^2_X \geq 0. \end{aligned} \tag{3.33}$$

Combining (3.32)–(3.33), we obtain

$$e^{-\tilde{\omega}T}\, (\tilde{E}\, |\pi(T)|^2_X - \limsup\ \tilde{E}\, |\pi_\lambda(T)|^2_X) \geq 0.$$

Since $\pi_\lambda(T)$ converges weakly to $\pi(T)$ (without loss of generality), $\tilde{E}\, |\pi(T)|^2_X - \liminf \tilde{E}\, |\pi_\lambda(T)|^2_X \leq 0$ and thus we have

$$\tilde{E}\, |\pi_\lambda(T) - \pi(T)|^2_X \to 0 \quad \text{as } \lambda \to 0.$$

Moreover, $\liminf\ S_\lambda = 0$ and from (3.31)

$$\tilde{E}\, |\pi_\lambda(t) - \pi(t)|_{L^2(0,T;\widetilde{V})} \to 0 \quad \text{as } \lambda \to 0. \qquad \square$$

4. Convergence rate. In this section we establish the convergence estimate (3.12) of π^λ_t to $\pi(t)$ by Corollary 3.5. First, we have the following estimate.

LEMMA 4.1. *For all m and $\lambda = T/m$,*

$$\sup_{1 \leq k \leq m} \tilde{E} : |p^k|^4_X \leq C\, \tilde{E} : |p_0|^4_X.$$

Proof. Note that in (2.15) $\alpha^k = \alpha^k(t_k)$, where

$$\alpha^k(t) = p^{k-1} + \int_{t_{k-1}}^t h\alpha^k(s)\, dy(s) \quad t \in [t_{k-1}, t_k]. \tag{4.1}$$

Multiplying (2.15) by $p_k|p_k|_X^2$, we obtain

(4.2)
$$\frac{1}{4}\left(|p^k|_X^4 - |\alpha^k|_X^4 + (|p_k|_X^2 - |\alpha^k|_X^2)^2\right) + \frac{1}{2}\,|p^k|_X^2|p^k - \alpha^k|_X^2 + \lambda\,\langle Ap^k, p^k\rangle\,|p^k|_X^2 = 0.$$

From (4.1) and Ito's lemma

$$d|\alpha^k(t)|_X^2 = 2\,(h\,\alpha^k(t), \alpha^k(t))_X\,dy(t) + |h\,\alpha^k(t)|_X^2\,dt.$$

Thus, by Ito's lemma,

$$\tilde{E}\,|\alpha^k(t)|_X^4 - \tilde{E}\,|p^{k-1}|_X^4 = \tilde{E}\int_{t_{k-1}}^{t} (4\,|(h\,\alpha^k(s), \alpha^k(s))_X|^2 + 2\,|h\,\alpha^k(s)|_X^2|\alpha^k(s)|_X^2)\,ds$$

$$\leq \int_{t_{k-1}}^{t} 6|h|_\infty^2\,\tilde{E}\,|\alpha^k(s)|_X^4\,ds.$$

By Gronwall's lemma,

$$\tilde{E}\,|\alpha^k|_X^4 \leq e^{6|h|_\infty^2\lambda}\,\tilde{E}\,|p^{k-1}|_X^4.$$

As shown in section 3.2, it follows from (2.7) and (4.3) that

$$e^{-\tilde{\omega}\lambda}\tilde{E}\,|p^k|_X^4 + 2e^{-6\lambda|h|_\infty^2}\beta\,\tilde{E}\,|p^k|_{\widetilde{V}}^2|p^k|_X^2 \leq \tilde{E}\,|p^{k-1}|_X^4,$$

where $\tilde{\omega} = 1 + 4\tilde{\rho} + 6\,|h|_\infty^2$. Hence we obtain

$$\sup_k \tilde{E}\,|p^k|_X^4 + \sum_{k=1}^{m} \tilde{E}\,\beta\,|p^k|_{\widetilde{V}}^2|p^k|_X^2 \leq e^{-\tilde{\omega}T}\,\tilde{E}\,|p_0|_X^4. \qquad \square \tag{4.3}$$

Next, we have the convergence estimate.

THEOREM 4.2. *Assume that the solution $p(t)$ of (1.7) satisfies the following regularity:*

$$\tilde{E}\,|A(hp(t))|_X^2 + \tilde{E}\,|Ap(t)|_X^2 \leq M \tag{4.4}$$

for some $M > 0$ and all $t \in [0, T]$. Then for all m and $\lambda = T/m$,

$$\sup_{1\leq k\leq m} \tilde{E}\,|p^k - p(t_k)|_X^2 + \beta\sum_{k=1}^{m} |p^k - p(t_k)|_{\widetilde{V}}^2 \leq C\,\lambda^2.$$

Proof. For $k \geq 0$ define the approximation error ϵ_k by

$$\epsilon_k = \int_{t_{k-1}}^{t_k} A(p(t) - p(t_k))\,dt - \int_{t_{k-1}}^{t_k} h(p(t) - \bar{\alpha}(t))\,dy(t), \tag{4.5}$$

where $p(t)$ is the unique solution to the Zakai equation (1.7) and $\bar{\alpha}(t)$ satisfies

$$d\bar{\alpha}(t) = h(x)\bar{\alpha}(t)\,dy(t), \quad \bar{\alpha}(t_{k-1}) = p(t_{k-1})$$

on $[t_{k-1}, t_k]$. Since (1.7) is equivalently written as

$$p(t_k) - p(t) + \int_t^{t_k} Ap(s)\, ds = \int_t^{t_k} hp(s)\, dy(s), \tag{4.6}$$

we have $\epsilon_k = \epsilon_k^{(1)} - \epsilon_k^{(2)}$ with

$$\begin{aligned} \epsilon_k^{(1)} &= \int_{t_{k-1}}^{t_k} \int_t^{t_k} A^2 p(s)\, ds\, dt, \\ \epsilon_k^{(2)} &= \int_{t_{k-1}}^{t_k} \int_t^{t_k} A(hp(s))\, dy(s)\, dt + \int_{t_{k-1}}^{t_k} h(p(t) - \bar\alpha(t))\, dy(t). \end{aligned} \tag{4.7}$$

Since from (4.5)–(4.6),

$$p(t_k) - p(t_{k-1}) + Ap(t_k) = \int_{t_{k-1}}^{t_k} h\bar\alpha(t)\, dy(t) - \epsilon_k,$$

the error function $\delta p_k = p^k - p(t_k)$ satisfies

$$\delta p_k - \delta p_{k-1} + \lambda\, A\delta p_k = \int_{t_{k-1}}^{t_k} h\delta\alpha_k(t)\, dy(t) + \epsilon_k, \tag{4.8}$$

where $\delta\alpha_k(t) = \alpha^k(t) - \bar\alpha(t)$ on $[t_{k-1}, t_k)$. Multiplying (4.4) by δp_k, we obtain

$$\begin{aligned} &\frac{1}{2}\left(|\delta p_k|_X^2 - |\delta p_{k-1}|_X^2 + |\delta p_k - \delta p_{k-1}|_X^2\right) + \langle \lambda\, A\delta p_k - \epsilon_k^{(1)},\, \delta p_k\rangle_X \\ &\quad = (\delta p_k, \gamma_k - \epsilon_k^{(2)}), \end{aligned} \tag{4.9}$$

where

$$\gamma_k = \int_{t_{k-1}}^{t_k} h\delta\alpha_k(t)\, dy(t).$$

Since $\tilde E\,(\delta p_{k-1}, \gamma_k - \epsilon_k^{(2)}) = 0$, we have

$$\tilde E\,(\delta p_k, \gamma_k + \epsilon_k^{(2)}) \le \frac{1}{2}\,(\tilde E\,|\delta p_k - \delta p_{k-1}|_X^2 + \tilde E\,|\gamma_k - \epsilon_k^{(2)}|_X^2),$$

and from (4.9) and (2.7),

$$\begin{aligned} &(1 - 2\lambda\tilde\rho)\,\tilde E\,|\delta p_k|_X^2 - \tilde E\,|\delta p_{k-1}|_X^2 \\ &\qquad \le -\lambda\beta\,|\delta p_k|_{\widetilde V}^2 + 2\tilde E\,\langle \epsilon_k^{(1)},\, \delta p_k\rangle_X + \tilde E\,|\gamma_k - \epsilon_k^{(2)}|_X^2 \\ &\qquad \le -\frac{\beta\lambda}{2}\,|\delta p_k|_{\widetilde V}^2 + \frac{2\lambda}{\beta}\left|\frac{\epsilon^{(1)}}{\lambda}\right|_{\widetilde V^*}^2 + 2\,(\tilde E\,|\gamma_k|_X^2 + \tilde E\,|\epsilon_k^{(2)}|_X^2). \end{aligned} \tag{4.10}$$

Note that

$$\tilde E\,|\gamma_k|^2 \le |h|_\infty^2 \int_{t_{k-1}}^{t_k} \tilde E\,|\delta\alpha^k(s)|^2\, ds \le |h|_\infty^2 e^{|h|_\infty^2 \lambda}\lambda\,\tilde E\,|\delta p_{k-1}|^2 \tag{4.11}$$

since $\delta\alpha_k$ satisfies $d\delta\alpha_k(t) = h\,\delta\alpha_k(t)\,dy(t)$ with $\delta\alpha_k(t_{k-1}) = \delta p_{k-1}$. Now we evaluate the error terms $\tilde{E}\,|\epsilon^{(2)}|_X^2$ and $\tilde{E}\,|\epsilon^{(1)}|_{\widetilde{V}^*}^2$. To this end we use

$$\tilde{E}\left|\int_{t_{k-1}}^{t}\eta(s)\,ds\right|^2 \le e\,(t-t_{k-1})\int_{t_{k-1}}^{t}\tilde{E}\,|\eta(s)|^2\,ds \tag{4.12}$$

for any square integrable $\mathcal{F}_t$-adapted process η and $t > t_{k-1}$. In fact if $\phi(t) = \int_{t_{k-1}}^{t}\eta(s)\,ds$, then

$$\tilde{E}|\phi(t)|^2 = 2\int_{t_{k-1}}^{t}\tilde{E}(\phi(s),\eta(s))\,ds \le \int_{t_{k-1}}^{t}\left(\frac{1}{t-t_{k-1}}\tilde{E}\,|\phi(s)|^2 + (t-t_{k-1})\tilde{E}\,|\eta(s)|^2\right)ds.$$

Thus (4.12) follows from Gronwall's lemma. From (4.12),

$$\begin{aligned}\tilde{E}\,|\epsilon^{(1)}|_{\widetilde{V}^*}^2 &\le e\lambda\int_{t_{k-1}}^{t_k}\tilde{E}\left|\int_t^{t_k}Ap(s)\,ds\right|_{\widetilde{V}^*}^2 dt\\ &\le (e\lambda)^2\int_{t_{k-1}}^{t_k}\int_t^{t_k}\tilde{E}\,|Ap(s)|_{\widetilde{V}^*}^2\,ds\,dt \le \frac{1}{2}(e\lambda^2)^2\max_{s\in[0,T]}\tilde{E}\,|Ap(s)|_{\widetilde{V}^*}^2\end{aligned}$$

and

$$\begin{aligned}\tilde{E}\left|\int_{t_{k-1}}^{t_k}\int_t^{t_k}A(hp(s))\,dy(s)\,dt\right|_X^2 &\le e\lambda\int_{t_{k-1}}^{t_k}\tilde{E}\left|\int_t^{t_k}A(hp(s))\,dy(s)\right|_X^2 dt\\ &\le e\lambda\int_{t_{k-1}}^{t_k}\int_t^{t_k}\tilde{E}\,|A(hp(s))|_X^2 \le \frac{1}{2}e\lambda^3\max_{s\in[0,T]}\tilde{E}\,|A(hp(s))|_X^2.\end{aligned}$$

Note that

$$d(p(t)-\bar{\alpha}(t)) = h\,(p(t)-\bar{\alpha}(t))\,dy(t) + Ap(t)\,dt$$

and thus by Ito's lemma and the Hölder inequality,

$$\tilde{E}\,|p(t)-\bar{\alpha}(t)|_X^2 \le \int_{t_{k-1}}^{t}\left(\lambda\,\tilde{E}\,|Ap(t)|_X^2 + \left(\frac{1}{\lambda}+|h|_\infty^2\right)\tilde{E}\,|p(s)-\bar{\alpha}(s)|_X^2\right)ds.$$

By Gronwall's lemma,

$$\tilde{E}\,|p(t)-\bar{\alpha}(t)|_X^2 \le e^{1+|h|_\infty^2\lambda}\,\lambda^2\max_{s\in[0,T]}\tilde{E}\,|Ap(s)|_X^2$$

and thus

$$\begin{aligned}\tilde{E}\left|\int_{t_{k-1}}^{t_k}h(p(t)-\bar{\alpha}(t))\,dy(t)\right|_X^2 &\le |h|_\infty^2\int_{t_{k-1}}^{t_k}\tilde{E}\,|p(t)-\bar{\alpha}(t)|_X^2\\ &\le e^{1+|h|_\infty^2\lambda}\,|h|_\infty^2\lambda^3\max_{s\in[0,T]}\tilde{E}\,|Ap(s)|_X^2.\end{aligned}$$

Hence there exists a constant c_2 such that

$$\frac{2}{\lambda}\,\tilde{E}\,|\epsilon^{(2)}|_X^2 + \frac{1}{\beta}\,\tilde{E}\,\left|\frac{\epsilon^{(1)}}{\lambda}\right|_{\widetilde{V}^*}^2 \le c_2\,\lambda^2 \tag{4.13}$$

provided that (4.4) holds. It thus follows from (4.10)–(4.13) that

$$(1-2\tilde{\rho}\lambda)\,(\tilde{E}\,|\delta p_k|_X^2 + \beta\,\tilde{E}\,|\delta p_k|_{\widetilde{V}}^2) \le (1+c\lambda)\,\tilde{E}\,|\delta p_{k-1}|_X^2 + c_2\,\lambda^3$$

for some positive constants c. Thus we obtain

$$\sup_{1\le k\le m}\ \tilde{E}\,|\delta p_k|_X^2 + \beta\sum_{k=1}^{m}\lambda\tilde{E}\,|\delta p_k|_{\widetilde{V}}^2 \le M\frac{e^{\tilde{c}T}-1}{\tilde{c}}\,\lambda^2$$

for $\tilde{c} = 1 + c + 2\tilde{\rho}$. □

The regularity assumption (4.4) of the solution to (1.7) can be verified under certain smoothness assumptions on the functions f, σ, h, and the initial condition p_0 (e.g., see [20]). Now the estimate (3.12) follows from Corollary 3.5, Lemma 4.1, and Theorem 4.2.

REFERENCES

[1] J. Baras, *Real-time architectures for the Zakai equation and applications*, in Stochastic Analysis, E. Mayer-Wolf et al., eds, Academic Press, Boston, 1991, pp. 15–38.

[2] A. Bensoussan, *Nonlinear filtering theory*, in Progress in Automation and Information Systems, Recent Advances in Stochastic Calculus, J. S. Baras and V. Mirelli, eds., Springer-Verlag, New York, 1990.

[3] A. Bensoussan, R. Glowinski, and A. Rascanu, *Approximation of the Zakai equation by the splitting up method*, SIAM J. Control Optim., 28 (1990), pp. 1420–1431.

[4] J. M. C. Clark, *The design of robust approximation to the stochastic differential equation of nonlinear filtering*, in Communication Systems and Random Process Theory, J. Skwirzynski, ed., Sijthoff and Noordhoff, Alphen aan den Rijn, The Netherlands, 1978, pp. 721–734.

[5] G. B. DiMasi and W. J. Runggaldier, *On measure transformations for combined filtering and parameter estimation in discrete time*, Systems Control Lett., 2 (1982), pp. 57–62.

[6] R. J. Elliot and R. Glowinski, *Approximations to solutions of the Zakai filtering equation*, Stochastic Anal. Appl., 7 (1989), pp. 145–168.

[7] M. Fujisaki, G. Kallianpur, and H. Kunita, *Stochastic differential equations for the nonlinear filtering problem*, Osaka J. Math., 9 (1972), pp. 19–44.

[8] P. Florchinger and F. LeGland, *Time-discretization of the Zakai equation for diffusion processes observed in correlated noise*, in 9th Conference on Analysis and Optimization of Systems, Stochastics, Stochastic Rept., 35 (1991), pp. 233–256.

[9] A. Germani and M. Piccioni, *Finite-dimensional approximations for the equations of nonlinear filtering derived in the mild form*, Appl. Math. Optim., 16 (1987), pp. 51–72.

[10] R. Glowinski and P. Le Tallec, *Augmented Lagrangian and Operator-Splitting Methods in Nonlinear Mechanics*, SIAM Stud. Appl. Math. 9, SIAM, Philadelphia, PA, 1989.

[11] K. Ito, *Approximation of the Zakai equation for nonlinear filtering*, SIAM J. Control Optim., 34 (1996), pp. 620–634.

[12] N. V. Krylov and B. L. Rozovskii, *Stochastic evolution equations*, J. Soviet Math., 16 (1981), pp. 1233–1276.

[13] H. J. Kushner, *Dynamical equations for nonlinear filtering*, J. Differential Equations, 3 (1967), pp. 179–190.

[14] H. J. Kushner, *Probability Methods for Approximations in Stochastic Control and for Elliptic Equations*, Academic Press, New York, 1977.

[15] H. J. Kushner and P. G. Dupuis, *Numerical Methods for Stochastic Control Problems in Continuous Time*, Academic Press, New York, 1977.

[16] S. Lototsky, R. Mikulevicius, and B. L. Rozovskii, *Nonlinear filtering revisited: A spectral approach*, SIAM J. Control Optim., 35 (1997), pp. 435–461.

[17] R. Sh. Lipster and A. N. Shiryayev, *The Statistics of Random Process*, I, II, Springer-Verlag, Berlin, 1977.
[18] G. I. Marchuk, *Splitting and alternating directions methods*, in Handbook of Numerical Analysis 1, P. G. Ciarlet and J.-L. Lions, eds., North-Holland, Amsterdam, 1990, pp. 197–462.
[19] E. Pardoux, *Équations du filtrage non linéaire, de la prédiction et du lissage*, Stochastics, 3 (1979), pp. 127–168.
[20] B. L. Rozovskii, *Stochastic Evolution Systems, Linear Theory and Application to Nonlinear Filtering*, Kluwer Academic Publishers, Dordrecht, The Netherlands, 1991.
[21] B. L. Rozovskii, *A simple proof of uniqueness for Kushner and Zakai equations*, in Stochastic Analyisis, E. Mayer-Wolf et al., eds., Academic Press, Boston, 1991, pp. 449–458.
[22] H. Tanabe, *Equations of Evolution*, Pitman, London, 1979.

SIAM J. CONTROL OPTIM.
Vol. 38, No. 3, pp. 916–946

REGULARITY RESULTS FOR THE MINIMUM TIME FUNCTION OF A CLASS OF SEMILINEAR EVOLUTION EQUATIONS OF PARABOLIC TYPE*

PAOLO ALBANO†, PIERMARCO CANNARSA†, AND CARLO SINESTRARI†

Abstract. Under suitable controllability and smoothness assumptions, the minimum time function $T(x)$ of a semilinear control system is proved to be locally Lipschitz continuous and semiconcave on the controllable set. These properties are then applied to derive optimality conditions relating optimal trajectories to the superdifferential of T.

Key words. time optimal control, semilinear parabolic problem, semiconcavity, optimality conditions

AMS subject classifications. 49L20, 49K20, 93C10

PII. S0363012998335176

1. Introduction.

In this paper we study a time optimal control problem for the system

$$y'(t) = Ay(t) + f(y(t)) + u(t) \quad (t > 0), \quad y(0) = x.$$

Here y belongs to a real separable Banach space X, A is the generator of an analytic semigroup of negative type, f is a Lipschitz perturbation, and $u(\cdot)$ is a control taking values in a closed set $U \subset X$. A model problem for such an equation is a semilinear parabolic system under the action of a distributed control.

The time optimal control problem associated with this system consists of finding, for any initial condition $x \in X$, the trajectories reaching in minimum time a given set $K \subset X$, called the target set. If such a trajectory exists, it is called optimal for the point x.

The history of this problem dates back to the works [17], [18], [20], [21], [5], [16], [22] (see also [24]), where it was studied for linear systems ($f \equiv 0$) using the Pontryagin maximum principle. The linear problem is also considered in [12] from the point of view of dynamic programming. Nonlinear time optimal control problems in infinite dimensions were treated in [6], [7], [14], [15] following the dynamic programming approach. A comprehensive description of many of these results is given in [23].

In this paper we restrict our attention to the case where both the control set U and the target K are closed balls in X. In addition, we require that the control set be large enough to ensure local controllability on ∂K. Our approach can possibly be extended to more general control systems, where K and U are smooth bounded sets with nonempty interior, but we do not treat such cases here.

The main object of our analysis is the minimum time function $T(x)$ of the above system, defined as the infimum of the time taken to steer the trajectories of the state equation from x to K. It is well known that this function provides the basis of the dynamic programming method in optimal control. In section 3 we give the

*Received by the editors March 5, 1998; accepted for publication (in revised form) July 14, 1999; published electronically March 15, 2000.
http://www.siam.org/journals/sicon/38-3/33517.html
†Dipartimento di Matematica, Università di Roma "Tor Vergata", Via della Ricerca Scientifica, 00133 Roma, Italy (albano@mat.uniroma2.it, cannarsa@mat.uniroma2.it, sinestra@mat.uniroma2.it).

precise definition and state some basic properties of $T(x)$. In section 4 we analyze the regularity of the minimum time function. Our main result (Theorem 4.3) is that T satisfies a semiconcavity estimate of the form

$$T(x) + T(y) - 2T\left(\frac{x+y}{2}\right) \leq C|x-y|^{1+\alpha}$$

for some $C \geq 0$, $\alpha \in]0,1]$. This result holds under suitable regularity assumptions on the norm of the space X and on the nonlinear term f which appears in the state equation. We have to assume, in fact, that the norm of X is sufficiently smooth away from 0 and that, roughly speaking, f is of class $C^{1,\alpha}$ as a function from Y to X, where $Y \subset X$ is the domain of a suitable fractional power of $-A$. We could just assume $f \in C^{1,\alpha}(X,X)$, but this would be too restrictive for the application to the model problem mentioned at the beginning. A similar semiconcavity result was obtained by two of the authors in [11] for a finite dimensional setting; the main ideas in the proof are similar, but the extension of the technique of [11] to infinite dimensions requires a substantial improvement of the method.

Once we have proved the semiconcavity of T, we are in a position to apply the regularity results on semiconcave functions proved in [2]. These results are recalled in section 2 of this paper and concern the set of nondifferentiability (or singular set) of a semiconcave function. In particular, the singular set can be covered by a countable family of Lipschitz surfaces in X. Moreover, it is possible to give conditions for the propagation of singularities starting from a given point of nondifferentiability. In the case of the minimum time function, these results become interesting in connection with the optimality conditions that we derive in section 5 of this paper. Our main result (see Corollary 5.12 and Theorem 5.13) states that there is a one-to-one correspondence between the optimal trajectories starting at a given point x and the elements of a suitable generalized differential of T at x. In particular, it follows that the differentiability points of T are exactly the starting points of a unique time optimal trajectory. Therefore, the aforementioned result on the singular sets of T implies that the initial conditions of multiple time optimal trajectories form a "small" set in X. On the other hand, this set is in general nonempty even for a system associated with a parabolic equation (see [1]).

2. Preliminaries. Throughout this paper we denote by X a *real separable* Banach space with *separable dual* X^*, and by

$$E = \{e_j\}_{j\in\mathbb{N}}$$

a dense subset of X. Then, X is an Asplund space and, in particular, X possesses an equivalent norm that is Fréchet differentiable on $X \setminus \{0\}$. We denote by $|\cdot|$ such a norm, by $\langle\cdot,\cdot\rangle$ the duality pairing between X and X^*, and by $\|\cdot\|$ the standard norm of a linear operator between Banach spaces. For any $R > 0$ and $x \in X$, we set

$$B_R(x) = \{y \in X : |x-y| < R\},$$

and we abbreviate $B_R = B_R(0)$.

Let $\Omega \subset X$ be an open set and W another Banach space. We denote by $C(\Omega;W)$ the space of all continuous functions $g : \Omega \to W$, and by $C^1(\Omega;W)$ the space of all functions that are Fréchet differentiable in Ω with continuous derivative Dg. Furthermore, for any $\theta \in [0,1]$, $C^{1,\theta}(\Omega;W)$ stands for the set of all functions $g \in C^1(\Omega;W)$ with Hölder continuous derivative; that is

$$\|Dg(x) - Dg(z)\| \leq C|x-z|^\theta \quad \forall x,z \in \Omega$$

for some constant $C > 0$. Finally, in the above notation, we drop the arrival set when $W = \mathbb{R}$.

We now recall the definition of a semiconcave function in Ω, a notion that is classical in finite dimensions and that was applied to infinite dimensional problems in [26], [8].

DEFINITION 2.1. *Given $\alpha \in]0,1]$, a function $g \in C(\Omega)$ is said to be* semiconcave *with exponent α if, for any point $x_0 \in \Omega$, there exist $\rho > 0$ and $C \geq 0$ such that*

$$g(x) + g(y) - 2g\left(\frac{x+y}{2}\right) \leq C|x-y|^{1+\alpha}$$

for all $x, y \in B_\rho(x_0)$. We denote by $SC^\alpha(\Omega)$ the set of all functions $g : \Omega \to \mathbb{R}$ which are semiconcave with exponent α and we set $SC(\Omega) = \cup_{\alpha \in]0,1]} SC^\alpha(\Omega)$.

Semiconcave functions share many properties with concave functions. For instance, arguing as in the finite dimensional case (see e.g., [3]), it is easy to show that any semiconcave function in Ω is locally Lipschitz continuous. Consequently, g is Fréchet differentiable on a dense set by a result of [28], and the gradient Dg is locally bounded. Now, let us denote by $D_*g(x)$ the set of all points $p \in X^*$ for which a sequence $\{x_k\}_{k\in\mathbb{N}} \subset \Omega$ exists such that

(i) $x = \lim_{k\to\infty} x_k$,

(ii) g is Fréchet differentiable at x_k,

(iii) $Dg(x_k)$ weakly–* converges to p as $k \to \infty$.

In view of the above remarks we have that $D_*g(x) \neq \emptyset$ for any $x \in \Omega$.

Just like the concave case, semiconcave functions possess a natural notion of generalized gradient, given by the *superdifferential*

$$D^+g(x) = \left\{ p \in X^* \,:\, \limsup_{y\to x} \frac{g(y) - g(x) - \langle p, y-x\rangle}{|y-x|} \leq 0 \right\} \quad \forall x \in \Omega.$$

Actually, a similar generalization of the gradient is the *subdifferential* of g defined as

$$D^-g(x) = \left\{ p \in X^* \,:\, \liminf_{y\to x} \frac{g(y) - g(x) - \langle p, y-x\rangle}{|y-x|} \geq 0 \right\}.$$

However, for a semiconcave function, the superdifferential plays a more important role than the subdifferential. Indeed, in view of the proposition below, D^+g is nonempty at every point. Therefore, given any $x \in \Omega$, either $D^-g(x)$ is empty, or g is differentiable at x.

PROPOSITION 2.2. *Let $g \in SC^\alpha(\Omega)$ for some $\alpha \in]0,1]$ and let $x_0 \in \Omega$. Then $D^+g(x_0)$ is nonempty and satisfies*

$$D^+g(x_0) = \overline{\text{co}}\, D_*g(x_0), \tag{2.1}$$

where $\overline{\text{co}}$ denotes the closed convex hull. Moreover, there exist $R, C > 0$ such that

$$g(y) - g(x) - \langle p, y-x\rangle \leq C|y-x|^{1+\alpha} \quad \forall\, y, x \in B_R(x_0),\ p \in D^+g(x), \tag{2.2}$$

$$\langle p-q, y-x\rangle \leq 2C|y-x|^{1+\alpha} \quad \forall\, y, x \in B_R(x_0),\ p \in D^+g(y),\ q \in D^+g(x). \tag{2.3}$$

For the proof of the above proposition the reader is referred to [8] for semiconcave functions with exponent $\alpha = 1$, the argument in the general case being similar.

We now introduce the singular sets of a semiconcave function on X. Let P be a subspace of X^* and $r > 0$. We denote by B_r^P the ball with radius r and center at 0 in P. We say that B is a *ball of dimension* $n \in \mathbb{N}$ in X^* if there exist a subspace $P \subset X^*$, with $\dim P = n$, a covector $p \in X^*$, and a radius $r > 0$ such that

$$B = p + B_r^P. \tag{2.4}$$

Similarly, we say that B is a *ball of codimension* n in X^* if (2.4) holds for a subspace P such that $\operatorname{codim} P = n$. Let $g : \Omega \to \mathbb{R}$ be a semiconcave function. We introduce the following singular sets of g.

DEFINITION 2.3. *The set of all points $x \in \Omega$ such that $D^+g(x)$ fails to be a singleton is called the* singular set *of g and is denoted by $\Sigma(g)$. The points of $\Sigma(g)$ are called the* singular points *of g. Moreover, for any $n \in \mathbb{N}$, $n > 0$ we denote by $\Sigma_n(g)$ (resp., $\Sigma_{\infty-n}(g)$) the set of all points $x \in \Omega$ such that $D^+g(x)$ contains a ball of dimension n (resp., codimension n).*

To study the singular set we will need the following definition (see e.g., [19]).

DEFINITION 2.4. *We will say that a set $\Sigma \subset X$ is n* rectifiable *iff a bounded set $A \subset \mathbb{R}^n$ and a Lipschitz function, f, exist so that*

$$f : A \to \Sigma$$

is surjective. Moreover, we will say that a set $\Sigma \subset X$ is countably n rectifiable *iff*

$$\Sigma = \bigcup_{j \in \mathbb{N}} Q_j \tag{2.5}$$

where the sets $Q_j \subset X$, $j \in \mathbb{N}$, are n rectifiable.

By analogy with the previous definition, we will say that $\Sigma \subset X$ is $\infty - n$ rectifiable if a subspace $Y \subset X$, of codimension n, and a bounded set $A \subset Y$ exist so that Σ is the image of A under a Lipschitz map f. Similarly, we will say that Σ is countably $\infty - n$ rectifiable if it can be represented as in (2.5) for some family of $\infty - n$ rectifiable sets $\{Q_j\}$. The following theorem is proved in [2].

THEOREM 2.5. *Let $g \in SC(\Omega)$ and fix $n \in \mathbb{N}$, $n > 0$. Then,*

(i) *$\Sigma_n(g)$ is countably $\infty - n$ rectifiable;*

(ii) *$\Sigma_{\infty-n}(g)$ is countably n rectifiable.*

The following theorem, again proved in [2], describes the propagation of singularities of a semiconcave function with exponent 1.

THEOREM 2.6. *Let X be a Hilbert space, let $g \in SC^1(\Omega)$, and let $x_0 \in \Sigma(g)$. Let*

$$p_0 \in D^+g(x_0) \setminus D_*g(x_0)\,, \tag{2.6}$$

and suppose that, for some vector $q \in X \setminus \{0\}$ and some $T_0 > 0$,

$$p_0 + tq \notin D^+g(x_0) \quad \forall 0 < t \le T_0\,. \tag{2.7}$$

Then, a number $0 < T \le T_0$ and two Lipschitz arcs $x, p : [0, T] \to X$ exist so that

$$\text{(i)} \quad \langle x(t) - x_0, q\rangle < 0 \ \ \forall 0 < t \le T,$$

$$\text{(ii)} \quad (x(0), p(0)) = (x_0, p_0),$$

$$\text{(iii)} \quad (x(t), p(t)) \in \Sigma(g) \times D^+g(x(t)) \quad \forall t \in [0, T]\,.$$

A point $x_0 \in \Sigma(g)$ for which x and p as above exist will be called a *propagation point* of $\Sigma(g)$.

3. A time optimal control problem. Let us consider the controlled evolution equation in X

$$\begin{cases} y'(t) = Ay(t) + f(y(t)) + u(t), & t > 0, \\ y(0) = x, \end{cases} \tag{3.1}$$

under the following assumptions:

(H1) $A : D(A) \subset X \to X$ is the infinitesimal generator of an analytic semigroup of bounded linear operators on X, e^{tA}, satisfying

$$\|e^{tA}\| \le e^{-\omega t} \quad \forall t \ge 0 \tag{3.2}$$

for some constant $\omega > 0$;

(H2) $f : X \to X$ is a Lipschitz continuous function satisfying

$$|f(x) - f(y)| \le L|x - y| \quad \forall x, y \in X,$$
$$f(0) = 0.$$

We assume that the *control set* U is a closed ball in X of the form

$$U = \overline{B}_r \tag{3.3}$$

for some given $r > 0$. A measurable function $u : [0, +\infty[\to U$ will be called a *control strategy* (or, simply a *control*). We denote by $\mathcal{U}$ the set of all control strategies. It is well known that, for any $x \in X$ and any control u, state equation (3.1) has a unique mild solution. We recall that a mild solution of (3.1) is a function $y \in C([0, \infty[; X)$ satisfying

$$y(t) = e^{tA}x + \int_0^t e^{(t-s)A}[f(y(s)) + u(s)]ds\,, \quad t \ge 0. \tag{3.4}$$

Such a solution is called the *trajectory* of system (3.1) starting from x with control u and is denoted by $y(\cdot; x, u)$.

Remark 3.1. The assumption that ω in (H1) is positive is not restrictive, as one can always reduce the system to this case by adding a bounded linear term to f.

Remark 3.2. Assumption (H1) implies (see e.g., [27]) that the *fractional powers* of $-A$, denoted by $(-A)^\theta$, are well defined for any $\theta \in [0, 1]$ and satisfy

$$|(-A)^\theta e^{tA}x| \le \frac{M_\theta}{t^\theta}|x| \quad \forall x \in X, t > 0, \tag{3.5}$$

$$|x| \le M'_\theta|(-A)^\theta x| \quad \forall x \in D((-A)^\theta) \tag{3.6}$$

for suitable constants $M_\theta, M'_\theta > 0$. Moreover, a well-known interpolation inequality (see e.g., [27]) ensures that, for any $\theta \in [0, \frac{1}{2}]$, there exists $\kappa_\theta > 0$ such that

$$|(-A)^\theta x| \le |(-A)^{\frac{1}{2}}x| + \kappa_\theta|x| \qquad \forall x \in D((-A)^{\frac{1}{2}})\,. \tag{3.7}$$

We are interested in the time optimal control problem for system (3.1) with a closed ball $\overline{B}_R$ as a target. More precisely, for any $x \in X$ and any control strategy u we denote by

$$\tau(x, u) = \min\{t \ge 0 : y(t; x, u) \in \overline{B}_R\} \in [0, +\infty] \tag{3.8}$$

the transition time from x to $\overline{B}_R$. We define the *controllable set* $\mathcal{R}$ to be the set of all points x such that $\tau(x,u) < +\infty$ for some u. Then, for all $x \in \mathcal{R}$, the time optimal control problem consists of minimizing $\tau(x,u)$ over all controls u. A control u at which $\tau(x,\cdot)$ attains the minimum is said to be optimal for x and the corresponding solution $y(\cdot;x,u)$ of (3.1) is called an optimal trajectory.

The dynamic programming approach to the above problem is based on the properties of the *value function* — called minimum time function in this case — defined as

$$T : \mathcal{R} \to [0,+\infty[\,, \quad T(x) = \inf_{u\in\mathcal{U}} \tau(x,u). \tag{3.9}$$

The minimum time function encompasses most of the information about the control problem. An important property of this function is the dynamic programming principle which says that, for any $x \in \mathcal{R}$ and any control u,

$$T(x) \le t + T(y(t;x,u)) \qquad \forall t \in [0,\tau(x,u)]\,. \tag{3.10}$$

Furthermore, equality holds in (3.10) iff u is optimal for x.

Our goal is to relate the optimal trajectories of the above problem to the structure of superdifferentials of T. Such generalized differentials enjoy the properties that we have described in the previous section, once it is shown that T is semiconcave. Therefore, one of the main steps of our analysis will be to prove the semiconcavity of T, which we do in the next section. For this purpose we impose, in addition to (H1) and (H2), the conditions below.

(H3) The constants ω, L, r, R which appear in (H1), (H2), (3.3), and (3.8), resp., satisfy

$$0 < (L-\omega)R < r.$$

(H4) There exists $\alpha \in]0,1]$ such that the norm in X is of class $C^{1,\alpha}(X \setminus \overline{B}_M)$, for any $M > 0$.

(H5) The function f is Gâteaux differentiable at all points $x \in X$. The Gâteaux differential $\delta f(x)$ satisfies the following property: there exist $\theta_1 \in]0,1/2[$, $\beta \in]0,1]$, and $\hat{C} > 0$ such that

$$|f(x+y) - f(x) - \delta f(x)y| \le \hat{C}|(-A)^{\theta_1} y|^{1+\beta} \tag{3.11}$$

for all $x, y \in D((-A)^{\theta_1})$. Moreover, the map $x \mapsto \delta f(x)$ is strongly continuous on X; that is, for any sequence $\{x_n\}$ in X,

$$\lim_{n\to\infty} x_n = x_\infty \quad \Rightarrow \quad \lim_{n\to\infty} \delta f(x_n)x = \delta f(x_\infty)x \ \ \forall x \in X. \tag{3.12}$$

Remark 3.3. The rightmost inequality in (H3) is a controllability assumption and will ensure that $\mathcal{R}$ is an open neighbourhood of $\overline{B}_R$. The first inequality holds provided $L > \omega$ and has been assumed only in order to simplify the exposition. In fact, all the results of this paper hold in the case of $L \le \omega$ as well, and some of them even in a stronger form. For instance, if $L \le \omega$, one could prove that $\mathcal{R} = X$.

Remark 3.4. Let us denote by $d(x)$ the distance of x from the target

$$d(x) = \inf_{y\in\overline{B}_R} |y - x| = (|x| - R)^+\,.$$

Then, from (H4) and the identity

$$|x+h|+|x-h|-2|x|=\int_0^1 \langle D|x+th|-D|x-th|,h\rangle dt,$$

it follows that d is semiconcave on $X\setminus \overline{B}_R$ with exponent α. In addition, the numbers C,ρ in Definition 2.1 can be chosen independently on $x_0\in X\setminus \overline{B}_R$.

Condition (H4) holds true in several important examples. For instance, let $X=L^p(\mathcal{O})$ where $p\in]1,\infty[$ and $\mathcal{O}$ is a bounded domain $\mathbb{R}^n$. Then, it is well known that the norm satisfies (H4) with $\alpha=p-1$ if $p\in]1,2[$ and with $\alpha=1$ if $p\in[2,\infty[$.

Remark 3.5. Hypothesis (H5) is trivially satisfied when $f\in C^{1,\beta}(X;X)$, since we can take in this case $\delta f=Df$ and any $\theta_1\in]0,1/2[$. However, there are interesting examples where f satisfies our assumptions without being Fréchet differentiable, as shown in the next example. Let us also observe that the Lipschitz continuity of f and property (3.11) imply that, for some constant $L^*>0$,

$$|f(x)+f(y)-2f(z)|\le L^*(|(-A)^{\theta_1}(x-y)|^{1+\beta}+|x+y-2z|) \tag{3.13}$$

for all $x,y,z\in D((-A)^{\theta_1})$.

Example. Let $\mathcal{O}$ be a bounded domain in $\mathbb{R}^n$ with sufficiently smooth boundary, and let $\phi\in \mathrm{Lip}(\mathbb{R})\cap C^{1,1}(\mathbb{R})$ be such that $\phi(0)=0$. In this example we show that the control system associated with the *parabolic* state equation

$$\begin{cases} y_t(t,\xi)=\Delta y(t,\xi)+\phi(y(t,\xi))+u(t,\xi), & (t,\xi)\in]0,+\infty[\times\mathcal{O}\,,\\ y(t,\xi)=0, & (t,\xi)\in]0,+\infty[\times\partial\mathcal{O}\,,\\ y(0,\xi)=x(\xi), & \xi\in\mathcal{O}, \end{cases} \tag{3.14}$$

rewritten in abstract form, satisfies (H1), (H2), and (H5).

Let us set $X=L^2(\mathcal{O})$, and define $A:D(A)\subset X\to X$ as

$$D(A)=W^{2,2}(\mathcal{O})\cap W_0^{1,2}(\mathcal{O})\,;\qquad Ax(\xi)=\Delta x(\xi)\quad(\xi\in\mathcal{O},x\in D(A)).$$

Moreover, consider the composition map $f:X\to X$ defined by

$$f(x)(\xi)=\phi(x(\xi)),\quad(\xi\in\mathcal{O},x\in X).$$

Then, (3.1) is the abstract version of system (3.14). The fact that A satisfies (H1) is well known; (H2) is immediate to check. Let us show that (H5) is also satisfied. It is well known (see e.g., [4]) that f is Gâteaux differentiable with

$$\delta f(y)(z)(\xi)=\phi'(y(\xi))z(\xi)\text{ for any }y,z\in X\,,\ \xi\in\mathcal{O},$$

but f is nowhere Fréchet differentiable unless ϕ is linear. On the other hand, if L and M are the Lipschitz constants of ϕ and ϕ', then, for any $\beta\in[0,1]$, we have that

$$\begin{aligned} &||f(x+y)-f(x)-\delta f(x)y||^2_{L^2(\mathcal{O})}\\ &=\int_\Omega y^2(\xi)\left(\int_0^1(\phi'(x(\xi)+\lambda y(\xi))-\phi'(x(\xi)))\,d\lambda\right)^2 d\xi\\ &\le\int_\Omega y^2(\xi)\left(\int_0^1(|\phi'(x(\xi)+\lambda y(\xi))|+|\phi'(x(\xi))|)\,d\lambda\right)^{2-2\beta}\left(\int_0^1 M\lambda|y(\xi)|d\lambda\right)^{2\beta}d\xi\\ &\le(2L)^{2-2\beta}(M/2)^{2\beta}\int_\Omega|y(\xi)|^{2+2\beta}d\xi=C||y||^{2(1+\beta)}_{L^{2+2\beta}(\mathcal{O})}. \end{aligned}$$

Recalling that, by the Sobolev embedding theorem,

$$D((-A)^\theta) \subset W^{2\theta,2}(\mathcal{O}) \subset L^{2n/(n-4\theta)}(\mathcal{O}),$$

we obtain property (3.11) with

$$\beta = 1\,, \quad \theta_1 = \frac{3}{8} \quad \text{if} \quad n \le 3,$$

and with

$$\beta \in \left]0, \frac{1}{n-1}\right[\,, \quad \theta_1 = \frac{1}{4} \quad \text{if} \quad n > 3.$$

Finally, property (3.12) follows from the Lebesgue dominated convergence theorem.

In the next lemma we collect some basic properties about the trajectories of (3.1).

LEMMA 3.6. *Assume* (H1) *and* (H2). *Then the following properties hold.*

(i) *For any $x \in X$ a control strategy u^* exists such that the corresponding trajectory of* (3.1) *satisfies*

$$|y(t;x,u^*)| \le e^{(L-\omega)t}\left(|x| - \frac{r}{L-\omega}\right) + \frac{r}{L-\omega} \tag{3.15}$$

for any $0 \le t \le \omega^{-1}\log(1 + r^{-1}\omega|x|)$.

(ii) *For any $x \in X$ and any control strategy u the corresponding trajectory of* (3.1) *satisfies*

$$|y(t;x,u)| \le e^{(L-\omega)t}\left(|x| + \frac{r}{L-\omega}\right) - \frac{r}{L-\omega} \quad \forall t \ge 0\,. \tag{3.16}$$

(iii) *For any $x, z \in X$ and any control strategy u we have that*

$$|y(t;x,u) - y(t;z,u)| \le e^{(L-\omega)t}|x-z| \quad \forall t \ge 0. \tag{3.17}$$

(iv) *Let $\theta \in [0,1[$ be given. Then the trajectories of* (3.1) *belong to the domain of $(-A)^\theta$ for any $t > 0$, and*

$$|(-A)^\theta y(t;x,u)| \le \frac{M_\theta}{t^\theta}|x| + \int_0^t \frac{M_\theta}{(t-s)^\theta}(L|y(s;x,u)| + r)ds. \tag{3.18}$$

Moreover, for any $T_0 > 0$, a constant $K = K(\theta, T_0) > 0$ exists such that

$$|(-A)^\theta(y(t;x,u) - y(t;z,u))| \le \frac{K}{t^\theta}|x-z| \tag{3.19}$$

for any $x, z \in X$ and $t \in [0, T_0]$.

(v) *If in addition* (H5) *holds, then, for any $T_0 > 0$, a constant $k > 0$ exists such that*

$$|y(t;x+h,u) + y(t;x-h,u) - 2y(t;x,u)| \le k|h|^{1+\beta} \tag{3.20}$$

for every $x, h \in X$ and $t \in [0, T_0]$.

Proof. (i): Given $x \in X$, $x \neq 0$, let us set, for $t \geq 0$,

$$u^*(t) = -\frac{r}{|x|} e^{t(A+\omega I)} x.$$

By (3.2), $|u^*(t)| \leq r$ for any t, and thus u^* is an admissible control strategy. From (3.4) we compute

$$y(t) = \left(1 - \frac{r}{|x|} \frac{e^{\omega t} - 1}{\omega}\right) e^{tA} x + \int_0^t e^{(t-s)A} f(y(s))\, ds.$$

Let us set

$$\phi(t) = e^{\omega t} |y(t)|\,.$$

Then, for any $t \leq \omega^{-1} \log(1 + r^{-1}\omega|x|)$,

$$\begin{aligned} \phi(t) &\leq \left[|x| - \frac{r}{\omega}\left(e^{\omega t} - 1\right)\right] + \int_0^t e^{\omega s} |f(y(s))|\, ds \\ &\leq |x| + \int_0^t \left[L\phi(s) - r e^{\omega s}\right] ds. \end{aligned}$$

Here we have used the Lipschitz assumption (H2) on f. We can now apply the Gronwall inequality to obtain

$$\phi(t) \leq e^{Lt}|x| + \frac{r}{L - \omega}\left(e^{\omega t} - e^{Lt}\right),$$

which implies (3.15).
(ii)–(iii): These assertions can be derived by similar computations to those of (i).
(iv): Inequality (3.18) is a straightforward consequence of (3.4) and (3.5). In order to prove (3.19), we apply properties (3.5) and (3.17) to obtain

$$\begin{aligned} |(-A)^\theta (y(t;x,u) - y(t;z,u))| &\leq \frac{M_\theta}{t^\theta}|x - z| + \int_0^t \frac{M_\theta}{(t-s)^\theta} |y(s;x,u) - y(s;z,u)| ds \\ &\leq \frac{M_\theta}{t^\theta}|x - z| + M_\theta |x - z| \int_0^t \frac{e^{(L-\omega)s}}{(t-s)^\theta} ds \\ &\leq M_\theta |x - z| \left[\frac{1}{t^\theta} + e^{(L-\omega)T_0} \frac{T_0^{1-\theta}}{1-\theta}\right]. \end{aligned}$$

(v): By (3.2), (3.13), and (3.19), we obtain

$$\begin{aligned} &|y(t;x+h,u) + y(t;x-h,u) - 2y(t;x,u)| \\ &= \left| \int_0^t e^{(t-s)A} (f(y(s;x+h,u)) + f(y(s;x-h,u)) - 2f(y(s;x,u)))ds \right| \\ &\leq L^* \int_0^t e^{-\omega(t-s)} \{|y(s;x+h,u) + y(s;x-h,u) - 2y(s;x,u)| \\ &\qquad + |(-A)^{\theta_1}(y(s;x+h,u) - y(s;x-h,u))|^{1+\beta}\} ds \end{aligned}$$

$$\leq L^* \int_0^t e^{-\omega(t-s)} |y(s;x+h,u) + y(s;x-h,u) - 2y(s;x,u)| ds$$

$$+L^* \int_0^t e^{-\omega(t-s)} \frac{K(\theta_1, T_0)^{1+\beta}}{s^{(1+\beta)\theta_1}} |h|^{1+\beta} ds$$

$$\leq L^* \int_0^t e^{-\omega(t-s)} |y(s;x+h,u) + y(s;x-h,u) - 2y(s;x,u)| ds + K'|h|^{1+\beta}$$

for some $K' > 0$. We conclude our proof by the Gronwall inequality. ∎

Remark 3.7. From (ii) above and (H2) we conclude that, for any $\rho > 0$, a constant $C_\rho > 0$ exists such that

$$|f(y(t;x,u))| \leq C_\rho \quad \forall t \in [0,1]$$

for any $|x| \leq \rho$ and any control strategy u. Therefore, all trajectories starting from $x \in B_\rho$ may be estimated from below as follows:

$$|y(t;x,u)| \geq |e^{tA}x| - (C_\rho + r)t \quad \forall t \in [0,1]\,.$$

The last inequality, together with the strong continuity of e^{tA}, implies that $T(x) > 0$ for any $x \in \mathcal{R} \setminus \overline{B}_R$.

The following proposition shows that, in a neighbourhood of the target, the minimum time function is bounded above by the distance function.

PROPOSITION 3.8. *Assume* (H1), (H2), *and* (H3). *Then, there exists a number* $\rho \in]0, r/(L-\omega) - R[$ *such that*

$$T(x) \leq \frac{d(x)}{r - (R+\rho)(L-\omega)} \tag{3.21}$$

for all points x *satisfying* $d(x) \leq \rho$. *Moreover,* $\mathcal{R}$ *is an open subset of* X.

Proof. First of all, we fix $\rho \in]0, r/(L-\omega) - R[$ so that

$$\frac{1}{L-\omega} \log \frac{\frac{r}{L-\omega} - R}{\frac{r}{L-\omega} - |x|} \leq \omega^{-1} \log(1 + r^{-1}\omega|x|) \tag{3.22}$$

for any $x \in X$ such that $0 < d(x) \leq \rho$. The existence of such a number ρ follows from the fact that, taking the limit as $|x| \to R$, the left-hand side of (3.22) tends to 0 while the right-hand side tends to $\omega^{-1} \log(1 + r^{-1}\omega R) > 0$.

Now, let $x \in X$ be fixed so that $0 < d(x) \leq \rho$, and let u^* be the corresponding control strategy, given by Lemma 3.6(i). Defining

$$t^* = \frac{1}{L-\omega} \log \frac{\frac{r}{L-\omega} - R}{\frac{r}{L-\omega} - |x|}\,,$$

we can apply (3.15) with $t = t^*$ in light of (3.22), obtaining $|y(t^*;x,u^*)| \leq R$. Therefore, by an easy computation,

$$T(x) \leq t^* \leq \frac{d(x)}{r - |x|(L-\omega)}.$$

This estimate, together with (3.17), easily implies that $\mathcal{R}$ is open. ∎

In section 6 we will use the following proposition. For a proof of this standard result see, e.g., [23, p. 68].

PROPOSITION 3.9. *Assume* (H1), (H2), (H5)*, and let y be a trajectory of* (3.1)*. Then, there exists a unique strongly continuous map $G : \Delta \to \mathcal{L}(X)$, where $\Delta := \{(t,s) : 0 \le s \le t \le T\}$, such that*

$$\begin{cases} G(t,t) = I & \forall t \in [0,T] \\ G(t,r)G(r,s) = G(t,s) & \forall r,s,t : 0 \le s \le r \le t \le T, \end{cases} \tag{3.23}$$

and

$$\begin{aligned} G(t,s)x &= e^{(t-s)A}x + \int_s^t e^{(t-r)A}\delta f(y(r))G(r,s)x\,dr \\ &= e^{(t-s)A}x + \int_s^t G(t,r)\delta f(y(r))e^{(r-s)A}x\,dr \end{aligned}$$

for $0 \le s \le t \le T$, $x \in X$.

Remark 3.10. By standard techniques one can show that the operator G above is the evolution operator of the linearization of (3.1) along the trajectory y. Equivalently, if $y(\cdot) = y(\cdot; x, u)$, then, for any $h \in X$, we have that

$$y(t; x+h, u) - y(t; x, u) = G(t,0)h + o(|h|)\,. \tag{3.24}$$

4. Regularity of the minimum time function. In this section we prove that the minimum time function T, defined in (3.9), is semiconcave. We begin showing a preliminary regularity result, which yields, in particular, the local Lipschitz continuity of T.

THEOREM 4.1. *Assume* (H1), (H2), (H3)*, and let $x_0 \in \mathcal{R} \setminus \overline{B}_R$ be fixed. Then, for some $\delta > 0$,*

$$0 < m(\delta) := \inf_{x \in B_\delta(x_0)} T(x) \le \sup_{x \in B_\delta(x_0)} T(x) =: M(\delta) < +\infty\,. \tag{4.1}$$

Moreover, for any $\theta \in [0,1[$ a constant $K_\theta = K_\theta(x_0)$ exists such that

$$|T(x) - T(z)| \le K_\theta |(-A)^{-\theta}(x-z)| \qquad \forall x, z \in B_\delta(x_0). \tag{4.2}$$

Proof. We first prove the rightmost inequality of (4.1); that is, we show that T is bounded from above in some neighborhood of x_0. Let u_0 be a control such that $\tau_0 := \tau(x_0, u_0) < \infty$. Recalling (3.17), we have that, for any $x \in X$,

$$d(y(\tau_0; x, u_0)) \le |y(\tau_0; x, u_0) - y(\tau_0; x_0, u_0)| \le e^{(L-\omega)\tau_0}|x - x_0|\,.$$

Now, let

$$\delta_0 = \rho e^{(\omega - L)\tau_0},$$

where $\rho \in]0, r/(L-\omega) - R[$ is given by Proposition 3.8. Then, from the dynamic programming principle and (3.21) we obtain

$$T(x) \le \tau_0 + T(y(\tau_0; x, u_0)) < \tau_0 + \frac{\rho}{r - (R+\rho)(L-\omega)} \qquad \forall x \in B_{\delta_0}(x_0),$$

and this shows that $M(\delta_0) < +\infty$. We now show that there exist numbers $\delta_1 \in]0, \delta_0]$ and $C > 0$ such that

$$|T(x) - T(z)| \le C|x - z| \qquad \forall x, z \in B_{\delta_1}(x_0)\,. \tag{4.3}$$

Let us define

$$\delta_1 = \min\left\{\frac{\rho}{2}e^{(\omega - L)M(\delta_0)}, \delta_0\right\}$$

and consider two points $x, z \in B_{\delta_1}(x_0)$ with $T(x) < T(z)$. Then, for any $\varepsilon \in]0, T(z) - T(x)[$ a control u_ε exists such that

$$\tau(x, u_\varepsilon) < T(x) + \varepsilon < T(z) \le M(\delta_0).$$

Let us set for simplicity $x' = y(\tau(x, u_\varepsilon); x, u_\varepsilon)$, $z' = y(\tau(x, u_\varepsilon); z, u_\varepsilon)$. From the dynamic programming principle it follows that

$$|T(z) - T(x)| = T(z) - T(x) \le T(z') + \varepsilon\,. \tag{4.4}$$

Now, estimate (3.17) gives

$$d(z') \le |x' - z'| \le e^{(L-\omega)M(\delta_0)}|x - z| < \rho.$$

Then we can apply Proposition 3.8 to obtain

$$\begin{aligned} |T(z) - T(x)| &\le T(z') + \varepsilon \le \frac{|x' - z'|}{r - (R + \rho)(L - \omega)} + \varepsilon \\ &\le \frac{e^{(L-\omega)M(\delta_0)}}{r - (R + \rho)(L - \omega)}|x - z| + \varepsilon\,. \end{aligned} \tag{4.5}$$

Since ε is arbitrary, (4.3) follows. We note that (4.3) is a special case of (4.2), namely (4.2) for $\theta = 0$. Recalling Remark 3.7, such a Lipschitz estimate yields the lower bound in (4.1) for a suitably small $\delta \in]0, \delta_1[$.

In order to complete the proof it remains to prove (4.2) for $\theta \in]0, 1[$. For this purpose, we note that, taking x, z, and u_ε as above,

$$\begin{aligned} &|y(t; x, u_\varepsilon) - y(t; z, u_\varepsilon)| \\ &\le \int_0^t e^{-\omega(t-s)}|f(y(s; x, u_\varepsilon)) - f(y(s; z, u_\varepsilon))|ds + |e^{tA}(x - z)| \\ &\le L\int_0^t e^{-\omega(t-s)}|y(s; x, u_\varepsilon) - y(s; z, u_\varepsilon)|ds + \frac{M_\theta}{t^\theta}|(-A)^{-\theta}(x - z)|\,. \end{aligned}$$

Then, applying the Gronwall inequality, we obtain

$$\begin{aligned} |y(t; x, u_\varepsilon) - y(t; z, u_\varepsilon)| &\le \left[\frac{M_\theta}{t^\theta} + LM_\theta\int_0^t \frac{e^{(L-\omega)(t-s)}}{s^\theta}ds\right]|(-A)^{-\theta}(x - z)| \\ &\le \left[\frac{M_\theta}{m(\delta)^\theta} + \frac{LM_\theta}{1 - \theta}M(\delta)^{1-\theta}e^{(L-\omega)M(\delta)}\right]|(-A)^{-\theta}(x - z)| \end{aligned}$$

for all $t \in [T(x), \tau(x, u_\varepsilon)]$. For $t = \tau(x, u_\varepsilon)$, the left-hand side of the inequality is $|x' - z'|$. Arguing as in (4.5), we then conclude that

$$|T(x) - T(z)| \le K_\theta |(-A)^{-\theta}(x - z)| + \varepsilon$$

for some $K_\theta > 0$. Since ε is arbitrary, the proof is complete. □

Remark 4.2. From estimate (4.2) it is easy to deduce that $D^+T(x) \subset D((-A^*)^\theta)$ for all $\theta \in [0, 1[$. Moreover, if $D((-A)^\theta)$ is compactly embedded into X for some $\theta \in]0, 1[$, then (4.2) implies that T is sequentially weakly continuous near x_0.

We can now prove our semiconcavity result for the minimum time function.

THEOREM 4.3. *Let assumptions* (H1), (H2), (H3), (H4), *and* (H5) *be satisfied, and let* α *and* β *be as in* (H4) *and* (H5), *resp. Then the minimum time function* T *belongs to* $SC^\gamma(\mathcal{R} \setminus \overline{B}_R)$, *where* $\gamma = \min\{\alpha, \beta\}$.

Proof. Let $\overline{x} \in \mathcal{R} \setminus \overline{B}_R$ be fixed. By (4.3) we can choose $m^*, M^*, \delta^* > 0$ such that

$$0 < m^* \le \inf_{x \in B_{\delta^*}(\overline{x})} T(x) < \sup_{x \in B_{\delta^*}(\overline{x})} T(x) \le M^* < +\infty \tag{4.6}$$

and

$$2\delta^* < \rho e^{M^*(\omega - L)}\,, \tag{4.7}$$

where ρ is given by Proposition 3.8. From (3.17) we obtain that, for all $x, z \in B_{\delta^*}(\overline{x})$ and all $u \in \mathcal{U}$,

$$|y(t; x, u) - y(t; z, u)| < \rho\,, \qquad t \in [0, M^*]\,. \tag{4.8}$$

Moreover, by (3.18), for any $\theta \in [0, 1[$ there exists $C_\theta > 0$ such that

$$|(-A)^\theta y(t; x, u)| \le C_\theta \tag{4.9}$$

for all $x \in B_{\delta^*}(\overline{x})$, all $u \in \mathcal{U}$, and all $t \in [m^*/2, T(x)]$.

Let us now take $x, x + h, x - h \in B_\delta(\overline{x})$, where $\delta \in]0, \delta^*]$ is a number which will be specified later in the proof. We have to prove that

$$T(x + h) + T(x - h) - 2T(x) \le C|h|^{1+\gamma} \tag{4.10}$$

for some constant $C > 0$ independent of x, h.

Without loss of generality, we can assume the existence of an optimal control, u, for x, i.e.,

$$T(x) = \tau(x, u)\,.$$

In the general case the conclusion follows by an approximation argument, as in the proof of Proposition 4.1.

We observe that, if $T(x) \ge \max\{\tau(x - h, u), \tau(x + h, u)\}$, then (4.10) is trivial as

$$T(x + h) + T(x - h) - 2T(x) \le \tau(x - h, u) + \tau(x + h, u) - 2T(x) \le 0\,.$$

Therefore, it is enough to consider the two cases $T(x) \le \min\{\tau(x - h, u), \tau(x + h, u)\}$ and $\tau(x - h, u) < T(x) < \tau(x + h, u)$. We will denote by $c_0, c_1, \ldots$ positive constants, independent of x and h.

Case 1. Suppose that

$$T(x) \le \min\{\tau(x - h, u), \tau(x + h, u)\}. \tag{4.11}$$

Let us set

$$x_0 = y(T(x); x+h, u), \quad x_1 = y(T(x); x, u), \quad x_2 = y(T(x); x-h, u).$$

Then, by definition, $x_1 \in \partial B_R$. From the dynamic programming principle, Proposition 3.8, and property (4.8), we obtain

$$T(x+h) + T(x-h) - 2T(x) \le T(x_0) + T(x_2) \le \frac{d(x_0) + d(x_2)}{r - (R+\rho)(L-\omega)}.$$

In addition, from (3.17), (3.20), and the semiconcavity of d in $X \setminus \overline{B}_R$ (see Remark 3.4) we deduce

$$\begin{aligned} d(x_0) + d(x_2) &\le 2d\left(\frac{x_0+x_2}{2}\right) + c_0|x_0 - x_2|^{1+\alpha} \\ &\le |x_0 + x_2 - 2x_1| + c_0|x_0 - x_2|^{1+\alpha} \le c_1|h|^{1+\beta} + c_2|h|^{1+\alpha}. \end{aligned}$$

This proves (4.10) in this case.

Case 2. Suppose that

$$\tau(x-h, u) < T(x) < \tau(x+h, u). \tag{4.12}$$

Before starting our analysis let us point out that the main ideas of the proof are basically the same of the previous case. We will again exploit the regularity of f, the semiconcavity of the distance function, and the estimate of T near the target given by Proposition 3.8. However, if we proceed as in Case 1 and let the points $x-h$ and $x+h$ evolve according to a control u which is optimal for x, we obtain no useful information. We have to use instead controls which are suitable rescalings of u; this choice takes into account the fact that, by (4.12), the points $x-h$ and $x+h$ require a different amount of time to approach the target. For this reason, the analysis of this case is longer and requires more delicate estimates than the previous one.

Let us first set

$$\tau_0 = \frac{1}{2}\tau(x-h, u)$$

and

$$z_0 = y(\tau_0; x-h, u), \quad z_1 = y(\tau_0; x, u), \quad z_2 = y(\tau_0; x+h, u)\,.$$

We note that, for $j = 0, 1, 2$, $z_j \in D((-A)^\theta)$ for all $\theta \in [0, 1[$ in light of (4.9). In fact, this gain in regularity is the reason why we replace the points $x-h, x, x+h$ with z_0, z_1, z_2. By definition, we have that

$$T(z_0) \le \tau_0.$$

From the dynamic programming principle and (4.12) we obtain

$$\tau_0 + T(z_1) = T(x) > 2\tau_0 \ge \tau_0 + T(z_0),$$

and so

$$T(z_1) > T(z_0)\,. \tag{4.13}$$

Let us abbreviate

$$T_j = T(z_j), \quad j = 0, 1, 2.$$

Again by dynamic programming, we deduce that

$$T(x-h) + T(x+h) - 2T(x) \le T_0 + T_2 - 2T_1\,. \tag{4.14}$$

If $T_2 - T_1 \le T_1 - T_0$, then (4.10) follows from (4.14). Assume, on the contrary, that $T_2 - T_1 > T_1 - T_0$. Then a number a exists such that

$$a \in \left]\frac{T_1 - T_0}{T_1}, \frac{T_2 - T_1}{T_1}\right[\bigcap\,]0, 1[\,. \tag{4.15}$$

Indeed the above set can be empty only if $(T_1 - T_0)/T_1 \ge 1$ or if $(T_2 - T_1)/T_1 \le 0$. But both possibilities are excluded since $T_0 > 0$ and $T_2 - T_1 > T_1 - T_0 > 0$ by (4.13).

Next, having fixed a as in (4.15), we define

$$u_1(t) = u(t + \tau_0), \qquad u_0(t) = u_1\left(\frac{t}{1-a}\right), \qquad u_2(t) = u_1\left(\frac{t}{1+a}\right),$$

and

$$z_j(t) = y(t; z_j, u_j)\,, \quad j = 0, 1, 2.$$

Notice that, by (4.15), $(1+a)T_1 < T_2$ and $(1-a)T_1 < T_0$. We now claim that

$$d(z_2((1+a)T_1)) \le \rho, \qquad d(z_0((1-a)T_1)) \le \rho\,. \tag{4.16}$$

Assuming for a moment that (4.16) holds, we deduce from Proposition 3.8 and from the semiconcavity of d that

$$\begin{aligned} T_2 + T_0 - 2T_1 &\le T(z_2((1+a)T_1)) + T(z_0((1-a)T_1)) \\ &\le \frac{1}{r - (R+\rho)(L-\omega)}[d(z_2((1+a)T_1)) + d(z_0((1-a)T_1))] \\ &\le \frac{1}{r - (R+\rho)(L-\omega)} 2d\left(\frac{z_2((1+a)T_1) + z_0((1-a)T_1)}{2}\right) \\ &\quad + c_1 |z_2((1+a)T_1) - z_0((1-a)T_1)|^{1+\alpha}\,. \end{aligned} \tag{4.17}$$

Let us now prove our claim (4.16). To this end, define

$$\zeta_1(t) = z_2((1+a)t) - z_1(t), \quad \zeta_2(t) = z_0((1-a)t) - z_1(t)\,. \tag{4.18}$$

We note that $\zeta_1(t)$ is the solution of the problem

$$\begin{cases} \zeta_1'(t) &= (1+a)\,\{Az_2((1+a)t) + f(z_2((1+a)t)) + u_1(t)\} \\ &\quad - \{Az_1(t) + f(z_1(t)) + u_1(t)\}\,, \\ \zeta_1(0) &= z_2 - z_1\,. \end{cases}$$

Hence, we have that

$$\begin{aligned} \zeta_1(t) = e^{tA}(z_2 - z_1) &+ \int_0^t e^{(t-s)A}\{f(z_2((1+a)s)) - f(z_1(s))\}\,ds \\ &+ a\int_0^t e^{(t-s)A}\{Az_2((1+a)s) + f(z_2((1+a)s)) + u_1(s)\}\,ds\,, \end{aligned}$$

which yields

$$|\zeta_1(t)| \le e^{-\omega t}|z_2 - z_1| + L\int_0^t e^{-\omega(t-s)}|\zeta_1(s)|ds + a\varphi(t), \tag{4.19}$$

where we have set $\varphi(t) = \int_0^t e^{(t-s)A}\left(Az_2((1+a)s) + f(z_2((1+a)s)) + u_1(s)\right)ds$. Using (4.9), we obtain

$$\begin{aligned}|\varphi(t)| &\le \int_0^t \left(\frac{C_{1-\theta}M_\theta}{(t-s)^\theta} + e^{-\omega(t-s)}(LC_0 + r)\right)ds \\ &\le \frac{C_{1-\theta}M_\theta}{1-\theta}t^{1-\theta} + t(LC_0 + r) \le c_2(1+t)\end{aligned}$$

for some $c_2 > 0$. Now, the Gronwall inequality gives

$$\begin{aligned}|\zeta_1(t)| &\le e^{-\omega t}|z_2 - z_1| + ac_2(1+t) \\ &\quad + L\int_0^t e^{(L-\omega)(t-s)}\left(e^{-\omega s}|z_2 - z_1| + ac_2(1+s)\right)ds\,,\end{aligned} \tag{4.20}$$

whence, by (3.17),

$$|\zeta_1(T_1)| \le 2\delta e^{M^*(L-\omega)} + ac_2(1+M^*) + LM^*e^{(L-\omega)M^*}\left\{2\delta e^{M^*(L-\omega)} + ac_2(1+M^*)\right\}.$$

The above inequality can be rewritten as

$$|\zeta_1(T_1)| \le c_3(\delta + a) \tag{4.21}$$

for some $c_3 > 0$. Recalling that $T_1 = T(x) - \frac{1}{2}\tau(x-h,u)$ we obtain from (4.12)

$$T_1 \ge \frac{1}{2}T(x) \ge \frac{1}{2}m^*.$$

The Lipschitz continuity of $T(\cdot)$, together with (4.15) and (3.17), yields

$$a \le \frac{T_2 - T_1}{T_1} \le \frac{c_4|z_2 - z_1|}{m^*} \le c_5|h| \le 2c_5\delta. \tag{4.22}$$

Thus, from inequality (4.21) we conclude that if $\delta \in]0, \delta^*]$ is fixed small enough, then

$$|\zeta_1(T_1)| \le \rho.$$

This inequality implies our claim (4.16) as far as $z_2(\cdot)$ is concerned. To obtain the conclusion for $z_0(\cdot)$, it suffices to replace ζ_1 by ζ_2 in the above argument.

It now remains to estimate the right-hand side in (4.17). We first observe that the inequalities in (4.22) imply that both a and $|z_2 - z_1|$ are of order $O(|h|)$. From inequality (4.20) we deduce that

$$|\zeta_1(t)| < c_6|h|, \qquad t \in [0, T_1]. \tag{4.23}$$

An analogous property holds for $|\zeta_2(t)|$, $t \in [0, T_1]$. We then obtain

$$|z_2((1+a)t) - z_0((1-a)t)|^{1+\alpha} = |\zeta_1(t) - \zeta_2(t)|^{1+\alpha} \le 4c_6|h|^{1+\alpha} \tag{4.24}$$

for any $t \in [0, T_1]$. In view of (4.17), our proof will be concluded if we show that

$$d\left(\frac{z_2((1+a)T_1) + z_0((1-a)T_1)}{2}\right) \le c|h|^{1+\beta} \tag{4.25}$$

for some constant $c > 0$. As a first step we show that, for some $c_7 > 0$,

$$|(-A)^{\theta_1}(z_2((1+a)t) - z_0((1-a)t))| \le c_7|h|t^{-\theta_1}\,, \ \forall t \in [0, T_1], \tag{4.26}$$

with θ_1 as in (3.13). Indeed, set $\eta(t) = z_2((1+a)t) - z_0((1-a)t)$, and observe that

$$\begin{aligned}
\eta(t) &= \zeta_1(t) - \zeta_2(t) \\
&= e^{tA}(z_2 - z_0) + \int_0^t e^{(t-s)A}\{f(z_2((1+a)s)) - f(z_0((1-a)s))\}ds \\
&\quad + a\int_0^t e^{(t-s)A}A(z_2((1+a)s) - z_0((1-a)s))ds \\
&\quad + a\int_0^t e^{(t-s)A}\{2u_1(s) + f(z_2((1+a)s)) + f(z_0((1-a)s))\}ds\,.
\end{aligned}$$

Then

$$\begin{aligned}
|(-A)^{\theta_1}\eta(t)| &\le \frac{M_{\theta_1}}{t^{\theta_1}}|z_2 - z_0| + LM_{\theta_1}\int_0^t \frac{|\eta(s)|}{(t-s)^{\theta_1}}\,ds \\
&\quad + a\int_0^t |(-A)^{1+\theta_1}e^{(t-s)A}(z_2((1+a)s) + z_0((1-a)s))|ds \\
&\quad + a\int_0^t |(-A)^{\theta_1}e^{(t-s)A}(f(z_2((1+a)s)) + f(z_0((1-a)s)))|ds \\
&\quad + 2a\int_0^t |(-A)^{\theta_1}e^{(t-s)A}u_1(s)|ds. \qquad (4.27)
\end{aligned}$$

We recall that both $|z_2 - z_1|$ and a are of order $O(|h|)$. The last two integrals in (4.27) are easily estimated using (3.5) and the boundedness of z_0, z_2, u_1. In addition, we have, by (3.5) and (4.9),

$$\begin{aligned}
&a\int_0^t |(-A)^{1+\theta_1}e^{(t-s)A}(z_2((1+a)s) + z_0((1-a)s))|ds \\
&\le a\int_0^t |(-A)^{\frac{1+\theta_1}{2}}e^{(t-s)A}(-A)^{\frac{1+\theta_1}{2}}(z_2((1+a)s) + z_0((1-a)s))|ds \\
&\le a\int_0^t c_8(t-s)^{-\frac{1+\theta_1}{2}}ds \le c_9 a.
\end{aligned}$$

Finally, equality (4.24) shows that $|\eta(t)| = O(|h|)$, and our claim (4.26) follows.

Let us now define $\xi(t) = z_0((1-a)t) + z_2((1+a)t) - 2z_1(t)$ and note that, by (3.20) we have $\xi(0) \le c_{10}|h|^{1+\beta}$ for some $c_{10} > 0$. Moreover,

$$\begin{aligned}\xi'(t) &= (1-a)\{Az_0((1-a)t) + f(z_0((1-a)t)) + u_1(t)\}\\ &\quad + (1+a)\{Az_2((1+a)t) + f(z_2((1+a)t)) + u_1(t)\}\\ &\quad -2\{Az_1(t) + f(z_1(t)) + u_1(t)\}\\ &= A\xi(t) + a\{A(z_2((1+a)t) - z_0((1-a)t)) + f(z_2((1+a)t)) - f(z_0((1-a)t))\}\\ &\quad + (f(z_0((1-a)t)) + f(z_2((1+a)t)) - 2f(z_1(t)).\end{aligned}$$

Hence,

$$\begin{aligned}\xi(t) &= e^{tA}\xi(0) + \int_0^t 2e^{(t-s)A}\left\{f\left(\frac{z_0((1-a)t) + z_2((1+a)t)}{2}\right) - f(z_1(s))\right\}ds\\ &+ \int_0^t e^{(t-s)A}\{f(z_0((1-a)s)) + f(z_2((1+a)s))\\ &\qquad -2f\left(\frac{z_0((1-a)t) + z_2((1+a)t)}{2}\right)\Big\}ds\\ &+ a\int_0^t e^{(t-s)A}(f(z_0((1-a)s)) - f(z_2((1+a)s))ds + a\int_0^t e^{(t-s)A}A\eta(s)ds.\end{aligned}$$

Recalling the Lipschitz continuity of f and formula (3.13), we obtain

$$\begin{aligned}|\xi(t)| &\le c_{10}|h|^{1+\beta} + L\int_0^t |\xi(s)|ds + \int_0^t L^*|(-A)^{\theta_1}\eta(s)|^{1+\beta}\,ds\\ &\quad + aL\int_0^t |\eta(s)|ds + M_{1-\theta_1}\int_0^t \frac{1}{(t-s)^{1-\theta_1}}|(-A)^{\theta_1}\eta(s)|ds\,.\end{aligned}$$

Now, using (4.24), (4.26), and the Gronwall inequality, we conclude that

$$|\xi(t)| \le c_{11}|h|^{1+\beta}, \qquad t \in [0, T_1], \tag{4.28}$$

which implies (4.25) and completes our proof. □

Since the minimum time function T is semiconcave, its singular set enjoys the properties stated in Theorem 2.5. In the next section, we will use Theorem 2.6 to derive a criterion for the propagation of the singularities of T.

5. Optimality conditions. In this section we use the fact that T is semiconcave to derive some optimality conditions for the minimum time problem. From now on, we will suppose that all assumptions (H1)–(H5) are satisfied, and, moreover, that

(H6) X is a Hilbert space;

(H7) e^{tA} is a compact operator for every $t > 0$;

(H8) A is self-adjoint;

(H9) a constant $\hat{L} > 0$ exists such that

$$|[\delta f(x) - \delta f(y)]z| \le \hat{L}|(-A)^{\theta_1}(x-y)||(-A)^{\theta_1}z| \quad \forall x, y, z \in D((-A)^{\theta_1}), \tag{5.1}$$

where θ_1 is given by (H5). In what follows we omit recalling the above hypotheses.

Remark 5.1. We note that, if (H6) holds, then assumption (H4) is satisfied with $\alpha = 1$. Moreover, (H9) yields estimate (3.11) in (H5) with $\beta = 1$. In addition, (H1) implies that

$$\langle Ax, x\rangle \le -\omega|x|^2 \qquad \forall x \in D(A). \tag{5.2}$$

Remark 5.2. It is well known that the parabolic system (3.14) satisfies (H7) and (H8). Moreover, (H9) holds if $n \le 3$. In fact, in this case, $\theta_1 \in]0, 1/2[$ exists such that $W^{2\theta_1,2}(\mathcal{O}) \subset L^4(\mathcal{O})$. Then

$$|[\delta f(x) - \delta f(y)]z| \le L'||x-y||_{L^4(\mathcal{O})}||z||_{L^4(\mathcal{O})} \le \hat{L}||x-y||_{W^{2\theta_1,2}(\mathcal{O})}||z||_{W^{2\theta_1,2}(\mathcal{O})}.$$

We will now begin to analyze properties of optimal controls and optimal trajectories for problem (3.9). The existence of such optimal pairs can be proved, arguing as in [7, p. 365]. Let then $y(\cdot)$ be an optimal trajectory starting from a given point $x \in \mathcal{R}$. We consider the *adjoint system* associated with y

$$p'(t) = -Ap(t) - (\delta f(y(t)))^* p(t), \qquad t \in]0, T(x)] \tag{5.3}$$

together with the *transversality condition*

$$p(T(x)) = \frac{y(T(x))}{rR + |(-A)^{\frac{1}{2}} y(T(x))|^2 - \langle f(y(T(x))), y(T(x))\rangle}\,. \tag{5.4}$$

We observe that the right-hand side of (5.4) is well defined. In fact, $y(T(x)) \in D((-A)^{1/2})$ by Lemma 3.6(iv); in addition, by (H2), (H3), and (5.2), we have that

$$\begin{aligned} &rR + |(-A)^{\frac{1}{2}} y(T(x))|^2 - \langle f(y(T(x))), y(T(x))\rangle \\ &\ge rR + (\omega - L)|y(T(x))|^2 = R(r + (\omega - L)R) > 0. \end{aligned}$$

Moreover, applying Proposition 3.9, we can check easily that problem (5.3)–(5.4) has a unique solution given by

$$p(t) = \frac{G^*(T(x), t)y(T(x))}{rR + |(-A)^{\frac{1}{2}} y(T(x))|^2 - \langle f(y(T(x))), y(T(x))\rangle}. \tag{5.5}$$

Such a solution will be called the *dual arc* associated with $y(\cdot)$. The following inclusion is interesting in itself and has several important consequences.

THEOREM 5.3. *Let $y(\cdot) = y(\cdot; x, u)$ be a time optimal trajectory for x and p be the dual arc associated with $y(\cdot)$. Then*

$$p(t) \in D^+ T(y(t))$$

for any $t \in [0, T(x)[$.

To prove the above theorem we need a technical lemma.

LEMMA 5.4. *Let $\alpha \in [0, 1[$, $\theta \in]0, 1-\alpha[$ and $\overline{T} > 0$ be fixed. Then a constant $C > 0$ exists such that*

$$|(-A)^\alpha (y(t; \overline{x}, u)) - \overline{x})| \le C(1 + |(-A)^{\theta+\alpha}\overline{x}|)t^\theta \tag{5.6}$$

for all $\overline{x} \in D((-A)^{\theta+\alpha}) \cap B_{2R}$, for all controls $u \in \mathcal{U}$ and for all $t \in [0, \overline{T}]$.

Proof. First of all we note that

$$|(-A)^\alpha(y(t;\overline{x},u))-\overline{x})| \le |(-A)^\alpha(e^{tA}-I)\overline{x})| + \left|\int_0^t (-A)^\alpha e^{(t-s)A}(f(y(s;\overline{x},u))+u(s))ds\right|.$$

In addition we have, using (3.5),

$$\begin{aligned}|(-A)^\alpha(e^{tA}-I)\overline{x})| &= \left|\int_0^t (-A)^{1+\alpha}e^{\sigma A}\overline{x}d\sigma\right| \\ &= \left|\int_0^t (-A)^{1-\theta}e^{\sigma A}(-A)^{\theta+\alpha}\overline{x}d\sigma\right| \\ &\le M_{1-\theta}|(-A)^{\theta+\alpha}\overline{x}|\frac{t^\theta}{\theta}.\end{aligned}$$

On the other hand, using (H2) and (3.16), we obtain that there exists $c_1>0$ such that $|f(y(s;\overline{x},u))+u(s)|<c_1$ for any $\overline{x}\in B_{2R}$, $s\in[0,\overline{T}]$ and $u\in\mathcal{U}$. Therefore, using again (3.5), we find

$$\left|\int_0^t (-A)^\alpha e^{(t-s)A}(f(y(s;\overline{x},u))+u(s))ds\right| \le M_\alpha c_1\int_0^t \frac{ds}{s^\alpha} \le c_2 t^\theta,$$

which implies the conclusion. □

Proof of Theorem 5.3. It is enough to prove the result for the initial time $t=0$. In fact, if $t_0\in]0,T(x)[$, then, by the dynamic programming principle, the restriction of $y(\cdot)$ to the interval $[t_0,T(x)]$ is an optimal trajectory for the point $y(t_0)$ and the restriction of $p(\cdot)$ to the same interval is the associated dual arc. Thus, knowing that the inclusion holds for the initial time yields that $p(t_0)\in D^+T(y(t_0))$.

Therefore it suffices to show that $p(0)\in D^+T(x)$, or, equivalently, that

$$T(x+h)-T(x)-\langle p(0),h\rangle \le o(|h|).$$

For simplicity we set

$$T:=T(x) \quad\text{and}\quad \sigma := rR+|(-A)^{\frac{1}{2}}y(T)|^2-\langle f(y(T)),y(T)\rangle.$$

Then, by (5.5), $\langle p(0),h\rangle = \langle \sigma^{-1}y(T), G(T,0)h\rangle$. Therefore, all we need to show is the fact that

$$T(x+h)-T(x)-\left\langle \frac{y(T)}{\sigma}, G(T,0)h\right\rangle \le o(|h|), \tag{5.7}$$

where $h\in X$ is sufficiently small. Let us consider the trajectories $y_h(t):=y(t;x+h,u)$. In order to prove (5.7), we have to distinguish two cases.

Case 1. Suppose that

$$\tau(x+h,u)=:t_h<T(x).$$

Then

$$0<T-t_h\le T(x)-T(x+h)=O(|h|). \tag{5.8}$$

We note that

$$d(y(t_h)) = \Big\langle \frac{y(T)}{R}, y(t_h) - y(T) \Big\rangle + O(|y(t_h) - y(T)|^2). \tag{5.9}$$

Now, Lemma 5.4 (with $\alpha = 0$ and $\overline{x} = y(t_h)$) implies that $|y(T) - y(t_h)| = O((T - t_h)^\theta)$ for some $\theta > 1/2$, and so $|y(T) - y(t_h)|^2 = o(|h|)$. Moreover, we have that

$$\begin{aligned}
\Big\langle \frac{y(T)}{R}, y(t_h) - y(T) \Big\rangle &= -\int_{t_h}^T \frac{d}{dt} \Big\langle \frac{y(T)}{R}, y(t) \Big\rangle \, dt \\
&= -\int_{t_h}^T \Big\langle \frac{y(T)}{R}, Ay(t) + f(y(t)) + u(t) \Big\rangle \, dt \\
&\le -\int_{t_h}^T \left\{ \Big\langle \frac{y(T)}{R}, Ay(t) + f(y(t)) \Big\rangle - r \right\} dt \\
&= (T - t_h) \left\{ \frac{|(-A)^{\frac{1}{2}} y(T)|^2}{R} - \frac{\langle y(T), f(y(T)) \rangle}{R} + r \right\} \\
&\quad - \int_{t_h}^T \Big\langle (-A)^{\frac{3}{4}} \frac{y(T)}{R}, (-A)^{\frac{1}{4}} (y(T) - y(t)) \Big\rangle \, dt \\
&\quad + \int_{t_h}^T \Big\langle \frac{y(T)}{R}, f(y(T)) - f(y(t)) \Big\rangle \, dt = (\mathrm{I})_h + (\mathrm{II})_h + (\mathrm{III})_h \, .
\end{aligned}$$

From Lemma 5.4 we obtain that $|(-A)^\alpha (y(t) - y(T))| \to 0$ as $t \to T$ for any $\alpha \in [0, 1/2[$. Since $T - t_h = O(|h|)$ by (5.8), we have that $(\mathrm{II})_h + (\mathrm{III})_h = o(|h|)$ as $h \to 0$. Thus (5.9) implies that

$$d(y(t_h)) \le \frac{(T - t_h)\sigma}{R} + o(|h|) \, .$$

Then, we obtain from (5.8)

$$T(x + h) - T(x) \le t_h - T \le -\frac{R \, d(y(t_h))}{\sigma} + o(|h|) \, . \tag{5.10}$$

Next, by (3.24),

$$\begin{aligned}
d(y(t_h)) &= \Big\langle \frac{y_h(t_h)}{R}, y(t_h) - y_h(t_h) \Big\rangle + O(|y(t_h) - y_h(t_h)|^2) \\
&= \Big\langle \frac{y(T)}{R}, y(t_h) - y_h(t_h) \Big\rangle + \Big\langle \frac{y(t_h) - y(T)}{R}, y(t_h) - y_h(t_h) \Big\rangle \\
&\quad + O(|y(t_h) - y_h(t_h)|^2) \\
&= -\Big\langle \frac{y(T)}{R}, G(t_h, 0)h \Big\rangle + o(|h|) = -\Big\langle \frac{y(T)}{R}, G(T, 0)h \Big\rangle + o(|h|),
\end{aligned}$$

which, together with (5.10), proves claim (5.7) in this case.

Case 2. Suppose that

$$\tau(x+h,u) > T(x).$$

For simplicity we set $x_h = y(T(x); x+h, u)$, and we note that $x_h \notin \overline{B}_R$. Then, the dynamic programming principle gives

$$T(x+h) - T(x) \le T(x_h).$$

Our goal is to estimate $T(x_h)$ with $d(x_h)$. Since $x_h = y(T(x)) + G(T(x),0)h + o(|h|)$ by (3.24), we have

$$d(x_h) = \Big\langle \frac{y(T(x))}{R}, G(T(x),0)h \Big\rangle + o(|h|).$$

Therefore, to prove (5.7) in this case, we only need to show that

$$T(x_h) \le \frac{R\, d(x_h)}{\sigma} + o(|h|)\,. \tag{5.11}$$

For this purpose, we consider the problem

$$\begin{cases} \overline{y}_h'(t) = A\overline{y}_h(t) + f(\overline{y}_h(t)) + u^*(t), & t > 0, \\ \overline{y}_h(0) = x_h\,, \end{cases} \tag{5.12}$$

where u^* is a feedback control of the form $u^*(t) = -r\overline{y}_h(t)/|\overline{y}_h(t)|$. We are interested in studying the solution of system (5.12) only for $|\overline{y}_h(t)| \ge R$; hence we can consider the system as a semilinear evolution equation with a Lipschitz nonlinearity. First of all, observe that, by (H2) and (5.2),

$$\begin{aligned} \frac{d}{dt}\frac{|\overline{y}_h(t)|^2}{2} &= \langle A\overline{y}_h(t) + f(\overline{y}_h(t)), \overline{y}_h(t)\rangle - r|\overline{y}_h(t)| \\ &\le (L-\omega)|\overline{y}_h(t)|^2 - r|\overline{y}_h(t)|. \end{aligned}$$

If $|h|$ is small enough, then $|x_h| \le r + |y(T(x)) - x_h| < R$. Then, from (H3) and the above inequality it follows that $\tau(x_h, u^*) < +\infty$ for $|h|$ small, and that $\tau(x_h, u^*) \to 0$ as $h \to 0$. Let us set, for simplicity, $\tau_h = \tau(x_h, u^*)$. Then,

$$\begin{aligned} d(x_h) &= \Big\langle \frac{\overline{y}_h(\tau_h)}{R}, x_h - \overline{y}_h(\tau_h) \Big\rangle + O(|x_h - \overline{y}_h(\tau_h)|^2) \\ &= -\int_0^{\tau_h} \frac{d}{dt}\Big\langle \frac{\overline{y}_h(\tau_h)}{R}, \overline{y}_h(t) \Big\rangle\, dt + O(|x_h - \overline{y}_h(\tau_h)|^2) \\ &= -\int_0^{\tau_h} \Big\langle \frac{\overline{y}_h(\tau_h)}{R}, A\overline{y}_h(t) + f(\overline{y}_h(t)) - r\frac{\overline{y}_h(t)}{|\overline{y}_h(t)|} \Big\rangle\, dt + O(|x_h - \overline{y}_h(\tau_h)|^2). \end{aligned}$$

Arguing as in Case 1 and applying Lemma 5.4, we obtain

$$d(x_h) = \int_0^{\tau_h} \frac{\sigma}{R}\, dt + o(|h|) = \frac{\tau_h\, \sigma}{R} + o(|h|)\,,$$

which implies (5.11) and concludes our proof. □

We now state the Pontryagin maximum principle for our control system.

PROPOSITION 5.5. *Let $y(\cdot) = y(\cdot; x, u)$ be an optimal trajectory for x and $p(t)$ be the corresponding dual arc. Then, $p(t) \neq 0$ for all $t \in [0, T(x)[$, and*

$$u(t) = -r\frac{p(t)}{|p(t)|} \tag{5.13}$$

for almost everywhere (a.e.) $t \in [0, T(x)]$.

Proof. The maximum principle is a classical result in the theory of optimal control. For the minimum time problem, it is well known (see e.g., [7, p. 367]) that there exists at least one arc $p(\cdot)$ solution of (5.3) such that (5.13) holds. In our particular case, since the target set is a ball, we can prescribe the final condition for $p(\cdot)$ to be a normal vector to the ball at the terminal point of the trajectory, i.e., any scalar multiple of $y(T(x))$. This can be proved as in the finite dimensional case (see [9]). The normalization in the transversality condition (5.4) is chosen in such a way that Theorem 5.3 holds.

The fact that $p(t) \neq 0$ for any t is not obvious a priori. In fact, (5.3) is, in general, ill posed forward in time as $-A$ fails to be the generator of a C_0 semigroup. Nevertheless, a forward uniqueness property is proved in [10, Theorem A.1] for such equations, which ensures that $p(t)$ never vanishes if the final condition is nonzero, as in (5.4). □

Coupling state equation (3.1) with the adjoint system (5.3) and using (5.13), we obtain the Hamiltonian form of the maximum principle

$$\begin{cases} y'(t) = Ay(t) + f(y(t)) - r\dfrac{p(t)}{|p(t)|} \\ \\ p'(t) = -Ap(t) - (\delta f(y(t)))^* p(t)\,. \end{cases} \tag{5.14}$$

Once again, we note that the Cauchy problem for the above system is, in general, ill posed. However, as we show in the next proposition, the system satisfies a forward uniqueness property like the one recalled in the previous proof for the adjoint equation (5.3).

PROPOSITION 5.6. *Given $x, q \in X$, with $q \neq 0$ and $T > 0$, there exists at most one pair $(y, p) : [0, T] \to X \times (X \setminus \{0\})$ which solves system (5.14) in $[0, T[$, and satisfies the initial conditions*

$$\begin{cases} y(0) = x \\ p(0) = q\,. \end{cases} \tag{5.15}$$

The following proof uses a technique introduced by [25]; this method was adapted to systems by [13], in a different situation from the one of interest in this paper.

Proof. Suppose that there exist two solutions, say $(y_1(t), p_1(t))$ and $(y_2(t), p_2(t))$, of problem (5.14)–(5.15). Define $\bar{y}(t) := y_2(t) - y_1(t)$ and $\bar{p}(t) := p_2(t) - p_1(t)$. We note that the pair $(\bar{y}(t), \bar{p}(t))$ is a solution of the system

$$\begin{cases} \bar{y}'(t) = A\bar{y}(t) + f(y_2(t)) - f(y_1(t)) - r\dfrac{p_2(t)}{|p_2(t)|} + r\dfrac{p_1(t)}{|p_1(t)|}, \\ \\ \bar{p}'(t) = -A\bar{p}(t) - \delta f(y_2(t))^* p_2(t) + \delta f(y_1(t))^* p_1(t) \end{cases} \tag{5.16}$$

for $t \in]0, T[$, with initial conditions

$$\begin{cases} \bar{y}(0) = 0, \\ \bar{p}(0) = 0. \end{cases} \tag{5.17}$$

Now, let us take a function $\theta \in C^1(\mathbb{R})$ such that $\theta(t) = 1$ for $0 \le t \le \frac{T}{2}$, $\theta(t) = 0$ for $t = T$ and $|\theta'(t)| \le \frac{4}{T}$. We set

$$z(t) := e^{\frac{k(t-T)^2}{2}}\theta(t)\bar{y}(t) \qquad \text{and} \qquad q(t) := e^{\frac{k(t-T)^2}{2}}\theta(t)\bar{p}(t).$$

We have that

$$\begin{cases} z'(t) = Az(t) + k(t-T)z(t) + \phi_1(t)\,, \\ z(0) = z(T) = 0, \\ q'(t) = -Aq(t) + k(t-T)q(t) + \phi_2(t)\,, \\ q(0) = q(T) = 0, \end{cases} \tag{5.18}$$

where

$$\phi_1(t) = e^{\frac{k(t-T)^2}{2}}\theta(t)\left(f(y_2(t)) - f(y_1(t)) - r\frac{p_2(t)}{|p_2(t)|} + r\frac{p_1(t)}{|p_1(t)|}\right) + e^{\frac{k(t-T)^2}{2}}\theta'(t)\bar{y}(t) \tag{5.19}$$

and

$$\phi_2(t) = e^{\frac{k(t-T)^2}{2}}\theta(t)(\delta f(y_1(t))^* p_1(t) - \delta f(y_2(t))^* p_2(t)) + e^{\frac{k(t-T)^2}{2}}\theta'(t)\bar{p}(t)\,. \tag{5.20}$$

Then, multiplying the first equation of system (5.18) by $z'(t)$ and the second equation by $q'(t)$ we get

$$|z'(t)|^2 = \frac{1}{2}\frac{d}{dt}\{\langle Az(t), z(t)\rangle + k(t-T)|z(t)|^2\} - \frac{k}{2}|z(t)|^2 + \langle \phi_1(t), z'(t)\rangle$$

and

$$|q'(t)|^2 = \frac{1}{2}\frac{d}{dt}\{\langle -Aq(t), q(t)\rangle + k(t-T)|q(t)|^2\} - \frac{k}{2}|q(t)|^2 + \langle \phi_2(t), q'(t)\rangle.$$

Integrating the equations above on $[0,T]$ and recalling that z and q vanish at the endpoints, we obtain

$$\int_0^T \Big(|z'(t)|^2 + \frac{k}{2}|z(t)|^2\Big)dt \le \frac{1}{2}\int_0^T (|z'(t)|^2 + |\phi_1(t)|^2)dt$$

and

$$\int_0^T \Big(|q'(t)|^2 + \frac{k}{2}|q(t)|^2\Big)dt \le \frac{1}{2}\int_0^T (|q'(t)|^2 + |\phi_2(t)|^2)dt.$$

Hence,

$$k\int_0^T (|z(t)|^2 + |q(t)|^2)dt \le \int_0^T (|\phi_1(t)|^2 + |\phi_2(t)|^2)dt\,. \tag{5.21}$$

We have now to estimate the right-hand side of (5.21). Let us first note that, by Proposition 5.5,

$$e^{k(t-T)^2}\theta^2(t)\left|\frac{p_2(t)}{|p_2(t)|} - \frac{p_1(t)}{|p_1(t)|}\right|^2 \le C_1|q(t)|^2$$

for some $C_1 > 0$. Here and in the remainder of the proof $C_1, C_2, \dots$ denote positive numbers independent of k. We have

$$\begin{aligned}|\phi_1(t)|^2 &\le 2L^2|z(t)|^2 + 4e^{k(t-T)^2}|\theta'(t)\bar{y}(t)|^2 \\ &\quad + 4r^2 e^{k(t-T)^2}\theta^2(t)\left|\frac{p_2(t)}{|p_2(t)|} - \frac{p_1(t)}{|p_1(t)|}\right|^2 \\ &\le 2L^2|z(t)|^2 + 4e^{k(t-T)^2}|\theta'(t)\bar{y}(t)|^2 + C_2|q(t)|^2\end{aligned}$$

for some $C_2 > 0$. Moreover,

$$|\phi_2(t)|^2 \le 2e^{k(t-T)^2}|\theta'(t)\bar{p}(t)|^2 + 2e^{k(t-T)^2}\theta^2(t)|\delta f(y_1(t))^* p_1(t) - \delta f(y_2(t))^* p_2(t)|^2.$$

Now, the interpolation inequality (3.7) and assumption (5.1) yield

$$\begin{aligned}&|\delta f(y_1(t))^* p_1(t) - \delta f(y_2(t))^* p_2(t)| \\ &\le |[\delta f(y_1(t))^* - \delta f(y_2(t))^*]p_2(t) + \delta f(y_1(t))^*(p_1(t) - p_2(t))| \\ &\le C_3\{|(-A)^{\theta_1}(y_2(t) - y_1(t))||(-A)^{\theta_1}p_2(t)| + |p_1(t) - p_2(t)|\} \\ &\le C_4\{|(-A)^{\theta_1}(y_2(t) - y_1(t))| + |p_1(t) - p_2(t)|\} \\ &\le C_4\{|(-A)^{\frac{1}{2}}(y_2(t) - y_1(t))| + \kappa_{\theta_1}|y_2(t) - y_1(t)| + |p_1(t) - p_2(t)|\}\end{aligned}$$

for some $C_3, C_4 > 0$. Therefore, we obtain

$$|\phi_2(t)|^2 \le 2e^{k(t-T)^2}|\theta'(t)\bar{p}(t)|^2 + C_5(|(-A)^{\frac{1}{2}}z(t)|^2 + |z(t)|^2 + |q(t)|^2)$$

for some $C_5 > 0$. Multiplying by $z(t)$ the first equation of (5.18) and integrating, we obtain

$$\begin{aligned}\int_0^T |(-A)^{\frac{1}{2}}z(t)|^2 dt &\le \frac{1}{2}\int_0^T (|z(t)|^2 + |\phi_1(t)|^2)dt \\ &\le \frac{1}{2}\int_0^T ((1+2L^2)|z(t)|^2 + 4e^{k(t-T)^2}|\theta'(t)\bar{y}(t)|^2 + C_2|q(t)|^2)dt.\end{aligned}$$

Plugging the estimates of $|\phi_1(\cdot)|$ and $|\phi_2(\cdot)|$ into (5.21), we find

$$\begin{aligned}&k\int_0^T (|z(t)|^2 + |q(t)|^2)dt \\ &\le C_6\int_0^T (|z(t)|^2 + |q(t)|^2)dt + C_7\int_0^T e^{k(t-T)^2}|\theta'(t)|^2(|\bar{y}(t)|^2 + |\bar{p}(t)|^2)\,dt \\ &= C_6\int_0^T (|z(t)|^2 + |q(t)|^2)dt + C_7\int_{\frac{T}{2}}^T e^{k(t-T)^2}|\theta'(t)|^2(|\bar{y}(t)|^2 + |\bar{p}(t)|^2)\,dt \\ &\le C_6\int_0^T (|z(t)|^2 + |q(t)|^2)dt + C_7\frac{16}{T^2}e^{k\frac{T^2}{4}}\int_0^T (|\bar{y}(t)|^2 + |\bar{p}(t)|^2)\,dt\end{aligned}$$

for some $C_6, C_7 > 0$. On the other hand, if $k > C_6$,

$$(k - C_6)\int_0^T (|z(t)|^2 + |q(t)|^2)dt \ge (k - C_6)e^{k\frac{T^2}{4}}\int_0^{\frac{T}{2}} (|\bar{y}(t)|^2 + |\bar{p}(t)|^2)\,dt,$$

and therefore

$$\int_0^{\frac{T}{2}} (|\bar{y}(t)|^2 + |\bar{p}(t)|^2)dt \le \frac{16C_7}{(k - C_6)T^2} \int_0^T (|\bar{y}(t)|^2 + |\bar{p}(t)|^2)\, dt\,. \tag{5.22}$$

Sending $k \to \infty$, we obtain $|\bar{y}(t)| = |\bar{p}(t)| = 0$ on $[0, \frac{T}{2}]$. The conclusion follows by an easy iteration argument. □

The singular points of T can be characterized in terms of optimal trajectories. Before proving this fact, we need a preliminary result.

LEMMA 5.7. *Let x and x_n be given points of $\mathcal{R}\setminus\overline{B}_R$ such that $x_n \to x$ as $n \to \infty$. Let $y_n : [0, T(x_n)] \to X$ be optimal trajectories for x_n and let $p_n : [0, T(x_n)] \to X$ be the corresponding dual arcs. Then there exists a subsequence, $\{y_{n_k}(\cdot)\}$, converging uniformly on any interval $[0, T']$, with $T' < T(x)$, to a trajectory $y(\cdot)$, optimal for x. In addition, $\lim_{k\to\infty} p_{n_k}(t) = p(t)$ for all $t \in [0, T(x)[$, where $p(\cdot)$ is the dual arc associated with $y(\cdot)$.*

Proof. Let us denote by $u_n(\cdot)$ the optimal controls corresponding to $y_n(\cdot)$, and let us set for simplicity $T_n = T(x_n)$, $T = T(x)$. After extracting a subsequence, we find that there exists $u : [0, T] \to B_r$ such that $u_n \overset{*}{\rightharpoonup} u$ in $L^\infty([0, T'], B_r)$ for any $T' < T$. Then, using the compactness of the operators e^{tA}, we can see easily that $y(\cdot) := y(\cdot; x, u)$ is an optimal trajectory for x and that $y_n(\cdot) \to y(\cdot)$ uniformly in $[0, T']$ for any $T' < T$.

It remains to prove that $p_n(t) \to p(t)$ for all $t \in [0, T(x)[$. To this purpose, let us first show that $(-A)^{1/2}y_n(T_n)$ tends to $(-A)^{1/2}y(T)$. We have

$$\begin{aligned}|(-A)^{\frac{1}{2}}(y_n(T_n) - y(T))| &\le |(-A)^{\frac{1}{2}}(y_n(T_n) - y(T_n; x_n, u))| \\ &\quad + |(-A)^{\frac{1}{2}}(y(T_n; x_n, u) - y(T_n))| + |(-A)^{\frac{1}{2}}(y(T_n) - y(T))|.\end{aligned}$$

Using the representation formula (3.4) and property (3.5), we obtain that the first term on the right-hand side tends to 0 as $n \to \infty$. The second and third term also tend to 0, thanks to inequality (3.19) and to Lemma 5.4, resp. Thus we have that $(-A)^{1/2}y_n(T_n) \to (-A)^{1/2}y(T)$, and this implies, by (5.4), (H2), and (3.6), that $p_n(T_n) \to p(T)$.

Let us now prove that $p_n(t) \to p(t)$ for any $t \in [0, T(x)[$. To fix ideas, we suppose that $T_n > T$ for all n. Then we have, by (H5),

$$\begin{aligned}|p_n(t) - p(t)| &\le |e^{(T_n - t)A}(p_n(T_n) - p(T))| + |(e^{(T_n - t)A} - e^{(T-t)A})p(T)| \\ &\quad + \int_t^T |e^{(s-t)A}\delta f(y_n(s))^*(p_n(s) - p(s))|ds \\ &\quad + \int_T^{T_n} |e^{(s-t)A}\delta f(y_n(s))^* p_n(s)|ds \\ &\quad + \int_t^T |e^{(s-t)A}(\delta f(y_n(s))^* - \delta f(y(s))^*)p(s)|ds \\ &\le L\int_t^T |p_n(s) - p(s)|ds + o(1) \quad \text{as } n \to \infty.\end{aligned}$$

Then the conclusion follows by applying Gronwall's inequality. □

The following theorem provides a characterization of the singular points of T as the initial points of multiple time optimal trajectories.

THEOREM 5.8. *A point $x \in \mathcal{R} \setminus \overline{B}_R$ is a differentiability point for the minimum time function T iff there exist a unique optimal trajectory for system* (3.1) *with initial point x.*

Proof. Suppose first that T is differentiable at x. Let $y(\cdot)$ be any optimal trajectory starting at x, and let $p(\cdot)$ be the associated dual arc. Then, by (5.3), (5.13), and Theorem 5.3, we have that (y, p) is a solution of system (5.14) with initial conditions $y(0) = x, p(0) = DT(x)$. But Proposition 5.6 states that there exists at most a unique pair (y, p) with these properties. Therefore, x is the starting point of a unique optimal trajectory.

Conversely, let us assume that there exists a unique optimal trajectory starting at x, which we denote by $y(\cdot)$. We note that $D^+T(x) \neq \emptyset$ as the minimum time function is semiconcave. If we prove that $D^-T(x)$ is also nonempty, we obtain that T is differentiable at x. Hence, to conclude the proof it is enough to show that

$$p(0) \in D^-T(x), \tag{5.23}$$

where $p(\cdot)$ is the dual arc associated with $y(\cdot)$. For this purpose, let us consider a sequence x_k converging to x as $k \to \infty$ such that

$$\lim_{k\to\infty} \frac{T(x_k) - T(x) - \langle p(0), x_k - x\rangle}{|x_k - x|} = \liminf_{x'\to x} \frac{T(x') - T(x) - \langle p(0), x' - x\rangle}{|x' - x|}.$$

Now let $y_k(\cdot)$ be optimal trajectories for the points x_k, and let $p_k(\cdot)$ be the dual arcs associated with $y_k(\cdot)$. Since $y(\cdot)$ is the unique optimal trajectory for x, Lemma 5.7 implies that $y_k \to y$ and that $p_k \to p$ as $k \to \infty$. Using Theorem 5.3 and property (2.2) of semiconcave functions, we obtain

$$\begin{aligned}
&T(x_k) - T(x) - \langle p(0), x_k - x\rangle \\
&= T(x_k) - T(x) - \langle p_k(0), x_k - x\rangle - \langle p(0) - p_k(0), x_k - x\rangle \\
&\geq -C|x_k - x|^2 - |p(0) - p_k(0)|\,|x_k - x|\,.
\end{aligned}$$

Hence,

$$\lim_{k\to\infty} \frac{T(x_k) - T(x) - \langle p(0), x_k - x\rangle}{|x_k - x|} \geq 0,$$

and the proof is complete. □

We now proceed to show that, for the control systems under investigation, the minimum time function is differentiable along any optimal trajectory, except for the end points. For this purpose, we use the Hamilton–Jacobi–Bellman equation

$$r|DT(x)| - \langle DT(x), Ax\rangle - \langle DT(x), f(x)\rangle = 1. \tag{5.24}$$

The next lemma shows that the above equation is satisfied, in a suitable sense, along any optimal trajectory.

LEMMA 5.9. *Let $y(\cdot) = y(\cdot; x, u)$ be a time optimal trajectory for a point $x \in \mathcal{R} \setminus \overline{B}_R$. Then, for any $t \in]0, T(x)[\,, \theta \in]0, 1[$, and $q \in D^+T(y(t))$, we have that*

$$\langle (-A)^{1-\theta} q, (-A)^{\theta} y(t)\rangle - \langle q, f(y(t))\rangle + r|q| = 1. \tag{5.25}$$

Proof. First we observe that all the terms in (5.25) are well defined since, by Remark 4.2 and Lemma 3.6(iv), both q and $y(t)$ belong to $D((-A)^\theta)$ for any $\theta \in [0,1[$. Now, having fixed $t \in]0,T(x)[$, $\theta \in]0,1[$ and $q \in D^+T(y(t))$, we note that, for all $h>0$,

$$\frac{1}{h}\langle q, y(t)-y(t-h)\rangle$$

$$= \frac{1}{h}\langle q, (e^{hA}-I)y(t-h)\rangle + \frac{1}{h}\int_0^h \langle e^{(h-s)A}q, f(y(t+s-h))+u(t+s-h)\rangle ds.$$

Hence,

$$\lim_{h\downarrow 0}\frac{1}{h}\langle q,(e^{hA}-I)y(t-h)\rangle = -\lim_{h\downarrow 0}\frac{1}{h}\int_0^h \langle (-A)^{1-\theta}q, (-A)^\theta e^{sA}y(t-h)\rangle ds$$

$$= -\langle (-A)^{1-\theta}q, (-A)^\theta y(t)\rangle\,.$$

Moreover, we have that

$$\limsup_{h\downarrow 0}\frac{1}{h}\int_0^h \langle e^{(h-s)A}q, f(y(t+s-h))+u(t+s-h)\rangle ds \geq \langle q, f(y(t)\rangle - r|q|.$$

From the dynamic programming principle we obtain

$$1 = \lim_{h\downarrow 0}\frac{1}{h}\{T(y(t-h))-T(y(t))\} \leq \liminf_{h\downarrow 0}\frac{1}{h}\langle q, y(t-h)-y(t)\rangle$$

$$\leq \langle (-A)^{1-\theta}q, (-A)^\theta y(t)\rangle - \langle q, f(y(t))\rangle + r|q|. \tag{5.26}$$

To prove the converse inequality, we introduce the control

$$\tilde{u}(s) = \begin{cases} u(s) & \text{if } s \leq t,\\ -rq/|q| & \text{if } s > t.\end{cases}$$

Let us set $\tilde{y}(s) = y(s;x,\tilde{u})$. Arguing as in the first part of the proof we obtain

$$\lim_{h\downarrow 0}\frac{1}{h}\langle q, \tilde{y}(t+h)-\tilde{y}(t)\rangle = -\langle (-A)^{1-\theta}q, (-A)^\theta y(t)\rangle + \langle q, f(y(t))\rangle - r|q|.$$

Again by the dynamic programming principle, we have that

$$-1 \leq \liminf_{h\downarrow 0}\frac{1}{h}\{T(\tilde{y}(t+h))-T(\tilde{y}(t))\} \leq \liminf_{h\downarrow 0}\frac{1}{h}\langle q, \tilde{y}(t+h)-\tilde{y}(t)\rangle$$

$$= -\langle (-A)^{1-\theta}q, (-A)^\theta y(t)\rangle + \langle q, f(y(t))\rangle - r|q|\,,$$

which, together with (5.26), yields the conclusion. □

We also need the following easy result.

LEMMA 5.10. *Let $x \in D((-A)^\theta)$ for some $\theta \in]0,1[$, and let $\Gamma \subset D((-A)^{1-\theta})$ be a nonempty convex set such that*

$$-\langle (-A)^{1-\theta}q, (-A)^\theta x\rangle + \langle q, f(x)\rangle - r|q| = 1 \tag{5.27}$$

for any $q \in \Gamma$. Then Γ is a singleton.

Proof. Suppose on the contrary that Γ contains two distinct elements q_1 and q_2. Then, since Γ is convex, equality (5.27) holds with $q = \lambda q_1 + (1-\lambda)q_2$, for all $\lambda \in [0,1]$. This implies

$$|\lambda q_1 + (1-\lambda)q_2| = \lambda|q_1| + (1-\lambda)|q_2| \qquad \forall \lambda \in [0,1].$$

This is possible only if $q_2 = \alpha q_1$ for some $\alpha > 0$. But then (5.27), which is not homogeneous in q, cannot be satisfied both by q_1 and by q_2. □

The differentiability of T along any optimal trajectory follows as a corollary.

PROPOSITION 5.11. *Let $y(\cdot) = y(\cdot; x, u)$ be a time optimal trajectory for a point $x \in \mathcal{R} \setminus \overline{B}_R$. Then T is differentiable at any point of the form $y(t)$ for $t \in]0, T(x)[$.*

Proof. The two previous lemmas show that $D^+T(y(t))$ is a singleton for any $t \in]0, T(x)[$. On the other hand, a semiconcave function is differentiable at any point where its superdifferential is a singleton. Since T is semiconcave, the conclusion follows. □

From the above proposition we immediately obtain the following corollary.

COROLLARY 5.12. *Let $y(\cdot) = y(\cdot; x, u)$ be a time optimal trajectory for a point $x \in \mathcal{R} \setminus \overline{B}_R$ and p be the corresponding dual arc. Then*

$$p(0) \in D_*T(x).$$

Corollary 5.12, together with the following result, completes Theorem 5.8, giving a characterization of all time optimal trajectories starting at singular points of T.

THEOREM 5.13. *Let $x \in \mathcal{R} \setminus \overline{B}_R$. Then, for any $q \in D_*T(x)$, there exists a unique solution (y, p) of the Cauchy problem*

$$\begin{cases} y'(t) = Ay(t) + f(y(t)) - r\dfrac{p(t)}{|p(t)|}, & y(0) = x, \\ \\ p'(t) = -Ap(t) - (\delta f(y(t)))^* p(t), & p(0) = q. \end{cases} \tag{5.28}$$

Moreover, $y(\cdot)$ is a time optimal trajectory for x, and $p(\cdot)$ is the associated dual arc.

Proof. Given $q \in D_*T(x)$, let $\{x_n\}$ be a sequence of differentiability points for T such that $x_n \to x$ and $DT(x_n) \rightharpoonup q$. For any n, let us denote by $y_n(\cdot)$ the unique optimal trajectory for x_n and by $p_n(\cdot)$ the corresponding dual arc. By Theorem 5.3, we have that $p_n(0) = DT(x_n)$. By Lemma 5.7 we obtain that, after possibly extracting a subsequence, $y_n(\cdot) \to y(\cdot)$ and $p_n(\cdot) \to p(\cdot)$ pointwise, where $y(\cdot)$ is an optimal trajectory for x and $p(\cdot)$ is the associated dual arc. In particular, $p(0) = \lim p_n(0) = q$. Moreover, recalling the Hamiltonian form of the maximum principle (5.14), we conclude that (y, p) is a solution of (5.28). Finally, (y, p) is the unique solution of this problem in light of Proposition 5.6. □

The above proof shows that, if $q \in D_*T(x)$ for some x, then there exists $x_n \to x$ such that $DT(x_n) \to q$ in the strong topology, not only in the weak topology as required by the definition of D_*T. We remark that such a property would not hold without the compactness assumption on e^{tA} in (H7).

We will conclude this section with a propagation result for the singular set of T. For this purpose we need a lemma showing that the Hamilton–Jacobi–Bellman equation (5.24) is satisfied on a dense set.

LEMMA 5.14. *Let $\theta \in]0,1[$ and $x \in D((-A)^\theta) \cap (\mathcal{R} \setminus \overline{B}_R)$. Then, for all $p \in D_*T(x)$,*

$$\langle (-A)^{1-\theta} p, (-A)^\theta x \rangle - \langle p, f(x) \rangle + r|p| = 1. \tag{5.29}$$

Proof. Given $q \in D_*T(x)$, Theorem 5.13 yields the existence of a trajectory $y(\cdot) = y(\cdot; x, u)$, time optimal for x, such that the associated dual arc $p(\cdot)$ satisfies $p(0) = q$. Then,

$$\begin{aligned}
&|(-A)^\theta(y(t) - x)| \\
&= \left|(-A)^\theta \left((e^{tA} - I)x + \int_0^t e^{(t-s)A}(f(y(s)) + u(s))ds\right)\right| \\
&\le \left|(e^{tA} - I)(-A)^\theta x\right| + \int_0^t \frac{M_\theta}{(t-s)^\theta}|f(y(s)) + u(s)|ds \to 0
\end{aligned}$$

as $t \to 0$. Analogously, we find that

$$\begin{aligned}
|(-A)^{1-\theta}(p(t) - q)| \le &\left|\left(e^{(T-t)A} - e^{TA}\right)(-A)^{1-\theta}p(T)\right| \\
&+ \int_0^t |(-A)^{1-\theta}e^{sA}\delta f(y(s))^* p(s)|\, ds \\
&+ \int_t^T \left|(-A)^{1-\theta}\left(e^{(s-t)A} - e^{sA}\right)\delta f(y(s))^* p(s)\right| ds \to 0
\end{aligned}$$

as $t \to 0$. In addition, by Theorem 5.3 and Lemma 5.9, we have that

$$\langle(-A)^{1-\theta}p(t), (-A)^\theta y(t)\rangle - \langle p(t), f(y(t))\rangle + r|p(t)| = 1$$

for any $t > 0$. Letting $t \to 0$, we obtain the conclusion. □

We are now in a position to prove the aforementioned propagation result.

THEOREM 5.15. *Let $x_0 \in \Sigma(T)$. If x_0 belongs to $D((-A)^\theta)$ for some $\theta \in]0, 1]$, then x_0 is propagation point of $\Sigma(T)$.*

Proof. From Theorem 4.3 we know that T belongs to the class $SC^1(\mathcal{R} \setminus \overline{B}_R)$. We claim that, under the above assumptions, $D_*T(x_0)$ is a proper subset of $D^+T(x_0)$, i.e., assumption (2.6) of Theorem 2.6 holds. Let us argue by contradiction and suppose that $D^+T(x_0) = D_*T(x_0)$. Then, since $D^+T(x_0)$ is a convex set, we deduce, from Lemmas 5.10 and 5.14, that $D^+T(x_0)$ is a singleton. But this is impossible, since $x_0 \in \Sigma(T)$.

Having fixed $p_0 \in D^+T(x_0) \setminus D_*T(x_0)$, it remains to exhibit a vector q satisfying condition (2.7) of Theorem 2.6. For this purpose it suffices to take any $q \in X \setminus D((-A)^\theta)$. In fact, since $D^+T(x_0) \subset D((-A)^\theta), \theta \in [0, 1[$, we immediately obtain that $p_0 + tq \notin D^+T(x_0)$ for every $t \neq 0$. The proof is then completed applying Theorem 2.6. □

In this paper we include no example to show that the minimum time function may well possess a nonempty singular set. Such examples are well known in the literature: a typical case is that of the distance function from a closed set. One may wonder, however, whether the minimum time function associated with a parabolic system like (3.14) may be singular at some point. The affirmative answer to this question follows from an example of [1].

REFERENCES

[1] P. ALBANO AND P. CANNARSA, *Singularities of the minimum time function for semilinear parabolic systems*, in Control and Partial Differential Equations (Marseille–Luminy, 1997), ESAIM Proc. 4, Soc. Math. Appl. Indust., Paris, 1998, pp. 59–72.

[2] P. Albano and P. Cannarsa, *Singularities of semiconcave functions in Banach spaces*, in Stochastic Analysis, Control, Optimization and Applications, W. M. McEneaney, G. G. Yin, and Q. Zhang, eds., Birkhäuser, Boston, 1999, pp. 171–190.
[3] G. Alberti, L. Ambrosio, and P. Cannarsa, *On the singularities of convex functions*, Manuscripta Math., 76 (1992), pp. 421–435.
[4] A. Ambrosetti and G. Prodi, *A Primer of Nonlinear Analysis*, Cambridge University Press, Cambridge, UK, 1993.
[5] A. V. Balakrishnan, *Optimal control problems in Banach spaces*, SIAM J. Control, 3 (1965), pp. 152–180.
[6] V. Barbu, *The dynamic programming equation for the time-optimal control problem in infinite dimensions*, SIAM J. Control Optim., 29 (1991), pp. 445–456.
[7] V. Barbu, *Analysis and Control of Nonlinear Infinite Dimensional Systems*, Academic Press, Boston, 1993.
[8] P. Cannarsa, *Regularity properties of solutions to Hamilton Jacobi equations in infinite dimensions and nonlinear optimal control*, Differential Integral Equations, 2 (1989), pp. 479–493.
[9] P. Cannarsa, H. Frankowska, and C. Sinestrari, *Optimality conditions and synthesis for the minimum time problem*, J. Math. Systems Estim. Control, 8 (1998), pp. 123–126.
[10] P. Cannarsa and F. Gozzi, *On the smoothness of the value function along optimal trajectories*, in Boundary Control and Boundary Variation, J. P. Zolésio, ed., Lecture Notes in Control and Inform. Sci. 178, Springer–Verlag, Berlin, 1992, pp. 60–81.
[11] P. Cannarsa and C. Sinestrari, *Convexity properties of the minimum time function*, Calc. Var., 3 (1995), pp. 273–298.
[12] P. Cannarsa and C. Sinestrari, *An infinite dimensional time optimal control problem*, in Optimization Methods in Partial Differential Equations, S. Cox and I. Lasiecka, eds., Contemp. Math. 209, Amer. Math. Soc., 1997, pp. 29–41.
[13] P. Cannarsa and M. E. Tessitore, *Optimality conditions for boundary control problems of parabolic type*, in Control and Estimation of Distributed Parameter Systems: Nonlinear Phenomena (Vorau, 1993), Internat. Ser. Numer. Math. 118, Birkhäuser, Basel, 1994, pp. 79–96.
[14] O. Cârja, *On the minimal time function for distributed control systems in Banach spaces*, J. Optim. Theory Appl., 44 (1984), pp. 397–406.
[15] O. Cârja, *On continuity of the minimal time function for distributed control systems*, Boll. Un. Mat. Ital., 6 (1985), pp. 293–302.
[16] R. Conti, *Time-optimal solution of a linear evolution equation in Banach spaces*, J. Optim. Theory Appl., 2 (1968), pp. 277–284.
[17] Y. V. Egorov, *Optimal control in Banach spaces*, Dokl. Acad. Nauk SSSR, 150 (1963), pp. 241–244 (in Russian).
[18] Y. V. Egorov, *Certain problems in optimal control theory*, USSR Comput. Math. & Phis., 3 (1963), pp. 1209–1232 (in Russian).
[19] H. Federer, *Geometric Measure Theory*, Springer-Verlag, New York, 1969.
[20] H.O. Fattorini, *Time-optimal control of solutions of operational differential equations*, SIAM J. Control, 2 (1964), pp. 54–59.
[21] H.O. Fattorini, *The time-optimal control problems in Banach spaces*, Appl. Math. Optim., 1 (1974), pp. 163–188.
[22] A. Friedman, *Optimal control in Banach spaces with fixed end-points*, J. Math. Anal. Appl., 24 (1968), pp. 161–181.
[23] X. Li and J. Yong, *Optimal Control Theory for Infinite Dimensional Systems*, Birkhäuser, Boston, 1995.
[24] J. L. Lions, *Optimal Control of Systems Governed by Partial Differential Equations*, Springer-Verlag, Berlin, 1971.
[25] J. L. Lions and B. Malgrange, *Sur l'unicité rétrograde dans les problèmes mixtes paraboliques*, Math. Scand., 8 (1960), pp. 277–286.
[26] P. L. Lions, *Viscosity solutions of fully nonlinear second-order equations and optimal stochastic control in infinite dimensions. Part I: The case of bounded stochastic evolutions*, Acta Math., 161 (1988), pp. 243–278.
[27] A. Pazy, *Semigroups of Linear Operators and Applications to Partial Differential Equations*, Springer–Verlag, New York, 1983.
[28] D. Preiss, *Differentiability of Lipschitz functions on Banach spaces*, J. Funct. Anal., 91 (1990), pp. 312–345.

SIAM J. CONTROL OPTIM.
Vol. 38, No. 3, pp. 947–970

HOFFMAN'S ERROR BOUND, LOCAL CONTROLLABILITY, AND SENSITIVITY ANALYSIS*

ABDERRAHIM JOURANI†

Abstract. Our aim is to present sufficient conditions ensuring Hoffman's error bound for lower semicontinuous nonconvex inequality systems and to analyze its impact on the local controllability, implicit function theorem for (non-Lipschitz) multivalued mappings, generalized equations (variational inequalities), and sensitivity analysis and on other problems like Lipschitzian properties of polyhedral multivalued mappings as well as weak sharp minima or linear conditioning. We show how the information about our sufficient conditions can be used to provide a computable constant such that Hoffman's error bound holds. We also show that this error bound is nothing but the classical Farkas lemma for linear inequality systems. In the latter case our constant may be computed explicitly.

Key words. subdifferentials, Hoffman's bound, implicit function theorem, generalized equations, controllability, sensitivity analysis

AMS subject classifications. Primary, 46A30, 90C31, 49K24; Secondary, 49J52

PII. S0363012998339216

1. Introduction.

Consider the inequality system

$$f(x) \leq 0, \tag{1.1}$$

where $f : X \to R \cup \{\infty\}$ is an extended real-valued lower semicontinuous function and X is a Banach space.

Let S be a set of solutions to (1.1). Hoffman's error bound holds (globally) for (1.1) if there exists $a > 0$ such that

$$d(x, S) \leq a f_+(x) \quad \forall x \in X, \tag{1.2}$$

where $d(x, S) = \inf_{u \in S} \|x - u\|$, $f_+(x) = \max(f(x), 0)$ and $\| \ \|$ denotes the norm on X.

Hoffman's result [11] states that if f is a maximum of a finite number of affine functions in R^n, then (1.2) holds.

Various extensions in finite dimension of Hoffman's result were obtained for general convex inequality systems (see, for example, [36], [37], [12], [34], [35], [56]). For other related error bound results, see [4], [38], [13], [49]. Robinson [49] showed that in infinite dimension the error bound (1.2) holds for any convex differentiable inequality system which satisfies the Slater constraint qualification condition and where S is bounded. In [13], Ioffe gave a local version of (1.2) for nonconvex and nondifferentiable Lipschitz inequality system in infinite dimension. He used his result to obtain a new proof of Hoffman's inequality in linear programming in infinite dimension. Recently, Deng [9] studied any system of convex inequality in a reflexive Banach space which has an unbounded solution set.

In this paper we study the parametric inequality systems

$$f(x, y) \leq 0, \tag{1.1$'$}$$

*Received by the editors May 21, 1998; accepted for publication (in revised form) June 15, 1999; published electronically March 15, 2000.
http://www.siam.org/journals/sicon/38-3/33921.html
†Département de Mathématiques, Analyse Appliquée et Optimisation, BP 47 870, 21078 Dijon Cedex, France (jourani@u-bourgogne.fr).

where $f : X \times Y \to R \cup \{\infty\}$ is an extended real-valued lower semicontinuous function and Y is a Banach space.

For each y we let $S(y)$ be a set of solutions to (1.1′). We are concerned with the following analogous Hoffman's error bound: there exists $a > 0$ such that

$$d(x, S(y)) \leq a f_+(x, y) \quad \forall x \in X, \quad \forall y \in Y. \tag{1.2′}$$

Our aim is to present sufficient conditions ensuring (1.2′) for lower semicontinuous nonconvex inequality systems and to analyze its impact on the local controllability, sensitivity analysis, implicit function theorem for non-Lipschitz multivalued mappings, and generalized equations (variational inequalities) and on other problems like Lipschitzian properties of polyhedral multivalued mappings as well as weak sharp minima or linear conditioning. As in [9] and [13] we can show how the information about our sufficient condition can be used to provide a computable constant a such that (1.2′) holds. We also show that relation (1.2) is nothing but the classical Farkas lemma for linear inequality systems. In the latter case our constant may be computed explicitly. Note that the present paper is an infinite dimensional extension of [22] and a continuation of [18], published in *Mathematical Programming* in 1994, in which I indicate (see the end of section 3 of [18]) that we may use the partial approximate subdifferential (which works in any Banach spaces and more and the proposed proof works for any presubdifferential [55] as is noted in [24, Theorem 6.1]) and introduce similar regularity conditions to ensure metric regularity of systems defined by inclusions.

Our notation is basically standard. For any Banach space X and its topological dual X^* we denote by B_X and B_{X^*} their closed unit balls. As usual, $\mathrm{dom} f$ and $\mathrm{epi} f$ of an arbitrary extended real-valued function f stand for the domain and the epigraph

$$\begin{aligned} \mathrm{dom} f &= \{x : \quad f(x) < +\infty\}, \\ \mathrm{epi} f &= \{(x, r) : \quad f(x) \leq r\}. \end{aligned}$$

2. Error bound: The nonconvex case. In this section we present sufficient condition in terms of an abstract subdifferential. The partial subdifferential (as in [55]), in x with respect to y, on X is any operator ∂_x which satisfies the following properties.

For any lower semicontinuous function $f : X \times Y \to R \cup \{\infty\}$, any locally Lipschitz function $g : X \times Y \to R \cup \{\infty\}$, any $x \in X$, and any $y \in Y$,

(P_1) $\partial_x f(x, y) \subset X^*$ and $\partial_x f(x, y) = \emptyset$ if $f(x, y) = \infty$;

(P_2) $\partial_x g(x, y)$ coincides with the partial subdifferential in the sense of convex analysis whenever $g(\cdot, y)$ is convex, that is,

$$\partial_x g(x, y) = \{x^* \in X^* : \langle x^*, u - x \rangle \leq g(u, y) - g(x, y) \quad \forall u \in X\};$$

(P_3) $0 \in \partial_x f(x, y)$ whenever x is a local minimum for f with respect to y;

(P_4) $\partial_x f(x, y) = \partial_x w(x, y)$ whenever f and w coincide around (x, y);

(P_5)

$$\partial_x (f + g)(x, y) \subset \partial_x f(x, y) + \partial_x g(x, y).$$

In the case where $f(x, y) = f(x)$ for all $(x, y) \in X \times Y$, our subdifferential will be denoted by $\partial f(x)$.

Before stating our main results of this section, let us give some examples of partial subdifferentials.

Example 2.1 (partial limiting Fréchet subdifferential). Let X be an *Asplund space*, i.e., a Banach space on which every continuous convex function is Fréchet differentiable at a dense set of points. We refer the reader, for example, to the papers by Phelps [48]. The partial Fréchet ε-subdifferential of f at x with respect to y is the set

$$\partial_x^{\varepsilon} f(x,y) = \left\{ x^* \in X^* : \liminf_{h \to 0} \frac{f(x+h,y) - f(x,y) - \langle x^*, h \rangle}{\| h \|} \geq -\varepsilon \right\}$$

if $(x,y) \in \mathrm{dom} f$ and $\partial_x^{\varepsilon} f(x,y) = \emptyset$ if $(x,y) \notin \mathrm{dom} f$. *The partial limiting Fréchet subdifferential* $\partial_x^F f(x,y)$, which is first introduced in finite dimension by Jourani and Thibault [22], of f at x with respect to y is the set

$$\partial_x^F f(x,y) = \text{seq-} \limsup_{\substack{(u,v) \xrightarrow{f} (x,y) \\ \varepsilon \longrightarrow 0^+}} \partial_x^{\varepsilon} f(u,v).$$

In the case where $f(x,y) = f(x)$, this subdifferential (Kruger–Mordukhovich [30]) is an infinite dimensional extension of the nonconvex construction by Mordukhovich [39], [40] (see also [44]).

We may show that the partial limiting Fréchet subdifferential has the same properties as the limiting Fréchet subdifferential (see Kruger–Mordukhovich [31] and Kruger [29]) and hence it satisfies properties (P_1)–(P_5).

Example 2.2 (partial approximate subdifferential). The partial Dini subdifferential of f at x with respect to y is the set

$$\partial_x^- f(x,y) = \left\{ x^* \in X^* : \langle x^*, h \rangle \leq \liminf_{\substack{t \to 0^+ \\ u \to h}} t^{-1}(f(x+tu,y) - f(x,y)) \quad \forall h \in X \right\}$$

if $(x,y) \in \mathrm{dom} f$ and $\partial_x^- f(x,y) = \emptyset$ if $(x,y) \notin \mathrm{dom} f$.

The partial approximate subdifferential of f at x with respect to y is the set

$$\partial_x^A f(x,y) = \bigcap_{L \in \mathcal{F}(X)} \limsup_{(u,v) \xrightarrow{f} (x,y)} \partial_x^- f_{u+L}(u,v),$$

where $f_S(x,y) = f(x,y)$ if $x \in S$ and $f_S(x,y) = +\infty$ otherwise and $\mathcal{F}(X)$ denotes the collection of all finite dimensional subspaces of X.

In the case where $f(x,y) = f(x)$, this subdifferential (Ioffe [14], [15]) is an infinite dimensional extension of the nonconvex construction by Mordukhovich [39], [40]. In fact, Ioffe [16] showed that in finite dimensional spaces the approximate subdifferential and the limiting Fréchet subdifferentials coincide. Ioffe [14] used this representation to introduce the approximate subdifferential in infinite dimensional spaces. We have to note that the subdifferentials in Examples 2.1 and 2.2 are generally different in infinite dimensional case.

We may show that the partial approximate subdifferential has the same properties as the approximate subdifferential (see Ioffe [14], [15]) and hence it satisfies properties (P_1)–(P_5).

Example 2.3 (partial Clarke's subdifferential). Let f be locally Lipschitzian around (x,y). The partial Clarke's generalized directional derivative of f at x with respect to y in the direction $h \in X$ is given by

$$d_x^0 f((x,y);h) = \limsup_{\substack{(u,v) \longrightarrow (x,y) \\ t \to 0^+}} \frac{f(u+th,v) - f(u,v)}{t}.$$

The partial Clarke's subdifferential of f at x with respect to y is the set

$$\partial_x^C f(x,y) = \{x^* \in X^* : \quad \langle x^*, h\rangle \le d_x^0 f((x,y);h) \quad \forall h \in X\}.$$

We may show that the partial Clarke's subdifferential has the same properties as the Clarke's subdifferential (see Clarke [6]) and hence it satisfies properties (P_1)–(P_5) for locally Lipschitzian functions.

All these partial subdifferentials coincide, for a continuous convex function f, with the partial subdifferential in the sense of convex analysis

$$\partial_x f(x,y) = \{x^* \in X^* : \quad \langle x^*, h-x\rangle \le f(h,y) - f(x,y) \quad \forall h \in X\}.$$

We suppose in the following that our partial subdifferential ∂ satisfies (P_1)–(P_5).

Our proof is similar to that given by Ioffe in [13] in the local situation and in the case where f is locally Lipschitzian. We begin with the global error bound result.

THEOREM 2.4. *Let $f : X \times Y \to R \cup \{\infty\}$ be an extended real-valued function such that for each $y \in Y$, $f(\cdot, y)$ is lower semicontinuous. Suppose that*

$$\text{(2.1)} \qquad \text{there exists } a > 0 \quad \forall x \in X, \quad \forall y \notin S^{-1}(x), \quad d(0, \partial_x f(x,y)) \ge \frac{1}{a}.$$

Then (1.2′) *holds.*

Proof. Suppose that our relation (1.2′) is not valid. Then there exists $x' \in X$ and $y' \in Y$ such that

$$\text{(2.2)} \qquad d(x', S(y')) > a f_+(x', y').$$

Note that $x' \notin S(y')$. Then $f_+(x', y') > 0$. Set $\varepsilon = f(x', y')$ and $\lambda = (a+\alpha) f(x', y')$ where $\alpha > 0$ is such that $\lambda < d(x', S(y'))$. Then

$$f_+(x', y') \le \inf_{x \in X} f_+(x, y') + \varepsilon.$$

By the lower semicontinuity of $f(\cdot, y')$, the Ekeland's variational principle ensures the existence of $x \in X$ satisfying

$$\text{(2.3)} \qquad \|x - x'\| \le \lambda$$

$$\text{(2.4)} \qquad f_+(x, y') \le f_+(u, y') + \frac{\varepsilon}{\lambda}\|u - x\| \quad \forall u \in X.$$

Note that, by (2.2)–(2.3), $x \notin S(y')$. Since $f(\cdot, y')$ is lower semicontinuous, it coincides with $f_+(\cdot, y')$ in a neighborhood of x and hence by (2.4) and properties (P_1)–(P_5) we get

$$0 \in \partial_x f(x, y') + \frac{1}{(a+\alpha)} B_{X^*}$$

and this contradicts relation (2.1). □

Remark. Theorem 2.4 has been established by Clarke [5] (see also [7]) in Hilbert space using the decrease principle.

The proof of the following local result is similar to the previous one. It extends the result by Ioffe [13] to the non-Lipschitz case.

THEOREM 2.5. *Let $f : X \times Y \to R \cup \{\infty\}$ be an extended real-valued function such that for each $y \in Y$, $f(\cdot, y)$ is lower semicontinuous and let $\overline{x} \in S(\overline{y})$ for some $\overline{y} \in Y$. Suppose that there exist $r > 0$ and $a > 0$ such that*

$$(2.5) \qquad \forall x \in \overline{x} + rB_X,\ \forall y \notin S^{-1}(x) \text{ with } y \in \overline{y} + rB_Y, \quad d(0, \partial_x f(x,y)) \geq \frac{1}{a}.$$

Then

$$d(x, S(y)) \leq a f_+(x,y) \quad \forall x \in \overline{x} + \frac{r}{2}B_X,\ \forall y \in \overline{y} + rB_Y.$$

Remark. Note that Theorems 2.4 and 2.5 remain valid if Y is a metric space.

In the following corollary we give a sufficient condition for the solvability of inclusions: Given $(\overline{x}, \overline{y})$ with $\overline{y} \in F(\overline{x})$, find for each y in some neighborhood of $\overline{y}$ a point $x(y)$ near $\overline{x}$, solution of the inclusion

$$y \in F(x),$$

where $F : X \to 2^Y$ is a multivalued mapping.

COROLLARY 2.6. *Let $F : X \to 2^Y$ be a multivalued mapping with closed graph GrF containing $(\overline{x}, \overline{y})$. Suppose that*

$$\text{there exist } r > 0, a > 0 \quad \forall x \in \overline{x} + rB_X,\ \forall y \notin F(x) \text{ with } y \in \overline{y} + rB_Y,$$
$$d(0, \partial_x d(x, y; GrF)) \geq \frac{1}{a}.$$

Then

$$d(x, F^{-1}(y)) \leq a d(x, y; GrF) \quad \forall x \in \overline{x} + \frac{r}{2}B_X,\ \forall y \in \overline{y} + rB_Y.$$

Proof. Set $f(x,y) = d(x, y; \mathrm{Gr}F)$ and apply Theorem 2.5.

The following corollary gives us different characterizations of our sufficient condition in Corollary 2.6 for partial limiting Fréchet subdifferential in finite dimension.

COROLLARY 2.7. *Let X and Y be finite dimensional spaces and let $F : X \to 2^Y$ be a multivalued mapping with closed graph GrF containing $(\overline{x}, \overline{y})$. Then the following assertions are equivalent:*

(1) *if $(0, y^*) \in \partial^F d(\overline{x}, \overline{y}; GrF)$, then $y^* = 0$;*

(2) *there exist $r > 0$ and $a > 0$ $\forall x \in \overline{x} + rB_X$, $\forall y \notin F(x)$, with $y \in \overline{y} + rB_Y$,*

$$d(0, \partial_x^F d(x, y; GrF)) \geq \frac{1}{a};$$

(3) *(Graphical metric regularity) there exist $r > 0$ and $a > 0$ such that*

$$d(x, F^{-1}(y)) \leq a d(x, y; GrF) \quad \forall x \in \overline{x} + rB_X,\ \forall y \in \overline{y} + rB_Y;$$

(4) *(Metric regularity) there exist $r > 0$ and $a > 0$ such that*

$$d(x, F^{-1}(y)) \leq a d(y, F(x)) \quad \forall x \in \overline{x} + rB_X,\ \forall y \in \overline{y} + rB_Y.$$

Proof. (2) $\Longrightarrow$ (3) follows from Corollary 2.6.

(3) $\Longrightarrow$ (4) is obvious.

(4) $\Longrightarrow$ (1): see Mordukhovich [41], [42] and Jourani [18].

The proof of the implication (1) $\Longrightarrow$ (2) is similar to that of Theorem 3.14 in Jourani [18]. □

Remarks. (1) Recently Thibault [53] showed (by a direct method) that the equivalence (3) $\Longleftrightarrow$ (4) holds in any normed spaces.

(2) We may show that the corollary holds in infinite dimensional spaces for a special class of multivalued mappings (see for some applications of this type of result the papers by Ioffe [17], Mordukhovich and Shao [45], [46], and Jourani and Thibault [23], [24], [25], [26], [27], [18], [19], [20], [21]).

Suppose in addition to (P_1)–(P_5) that our subdifferential satisfies the following closedness assumption:

$$(P_6) \qquad \partial_x f(x,y) = \limsup_{\substack{(u,v)\to(x,y)\\ f(u,v)\to f(x,y)}} \partial_x f(u,v)$$

and

$$\partial_x f(x,y) = seq- \limsup_{\substack{(u,v)\to(x,y)\\ f(u,v)\to f(x,y)}} \partial_x f(u,v)$$

if X is Asplund.

Remarks. (1) Property (P_6) is satisfied for partial limiting Fréchet subdifferential (in Asplund spaces) and for partial approximate subdifferential for any lower semicontinuous function f and for partial Clarke's subdifferential for any locally Lipschitz function f.

(2) With these properties, our subdifferential satisfies the following inclusions for any locally Lipschitz function f around x uniformly in y (in some neighborhood V_y of y):

$$\partial_x^F f(x,y) \subset \partial_x^A f(x,y) \subset \partial_x f(x,y).$$

COROLLARY 2.8. *Let f be as in Theorem 2.5 and let $\overline{x} \in S(\overline{y})$ for some $\overline{y} \in Y$. Suppose that f is continuous around $(\overline{x},\overline{y})$, (P_6) is satisfied, and*

$$0 \notin \partial_x f(\overline{x},\overline{y}). \tag{2.6}$$

Then there exist $r > 0$ and $a > 0$ such that

$$d(x, S(y)) \le a f_+(x,y) \quad \forall x \in \overline{x} + rB_X,\ \forall y \in \overline{y} + rB_Y. \tag{2.7}$$

Proof. It suffices to show that (2.6) implies (2.5) and to apply Theorem 2.5. Indeed suppose that (2.5) is false. Then there are sequences $((x_n,y_n))$ and (x_n^*) such that $x_n \notin S(y_n)$ and $x_n^* \in \partial_x f(x_n,y_n)$ with $(x_n,y_n) \to (\overline{x},\overline{y})$ and $x_n^* \to 0$. Thus, by (P_6), $0 \in \partial_x f(\overline{x},\overline{y})$ and this contradiction completes the proof. □

3. Error bound: The convex case. We start this section by giving a characterization of Hoffman's error bound (see the excellent paper by Burke and Ferris [3]).

PROPOSITION 3.1. *Suppose X is a Hilbert space and f is convex and proper. If S is closed, then the following are equivalent:*

(i) *Relation* (1.2) *holds;*

(ii) $\partial d(S,x) \subset a\partial f_+(x) \quad \forall x \in S.$

In the following proposition we present a classical condition ensuring relation (1.2) which also can be obtained as a consequence of the results in [50] and [51].

PROPOSITION 3.2. *Suppose f is convex and $f(x_0) < 0$. Then for all $x \in X$*

$$d(x, S) \leq \frac{f_+(x)}{-f(x_0)} \|x - x_0\|.$$

If in addition there exists a nonempty set $C \subset X$, $\alpha > 0$, and $\gamma \geq 1$ such that

$$f(x) - f(x_0) \geq \alpha \|x - x_0\|^\gamma \quad \forall x \in C,$$

then

$$d(x, S) \leq \frac{f_+(x)}{\alpha} \quad \forall x \in C.$$

Let $g : X \to R \cup \{\infty\}$ be a function. The recession function g^∞ of g is defined by

$$\text{epi} g^\infty = (\text{epi} g)^\infty,$$

where $A^\infty = \{x : \quad A + x \subset A\}$. If g is convex, proper, and lower semicontinuous, then

$$g^\infty(u) = \sup_{x \in \text{dom} g} \{g(u + x) - g(x)\}.$$

The following result extends that of Deng [9], established in reflexive Banach spaces, to general Banach spaces.

THEOREM 3.3. *Let $g : X \to R \cup \{\infty\}$ be a convex, proper, and lower semicontinuous function and C be a closed convex subset of X such that $\text{dom} g \cap C \neq \emptyset$. Consider the solution set*

$$S' = \{x \in C : \quad g(x) \leq 0\}.$$

Suppose that

(i) *for all $\overline{x} \in \text{dom} g \cap C$, $\overline{x} \notin S'$,*

$$\partial(g + \Psi_C)(\overline{x}) = \partial g(\overline{x}) + N(C, \overline{x}),$$

where Ψ_C denotes the indicator function of C;

(ii) *there exist $a > 0$ and $\hat{x} \in C^\infty$ with $\|\hat{x}\| = 1$ such that*

$$g^\infty(\hat{x}) \leq \frac{-1}{a}.$$

Then

$$d(x, S') \leq a g_+(x) \quad \forall x \in C.$$

Proof. Put $f = g + \Psi_C$. It is an easy exercise to show that (i) and (ii) imply (2.1) and Theorem 2.4 completes the proof. □

Remarks. (1) In [9], Deng assumed that the function g is continuous and this is crucial in his proof.

(2) Deng [9] established this result in reflexive Banach space and for a finite number of inequality

$$g_i(x) \leq 0, \quad i = 1, \ldots, m.$$

But it suffices to put $g(x) = \max\{g_i(x) : \ i = 1, \ldots, m\}$ and it is easy to show that the assertion (ii) is satisfied.

(3) We know that the formula in (i) is valid under some constraint qualifications and there are many constraint qualifications in the literature ensuring this relation. It is valid, for example, when g is continuous.

4. Error bound: The linear case. In this section we are concerned with the linear equality-inequality systems

$$Ax = 0, \quad \langle x_i^*, x\rangle + b_i \le 0, \quad i = 1, \dots, m, \tag{4.1}$$

where $A : X \to Y$ is a linear continuous mapping such that $R(A)$, the rank of A, is closed, Y is a Banach space, $b_i \in R$, and $x_i^* \in X^*$ with $\|x_i^*\| = 1$.

We begin by showing that in the case of linear inequality systems, Hoffman's error bound is equivalent to the classical Farkas lemma. To do this, some notations are needed. We set

$\Delta_m = \{1, \dots, m\}$,
$I_m = \{E \subset \Delta_m : \quad (x_i^*)_{i\in E}$ are linearly independent $\}$,
$\forall E \in I_m$, $S_E = \cap_{i\in E} S_i$ and $f_E(x) = \sum_{i\in E} d(x, S_i)$ and
$\forall i \in \Delta_m$, $S_i = \{x \in H : \quad \langle x_i^*, x\rangle + b_i \le 0\}$.

THEOREM 4.1. *Consider the linear inequality system*

$$\langle x_i^*, x\rangle + b_i \le 0, \quad i = 1, \dots, m$$

with solution set S, $x_i^* \in H$, where H is a Hilbert space. Then the two following properties hold and are equivalent:

(i) *there exists $\alpha > 0$ depending only on $(x_i^*)_{i\in\Delta_m}$ such that*

$$d(x, S) \le \alpha f(x) \quad \forall x \in X.$$

(ii) *(Farkas lemma) for all u in S, $N(S, u) = R_+\partial f(u)$, where*

$$f(x) = \sum_{i=1}^{m} d(x, S_i) \qquad \text{and} \qquad N(S, u) = R_+\partial d(u, S).$$

Proof. The implication (i) $\Longrightarrow$ (ii) is obvious.

We establish the reverse implication in three steps.

First step. Let $E \in I_m$. Since $(x_i^*)_{i\in E}$ are linearly independent,

$$\min\left\{\left\|\sum_{i\in E}\lambda_i x_i^*\right\| : \quad 1 \ge \lambda_i \ge 0, i \in E, \sum_{i\in E}\lambda_i \ge 1\right\} > 0,$$

and hence for all $x \notin S_E$

$$d(0, \partial f_E(x)) \ge \frac{1}{\alpha_E},$$

where

$$\frac{1}{\alpha_E} = \min\left\{\left\|\sum_{i\in E}\lambda_i x_i^*\right\| : \quad 1 \ge \lambda_i \ge 0, i \in E, \sum_{i\in E}\lambda_i \ge 1\right\}.$$

So, by Theorem 2.4,

$$d(x, S_E) \le \alpha_E f_E(x) \quad \forall x \in H.$$

Second step. We will show that for each $x \notin S$ there exists $E \in I_m$ such that

$$d(x, S) = d(x, S_E).$$

Indeed, let $x \notin S$; then there exists u in S such that

$$d(x, S) = \|x - u\|$$

or, equivalently,

$$x - u \in N(S, u).$$

By (ii), $x-u \in R_+ \partial f(u)$ and as $\partial d(u, S_i) \subset [0, 1]x_i^*$, for $i \in \Delta_m$, there exist $1 \geq \lambda_i \geq 0$, $i \in \Delta_m$, not all equal to zero, such that

$$x - u = \sum_{i \in \Delta_m} \lambda_i x_i^*.$$

If $(x_i^*)_{i \in \Delta_m}$ are linearly independent the result follows from the first step. So suppose there exist $\mu_i \in R$, $i \in \Delta_m$, not all equal to zero such that

$$\sum_{i \in \Delta_m} \mu_i x_i^* = 0.$$

Hence for all $t \in R$

$$\sum_{i \in \Delta_m} (\lambda_i + t\mu_i) x_i^* = x - u.$$

Our problem is to find $t \geq 0$ and $i_0 \in \Delta_m$ such that $\lambda_{i_0} + t\mu_{i_0} = 0$ and $\lambda_i + t\mu_i \geq 0$ for $i \neq i_0$. Set $J = \{i \in \Delta_m : \quad \mu_i < 0\}$ and suppose that $J \neq \emptyset$. Set $t = \min_{i \notin J} \frac{-\lambda_i}{\mu_i}$. Then there exists $i_0 \in \Delta_m$ such that

$$\lambda_{i_0} + t\mu_{i_0} = 0 \quad \text{and} \quad \lambda_i + t\mu_i \geq 0 \quad \forall i \neq i_0.$$

By induction we show that $x - u$ is a positive combination of the linearly independent family of $(x_i^*)_{i \in E}$ and hence $x - u \in N(S_E, u)$, or equivalently, $\|x - u\| = d(x, S_E)$.

Third step. By the second step we have for each $x \notin S$ the existence of $E \in I_m$ such that

$$d(x, S) = d(x, S_E),$$

and by the first step we have

$$d(x, S) \leq \alpha_E f_E(x) \leq \alpha_E f(x).$$

Thus the proof is complete by taking

$$\alpha = \max_{E \in I_m} \alpha_E. \qquad \square$$

Remark. One of the referees pointed out that when the dimension of H is finite, a general result was proved by D. Klatte [28].

We use Theorem 4.1 to get a Lipschitzian property of polyhedral multivalued mappings. A multivalued mapping F from X into Y is polyhedral if its graph is the union of a finite (possibly empty) collection of polyhedral convex sets.

COROLLARY 4.2 (see [49]). *Let F be a multivalued mapping from X into Y. Suppose that Y is a Hilbert space and F is polyhedral. Then F is locally upper Lip-*

schitzian, i.e., there exists $a > 0$ such that for all $x \in X$ there exists $r > 0$ such that for all $x' \in x + rB_X$,

$$F(x') \subset F(x) + a\|x - x'\|B_Y.$$

Proof. Let $\mathrm{Gr}F$ be the graph of F. Then there are polyhedral sets $P_1, \dots, P_k \subset X \times Y$ such that $\mathrm{Gr}F = \cup\{P_i : \quad i = 1, \dots, k\}$. For each $i = 1, \dots, k$ there are $(b_{i,j})_{j=1,\dots,k_i} \subset R$, $(x^*_{i,j})_{j=1,\dots,k_i} \subset X^*$, and $(y^*_{i,j})_{j=1,\dots,k_i} \subset Y^*$ such that

$$P_i = \{(x, y) \in X \times Y : \quad \langle x^*_{i,j}, x\rangle + \langle y^*_{i,j}, y\rangle + b_{i,j} \le 0, \quad j = 1, \dots, k_i\}.$$

For each i let F_i be the multivalued mapping whose graph is P_i. for all $i = 1, \dots, k$ and $x \in X$, $F_i(x)$ is a polyhedral subset of Y.

If for all $i = 1, \dots, k$, $F_i(x)$ is empty, then there exists $r > 0$ such that for all $x' \in x + rB_X$ and for all $i = 1, \dots, k$, $F_i(x')$ are empty. So suppose that there exists i such that $F_i(x)$ is nonempty; then by Theorem 4.1, there exists $a_i > 0$, depending only on $(y^*_{i,j})_{j=1,\dots,k_i}$, such that for all $y \in Y$,

$$d(y, F_i(x)) \le a_i g^i(x, y),$$

where $g^i(x, y) = \sum_{j=1}^{k_i}(\langle x^*_{i,j}, x\rangle + \langle y^*_{i,j}, y\rangle + b_{i,j})_+$. So for all $x' \in X$ and for all $y \in F_i(x')$

$$d(y, F_i(x)) \le a_i g^i(x, y) \le a_i K_i \|x - x'\|,$$

where $K_i = \sum_{j=1}^{k_i} \|x^*_{i,j}\|$, and hence

$$F_i(x') \subset F_i(x) + a_i K_i \|x - x'\| B_Y.$$

Set $a = \max\{a_i K_i : \quad i = 1, \dots, k\}$. Then for all $x \in X$, there exists $r > 0$ such that for all $x' \in x + rB_X$

$$F(x') \subset F(x) + a\|x - x'\|B_Y. \qquad \square$$

Remark. If in Corollary 4.2 we have $F(x) \neq \emptyset$ for all x in some neighborhood V of x_0, then for all $x, x' \in V$

$$F(x') \subset F(x) + a\|x - x'\|B_Y.$$

The following result is due to Ioffe [13]. It extends that of Hoffman [11] from finite dimensional spaces to infinite dimensional ones.

THEOREM 4.3. *Let S be the solution set of the system* (4.1). *Then there exists $a > 0$ such that*

$$d(x, S) \le a f(x) \quad \forall x \in X,$$

where $f(x) = \|Ax\| + \sum_{i=1}^{m}(\langle x^*_i, x\rangle + b_i)_+$.

Proof. It suffices to show that the assumptions of Theorem 2.4 are satisfied (see Ioffe [13]). $\square$

5. Implicit function theorem and generalized equations. Let X be the product of two finite dimensional spaces U and V, and Y stands for a Banach space (in fact all the results of this section hold for Y a metric space) and let $F : U \times Y \to V$ be a multivalued mapping. Consider the inclusion

$$0 \in F(u, y).$$

We wish to solve this inclusion for u as a function of y around $(\overline{u}, \overline{y})$ at which this inclusion holds. Again our important results (Theorems 2.4 and 2.5) permit us to establish the implicit function theorem for non-Lipschitz multivalued mappings. Note that the infinite dimensional case will be studied in a forthcoming paper.

THEOREM 5.1. *Suppose that the graph GrF of the multivalued mapping F is closed and that*

$$(0, v^*) \in \partial^F_{(u,v)} h(\overline{u}, 0, \overline{y}) \quad \Longrightarrow \quad v^* = 0, \tag{5.1}$$

where $h(u, v, y) = d(u, v, GrF_y)$ and $F_y : u \to F(u, y)$. Suppose also that the multivalued mapping $y \to F_y(u)$ is pseudo-Lipschitzian around $\overline{y}$ uniformly in u in some neighborhood of $\overline{u}$, that is there exist $a > 0$ and $r > 0$ such that for all $y, y' \in \overline{y} + rB_Y$ and $u \in \overline{u} + rB_U$,

$$F_y(u) \cap rB_V \subset F_{y'}(u) + a\|y - y'\|B_V.$$

Then the multivalued mapping $(y, v) \to F_y^{-1}(v)$ is pseudo-Lipschitzian around $(\overline{y}, 0, \overline{u})$. Hence there exist $\alpha > 0$ and a Lipschitz continuous mapping $\phi : (\overline{y} + \alpha B_Y) \times \alpha B_V \to U$ with $\phi(\overline{y}, 0) = \overline{u}$ such that for all $(y, v) \in (\overline{y} + \alpha B_Y) \times \alpha B_V$,

$$v \in F(\phi(y, v), y).$$

Proof. For the first part, apply Theorem 2.5 to the function h with (u, v) playing the role of x. □

The proof of the second part follows from the following selection lemma.

LEMMA 5.2. *Let G be a multivalued mapping from some Banach space Z into R^m. Suppose that G is pseudo-Lipschitzian around $(\overline{z}, 0)$, that is, there exist $k_G > 0$ and $r > 0$ such that for all $z, z' \in \overline{z} + rB_Z$*

$$G(z) \cap rB_{R^m} \subset G(z') + k_G\|z - z'\|B_{R^m}.$$

Suppose also that for all $z \in \overline{z} + rB_Z$, $G(z)$ is closed. Then G admits a Lipschitz selection g near $\overline{z}$ with $g(\overline{z}) = 0$.

Proof. Without loss of generality we assume that $m = 1$. Set $s = r(1 + 2k_G)$ and consider the mappings $\overline{G} : \overline{z} + rB_Z \to [-s, s]$ defined by

$$\overline{G}(z) = G(z) \cap sB_R.$$

Then there exist $a > 0$ and $b > 0$ such that for all $z, z' \in \overline{z} + aB_Z$

$$\overline{G}(z) \cap bB_R \subset \overline{G}(z') \cap bB_R + k_G\|z - z'\|B_R.$$

Define the mapping $g : \overline{z} + aB_Z \to [-s, s]$ by

$$g(z) = \inf \overline{G}(z) \cap bB_R := \inf\{p : \quad p \in \overline{G}(z) \cap bB_R\}.$$

Then g is locally Lipschitzian near $\overline{z}$ and as $\overline{G}(z)$ is compact, $g(z) \in \overline{G}(z)$. □

Remarks. (1) Theorem 5.1 remains valid if Y is finite dimensional and if relation (5.1) is replaced by

$$(0, y^*, v^*) \in \partial^F d(\overline{u}, \overline{y}, 0; GrF) \quad \Longrightarrow \quad v^* = 0, \quad y^* = 0.$$

(2) The theorem is false if the uniform pseudo-Lipschitzness assumption is omitted. Indeed, consider the real-valued function F defined on R^2 by

$$F(u, y) = (\sqrt{|y|} + 1)u.$$

It is simple to see that all the hypotheses of the theorem are satisfied except the fact that the mapping

$$y \to F_y(u)$$

is not pseudo-Lipschitz around 0 uniformly in u in some neighborhood of 0 and the conclusion is that the mapping $(y, v) \to F_y^{-1}(v) = \frac{v}{\sqrt{|y|}+1}$ is not locally Lipschitz around $(0, 0)$.

(3) One of the referees pointed out that the first part of Theorem 5.1 was derived in [32] for infinite dimensional Fréchet smooth spaces and this is true. In my opinion their result is not an implicit function theorem but a metric regularity one. However, the authors of [32] may consult the paper by Jourani and Thibault [22], published in *Mathematical Programming* in 1990, established in finite dimensional spaces. In this paper we established a graphical metric regularity result but, as noted by Thibault [53], metric regularity and graphical metric regularity are equivalent. The present paper is an infinite dimensional extension of [22] and a continuation of the paper [18], published in *Mathematical Programming* in 1994, in which I indicate (see the end of section 3) that we may use the partial approximate subdifferential (which works in any Banach spaces and the proposed proof works for any subdifferential satisfying (P_1)–(P_6) as is noted in the paper [24, Theorem 6.1]) and introduce similar regularity conditions to ensure metric regularity of systems defined by inclusions. I am persuaded that the authors of the paper [32] might not be aware of the existence of the papers [22] and [18].

Now we consider the generalized equation of the form

$$0 \in f(u, y) + G(u), \tag{5.2}$$

where $f : U \times Y \to V$ is a mapping and $G : U \to V$ is a multivalued mapping of a closed graph. We are concerned with properties of the solution set

$$S(y) := \{u \in U : \quad 0 \in f(u, y) + G(u)\}$$

near a reference point. These questions are addressed to local sensitivity analysis of the generalized equation (5.2) under parameter perturbation y. There are many publications devoted to the study of the sensitivity analysis of generalized equations in forms (5.2) or in the form

$$0 \in f(u, y) + N(C, u),$$

where $N(C, u)$ is the normal cone operator in the sense of convex analysis [51], then in the latter case the generalized equation is reduced to the parametric variational inequality

$$\text{find} \quad u \in C \quad \text{such that} \quad \langle f(u, y), u - u' \rangle \geq 0 \quad \forall u' \in C.$$

For more details concerning these equations and their consequences, the reader can consult, for example, the works by Robinson [49], Mordukhovich [42], and references therein.

Most of the obtained results in the literature assume that f is differentiable. Our aim is to apply Theorem 2.5 in order to show that sensitivity analysis of these generalized equations may be established for f locally Lipschitzian in u uniformly in y around $(\overline{u}, \overline{y})$ solution of (5.2). Note that our result extends that of Mordukhovich [42] in which he assumes that f admits a so-called strong approximation.

THEOREM 5.3. *Let f be locally Lipschitzian around the point $(\overline{u}, \overline{y})$ solution of* (5.2). *Suppose that*

$$0 \in \partial_u^F (v^* \circ f)(\overline{u}, \overline{y}) + D^*G(\overline{u}, -f(\overline{u}, \overline{y})(v^*) \quad \Longrightarrow \quad v^* = 0,$$

where $D^*G(u,v)(v^*) = \{u^* \in U^* : \quad (u^*, -v^*) \in R_+ \partial^F d(u, v, GrG)\}$ is the coderivative multivalued mapping of G at (u,v). Then the multivalued mapping S is pseudo-Lipschitzian around $(\overline{u}, \overline{y})$, that is, there exist $r > 0$ and $b > 0$ such that for all $y, y' \in \overline{y} + rB_Y$

$$S(y) \cap (\overline{u} + rB_U) \subset S(y') + b\|y - y'\|B_U.$$

Proof. Apply Theorem 2.5 to the function $g(u,y) = d(u, -f(u,y), GrG)$ and use subdifferential calculus of composite functions for partial limiting Fréchet subdifferential which is the same as limiting Fréchet subdifferential. □

Remark. Theorem 5.3 may be obtained from Theorem 5.1 by setting $F(u,y) = f(u,y) + G(u)$.

When $G(u) = \{0\}$ for all u we obtain the following implicit function theorem for mappings which can be considered as a generalization of that of Clarke [6] since the partial limiting Fréchet subdifferential is smaller than Clarke's.

COROLLARY 5.4. *Let f be locally Lipschitzian around the point $(\overline{u}, \overline{y})$ solution of*

$$f(u,y) = 0.$$

Suppose that

$$0 \in \partial_u^F (v^* \circ f)(\overline{u}, \overline{y}) \quad \Longrightarrow \quad v^* = 0.$$

Then there exist $r > 0$ and a Lipschitz continuous mapping $\phi : (\overline{y} + rB_Y) \to U$ with $\phi(\overline{y}) = \overline{u}$ such that for all $y \in \overline{y} + rB_Y$

$$f(\phi(y), y) = 0.$$

6. Local controllability. Given a multivalued mapping $F : [0,1] \times R^n \to 2^{R^n}$ and a set $C_0 \subset R^n$, we call x a trajectory for F with respect to C_0 if it satisfies

$$(6.1) \quad x(0) \in C_0 \quad \text{and} \quad x'(t) \in F(t, x(t)) \quad \text{almost everywhere (a.e.) on } [0,1].$$

We denote by S the set of all trajectories of (6.1). We suppose in what follows this section that *S is a closed and nonempty subset* of the space $\mathcal{C}([0,1], R^n)$ of continuous functions on $[0,1]$ with values in R^n.

The parametrized reachable set of (6.1) (with parameter $z \in \mathcal{C}([0,1], R^n)$) is given by

$$R(z) = \{x(1) : \ x + z \in S\}.$$

The target is a closed nonempty subset of R^n and is denoted by C_1. Let A be the multivalued mapping from $R^n \times \mathcal{C}([0,1],R^n)$ into $\mathcal{C}([0,1],R^n)$ defined by

$$A(y,z) = \{x \in \mathcal{C}([0,1],R^n): \quad x(1)+y \in C_1,\ x+z \in S\}.$$

Let w be the linear continuous functional from $\mathcal{C}([0,1],R^n)$ into R^n defined by $w(x) = x(1)$ and let w^* be its adjoint functional. We denote by Ψ_C the indicator function of C.

Finally we shall say that the system is strongly locally controllable (s.l.c.) if there exists $r>0$ such that for all $z \in rB_{\mathcal{C}([0,1],R^n)}$,

$$rB_{R^n} \subset R(z) - C_1$$

and is locally controllable if this later holds for $z=0$.

There are many publications devoted to local controllability (see, for example, [10], [6], and bibliographies therein). Most works in the area conduct a local controllability under conditions expressed, in primal space in terms of Clarke's tangent cone or some derivative of multivalued mappings, or in dual space in terms of Clarke's normal cone. Our aim in this section is to present sufficient conditions in terms of the approximate subdifferential of the distance function ensuring strong local controllability. Note that the approximate subdifferential of locally Lipschitz functions is usually much smaller than Clarke's. It turns out that the sufficient condition below (6.6) is much more restrictive than the same one expressed in terms of Clarke's subdifferential and this latter does not hold for a broad class of sets C_1 and S important for applications (see, for example, [19]).

THEOREM 6.1. *Let $\overline{x} \in S$ with $\overline{x}(1) \in C_1$. Suppose that*

$$w^*(\partial^A d(\overline{x}(1),C_1)) \cap (-\partial^A \Psi_S(\overline{x})) = \{0\}. \tag{6.2}$$

Then there exist $r>0$ and $a>0$ such that for all $x \in \overline{x}+rB_{\mathcal{C}([0,1],R^n)}$, $y \in rB_{R^n}$, $z \in rB_{\mathcal{C}([0,1],R^n)}$,

$$d(x,A(y,z)) \le a(d(x(1)+y,C_1)+d(x+z,S)) \tag{6.3}$$

and hence the system is s.l.c.

Proof. It is not difficult to show that when $u \notin C_1$

$$\partial^A d(u,C_1) \subset S_{R^n}, \tag{6.4}$$

where $S_{R^n} = \{x^* \in R^n: \quad \|x^*\| = 1\}$. Set $f(x,y,z) = d(x(1)+y,C_1)+\Psi_S(x+z)$. We will show that (2.5) holds, and we apply Theorem 2.5. So, suppose the contrary. Then there are $(x_k,y_k,z_k) \to (\overline{x},0,0)$ and (x_k^*), with $\|x_k^*\| \to 0$ such that $x_k \notin A(y_k,z_k)$ and $x_k^* \in \partial_x^A f(x_k,y_k,z_k)$, for all integer k. By the definition of f and the subdifferential calculus we get $\partial_x^A f(x_k,y_k,z_k) \subset w^*(\partial^A d(x_k(1)+y_k,C_1)) + \partial_A \Psi_S(x_k+z_k)$. Then $x_k+z_k \in S$ and hence $x_k(1)+y_k \notin C_1$ and by (6.4) there exist $u_k^* \in \partial^A d(x_k(1)+y_k,C_1)$, with $\|u_k^*\| = 1$, and $v_k^* \in \partial^A \Psi_S(x_k+z_k)$ such that

$$x_k^* = w^*(u_k^*) + v_k^*.$$

We may suppose that $u_k^* \to u^*$, with $u^* \in \partial^A d(\overline{x}(1),C_1)$ and, by (6.4), $\|u^*\| = 1$ and $v_k^* \to v^*$, with $v^* \in \partial^A d(\overline{x},S)$ and

$$w^*(u^*) + v^* = 0,$$

and this contradicts (6.2) since $\|w^*(u^*)\| = 1$. □

Remark. In fact we showed that for all $x \in \overline{x} + rB_{\mathcal{C}([0,1],R^n)}$, $y \in rB_{R^n}$, $z \in rB_{\mathcal{C}([0,1],R^n)}$

$$d(x, A(y,z)) \le a(d(x(1)+y, C_1) + \Psi_S(x+z)) \tag{6.5}$$

for all $x \in \overline{x} + rB_{\mathcal{C}([0,1],R^n)}$, $y \in rB_{R^n}$, $z \in rB_{\mathcal{C}([0,1],R^n)}$, but this is equivalent (use Proposition 2.4.3 in [6]) to relation (6.3) with another constant a_1 instead of a.

In the following theorem we show that in relation (6.2), we may replace $\partial^A \Psi_S(\overline{x})$ by $\partial^A d(S, \overline{x})$.

THEOREM 6.2. *Theorem* 6.1 *remains valid if we replace* (6.2) *with*

$$w^*(\partial^A d(\overline{x}(1), C_1)) \cap (-\partial^A d(S, \overline{x})) = \{0\}. \tag{6.6}$$

Proof. Suppose that (6.5) does not hold. Then there are sequences $(x_k, y_k, z_k) \to (\overline{x}, 0, 0)$ such that $x_k + z_k \in S$ and

$$d(x_k, A(y_k, z_k)) > k d(x_k(1) + y_k, C_1). \tag{6.7}$$

Set $g_k(x) = d(x(1)+y_k, C_1)$ and $\varepsilon_k = \sqrt{g_k(x_k)}$. Note that by (6.7) $\varepsilon_k > 0$ and $\varepsilon_k \to 0$. Set $\lambda_k = \min(k\varepsilon_k^2, \varepsilon_k)$ and $s_k = \frac{\varepsilon_k^2}{\lambda_k} = \max(\frac{1}{k}, \varepsilon_k)$. Then

$$g_k(x_k) \le \inf_{x \in S - z_k} g_k(x) + \varepsilon_k^2$$

and hence by Ekeland's variational principle there exists $u_k \in S - z_k$ such that

$$\|u_k - x_k\| \le \lambda_k \tag{6.8}$$

and

$$g_k(u_k) \le g_k(x) + s_k \|x - u_k\| \quad \forall x \in S - z_k,$$

and hence u_k is a local minimum of the function

$$x \to g_k(x) + s_k\|x - u_k\| + (1 + s_k) d(x + z_k, S).$$

Thus

$$0 \in \partial^A g_k(u_k) + (1+s_k)\partial^A d(u_k + z_k, S) + s_k B_{(\mathcal{C}([0,1],R^n))^*}.$$

Note that by (6.8) $u_k(1) + y_k \notin C_1$ and hence by (6.4) there exist $u_k^* \in \partial^A d(u_k(1) + y_k, C_1)$, with $\|u_k^*\| = 1$, and $x_k^* \in (1+s_k)\partial^A d(u_k + z_k, S)$ such that

$$\|w^*(u_k^*) + x_k^*\| \le s_k,$$

and as in the proof of Theorem 6.1 we arrive at a contradiction with (6.6). □

In order to give other conditions ensuring relation (6.3), we give an estimate of the approximate subdifferential of the distance function to the solution set of the differential inclusion (6.1). We begin with some definitions and notations. A multivalued mapping $G : R^n \to R^n$ is said to be Lipschitz on a set A (of rank k) provided that for all $x, u \in A$

$$G(x) \subset G(u) + k\|x - u\| B_{R^n},$$

where $\|\cdot\|$ denotes the euclidean norm. Let Δ and Ω_t be the sets defined by

$$\Delta = \{t: \quad (t,x) \in \Omega\}$$
$$\Omega_t = \{x: \quad (t,x) \in \Omega\}.$$

Following Clarke [6], Ω is called a tube provided the set Δ is an interval (say, $\Delta = [0,1]$) and provided there exist a continuous function $\omega(t)$ and a continuous positive function ε on $[0,1]$ such that $\Omega_t = \omega(t) + \varepsilon(t)B_{R^n}$ for all $t \in [0,1]$. We call such tube a tube on $[0,1]$. If x is a given continuous function on $[0,1]$, the ε-tube about x, denoted $T(x,\varepsilon)$, is the tube on $[0,1]$ obtained by setting

$$\Omega = \{(t,u): \quad 0 \le t \le 1,\ u \in x(t) + \varepsilon B_{R^n}\}.$$

Let Ω be a tube on $[0,1]$. We say [6] that F is measurably Lipschitz on Ω provided that the following hold.

(i) For each $x \in R^n$, the mutivalued mapping $t \to F(t,x)$ is measurable on $[0,1]$.

(ii) There exists an integrable function $k(t)$ on $[0,1]$ such that for each $t \in [0,1]$ the multivalued mapping $x \to F(t,x)$ is nonempty and Lipschitz of rank $k(t)$ on Ω_t.

We assume in the following that F *is measurably Lipschitz on* Ω. To the multivalued mapping F we associate the quantities

$$\rho(t,x,y) = d(y,F(t,x)), \quad \rho_F(x) = I_F(x,x') = \int_0^1 \rho(t,x(t),x'(t))dt,$$

and we set $K = \exp(\int_0^1 k(t)dt)$. The set AC will refer to the set of absolutely continuous functions $x: [0,1] \to R^n$.

Now we may compute the subdifferential of the solution set of (6.1).

PROPOSITION 6.3. *Let S be the solution set of* (6.1) *and let $\overline{x} \in S$ and $\varepsilon > 0$ such that $T(\overline{x},\varepsilon) \subset \Omega$. Suppose that the values of F are closed. Then*

$$\partial^A d(\overline{x},S) \subset 2\partial^A[K\rho_F(\cdot) + (K\ln K + 1)d(u(\cdot),C_0)](\overline{x}),$$

where $u: x \to x(0)$.

The proof of this proposition is a consequence of the following lemma.

LEMMA 6.4 (compare with [6, Lemma 2, p. 124]). *Let $f: AC \to R$ be a locally Lipschitz function around $\overline{x}$, with constant $K_f > 0$, and let $\overline{x}$ be a local solution of the problem of minimizing f over S. Then $\overline{x}$ is a local solution of the problem*

$$\min\{f(x) + K_f K\rho_F(x) + K_f(K\ln K + 1)d(x(0),C_0): \quad x \in AC\}.$$

Remark. We may apply Theorem 2.4 in order to obtain $K_f(K+1)$ instead of $K_f(K\ln K + 1)$, since

$$d(x,u^{-1}(C_0)) \le d(u(x),C_0) \quad \forall x \in \mathcal{C}([0,1],R^n).$$

Proof of Proposition 6.3. Let $x^* \in \partial^A d(\overline{x},S)$. Then [15] for all $L \in \mathcal{F}(AC)$ there are nets $(x_i) \subset S$, $\varepsilon_i > 0$, and $x_i^* \in (1+\varepsilon_i)B_{(AC)^*}$ such that $x_i \to \overline{x}$, $x_i^* \to x^*$, $\varepsilon_i \to 0^+$ and such that the function

$$x \to -\langle x_i^*, x - x_i\rangle + \varepsilon_i\|x - x_i\| + (1+\varepsilon_i)d(x,x_i + L)$$

attains a local minimum at x_i on S. Then, by Lemma 6.4, the function

$$x \to -\langle x_i^*, x - x_i\rangle + \varepsilon_i\|x - x_i\| + (1+\varepsilon_i)d(x, x_i + L) + (2+2\varepsilon_i) \times [K\rho_F(x) + (K\ln K + 1)d(x(0), C_0)]$$

attains a local minimum at x_i. Thus

$$x_i^* \in (2+2\varepsilon_i)\partial^A[K\rho_F(\cdot) + (K\ln K+1)d(u(\cdot), C_0)](x_i) + \varepsilon_i B_{(AC)^*} + (1+\varepsilon_i)\partial^A d(x_i, x_i + L)$$

and hence

$$x^* \in 2\partial^A[K\rho_F(\cdot) + (K\ln K+1)d(u(\cdot), C_0)](\overline{x})$$

and the proof is complete. □

Set $K_1 = 2(K\ln K + 1)$. Then, with the help of Proposition 6.3, we have the following result.

COROLLARY 6.5. *Let F be as in Proposition* 6.3. *Suppose that*

$$if\ (v'(t), v(t)) \in K_1\partial^C\rho(t, \overline{x}(t), \overline{x}'(t)) \quad a.e.$$

with

$$v(0) \in K_1\partial^A d(\overline{x}(0), C_0),\ v(1) \in \partial^A d(\overline{x}(1), C_1), \quad then\ v(1) = 0.$$

Then the conclusion of Theorem 6.2 *holds.*

Here $\partial^C\rho$ refers to the Clarke's subdifferential of ρ with respect to the variables $(\overline{x}(t), \overline{x}'(t))$ for t fixed.

Proof. It suffices to show that (6.6) holds and to apply Theorem 6.2. Indeed consider (as in Thibault [54]) the mappings $\alpha : R^n \times L^1([0,1], R^n) \to R^n \times R^n$ and $\beta : R^n \times L^1([0,1], R^n) \to L^1([0,1], R^n) \times L^1([0,1], R^n)$ defined by

$$\alpha(x(0), x') = (x(0), x(1)), \quad \beta(x(0), x') = (x, x').$$

Let $b \in \partial^A d(\overline{x}(1), C_1)$ with $-w^*(b) \in \partial^A(\overline{x}, S)$. By Proposition 6.3 there exist $a \in K_1\partial^A d(\overline{x}(0), C_0)$ and $(u, v) \in K_1\partial^A I_L(\overline{x}, \overline{x}')$ such that

$$-\alpha^*(a, b) = \beta^*(u, v)$$

and hence (see Thibault [54])

$$b = -v(1), \quad a = v(0), \quad \text{and} \quad u(t) = v'(t) \text{ a.e.},$$

and hence $b = 0$ and the proof is complete. □

As a consequence, we obtain the following result which gives us an upper estimate of the solution set of (6.1).

COROLLARY 6.6. *There exist $a > 0$ and $r > 0$ such that*

$$d(x, B(z, v)) \le a(d(x(0) + z, C_0) + \rho_F(x + v))$$

for all $x \in \overline{x} + rB_{AC}$, $z \in rB_{R^n}$, and $v \in rB_{AC}$.

Here

$$B(z, v) = \{x \in AC : \quad x(0) + z \in C_0,\ x'(t) + v'(t) \in F(t, x(t) + v(t)),\ a.e.\}.$$

Corollary 6.5 may be stated in the following general form. The proof of this result is omitted since it is similar to those of Theorems 6.1 and 6.2. Note that in the following result, the perturbation appears only on the left-hand side of the differential inclusion.

THEOREM 6.7. *Under the assumptions of Corollary* 6.5 *there exist* $a := a(\overline{x}) > 0$ *and* $r > 0$ *such that*

$$d(x, C(y,z,v)) \le a(d(x(0)+z, C_0) + \hat{\rho}_F(x+v) + d(x(1)+y, C_1))$$

for all $x \in \overline{x} + rB_{AC}$, $z, y \in rB_{R^n}$, *and* $v \in rB_{AC}$.

Here

$$C(y,z,v) = \{x \in AC : \quad x(1)+y \in C_1,\ x(0)+z \in C_0,\ x'(t)+v'(t) \in F(t,x(t)), \text{ a.e.}\}$$

and

$$\hat{\rho}_F(x+v) = \int_0^1 d(x'(t)+v'(t), F(t,x(t)))dt.$$

This theorem implies in particular that there exist $a > 0$ and $r > 0$ such that for all $(y,z) \in r(B_{R^n} \times B_{R^n})$ and $v \in rB_{AC}$ there exists $u \in AC$ such that

$$u(0)+z \in C_0, \quad u(1)+y \in C_1, \quad u'(t)+v'(t) \in F(t,u(t)), \quad \text{and}$$
$$\|\overline{x}-u\|_{AC} \le a[\|z\| + \|y\| + \|v\|_{AC}].$$

7. Applications to sensitivity analysis. Throughout this section X and Y will be Banach spaces. Suppose that an optimization problem (P) is given in the following abstract form:

$$\min\{f(x) : \quad g(x) \le 0\}.$$

It often happens that (P) lends itself naturally to parametric perturbation, so that (P) is embedded in a family of optimization problems (P_y) indexed by a parameter y

$$\min\{f(x,y) : \quad g(x,y) \le 0\}.$$

The value of the problem (P_y) is denoted $v(y)$, and v is called the value function. For each y in the domain of v we consider the set of minimizers

$$S(y) := \{x \in X : \quad g(x,y) \le 0, f(x,y) = v(y)\}.$$

We proceed to examine a few typical properties of v that have a bearing on (P). We begin by the Lipschitzian property of v. For this we introduce a compactness assumption which will assure the stability of the parametrized problems (P_y). A stability assumption (SA) holds if $S(y) \neq \emptyset$ for y near 0 and

$$\text{if } y_n \to 0 \text{ and } x_n \in S(y_n)\ \forall n, \text{ then } (x_n) \text{ has an accumulation point.}$$

THEOREM 7.1. *Suppose that*

(i) $S(0)$ *is nonempty,*

(ii) f *(resp., g) is lower semicontinuous and locally Lipschitzian near each point* $(x_0, 0)$, *with Lipschitz constant* $k_f(x_0)$, *(resp., near* 0 *uniformly around* x_0*) for* $x_0 \in S(0)$,

(iii) (SA) *holds,*

(iv) *the Hoffman error bound holds at each point* x_0 *in* $S(0)$*, i.e., there exist* $a(x_0) > 0$ *and* $r > 0$ *such that*

$$d(x, \{u : \ g(u, y) \le 0\}) \le a(x_0) g_+(x, y) \quad \forall x \in x_0 + rB_X, \quad \forall y \in rB_Y.$$

Then v *is locally Lipschitzian near* 0*.*

Proof. First we show that v is continuous at 0. Since v is upper semicontinuous it suffices to show that it is s.l.c. at 0. So suppose the contrary; then there exist $\varepsilon > 0$ and a sequence (y_n) converging to 0 such that for n large enough

$$v(0) > v(y_n) + \varepsilon.$$

By (iii), there exists $x_n \in S(y_n)$, which we assume converging to some x_0 with, by (ii), $x_0 \in S(0)$ and hence by the lower semicontinuity of f, we get

$$v(0) \ge v(0) + \varepsilon$$

which leads to a contradiction. So v is s.l.c. at 0.

Now we proceed to show that v is locally Lipschitzian around 0. So suppose the contrary; then there are sequences $y_n \to 0$ and $y'_n \to 0$ such that for n large enough

$$|v(y_n) - v(y'_n)| > n\|y_n - y'_n\|.$$

We may assume that the set $I = \{n : \ v(y_n) - v(y'_n) > n\|y_n - y'_n\|\}$ is infinite (because (y_n) and (y'_n) play a symmetric role). For all $n \in I$ there exists, by (iii), $x'_n \in S(y'_n)$, which we suppose converges to $x_0 \in S(0)$ (by the continuity of v). Now, by (iv), for $n \in I$ large enough

$$d(x'_n, \{u : \ g(u, y_n) \le 0\}) \le a g_+(x'_n, y_n)$$

and hence there exists x_n, with $g(x_n, y_n) \le 0$, such that

$$\|x'_n - x_n\| \le a g_+(x'_n, y_n)$$

and since g is locally Lipschitzian near 0 uniformly in x'_n, with constant $k_g = k_g(x_0)$

$$\|x'_n - x_n\| \le a(g_+(x'_n, y_n) - g_+(x'_n, y'_n)) \le a k_g \|y_n - y'_n\|.$$

So if k_f is a Lipschitz constant of f around $(x_0, 0)$, then for all $n \in I$ sufficiently large

$$n\|y_n - y'_n\| < f(x_n, y_n) - f(x'_n, y'_n) \le k_f(1 + a k_g)\|y_n - y'_n\|,$$

and this contradiction completes the proof. ∎

We have the following estimate of the subdifferential of v.

THEOREM 7.2. *Suppose, in addition to the assumptions of Theorem* 7.1*, that* g *is locally Lipschitzian near each point* $(x_0, 0)$*, with* $x_0 \in S(0)$*. Then*

- $\forall x_0 \in S(0)$, $\partial^A v(0) \cap M^A(x_0) \ne \emptyset$,
- $\partial^A v(0) \subset M^A(0)$.

Where

$$M^A(x_0) = \{y^* + z^* : \ (x^*, y^*) \in \partial^A f(x_0, 0), \ (-x^*, z^*) \in \alpha(x_0)[0, 1]\partial^A g(x_0, 0)\},$$

$$M^A(0) = \bigcup_{x_0 \in S(0)} M^A(x_0)$$

$\alpha(x_0) = (k_v + k_f(x_0))a(x_0)$ and k_v is a Lipschitz constant of v near 0.

Proof. The proof of the second part is omitted because it is similar to that given by Jourani in [19] for the so-called G-subdifferential.

For all $x_0 \in S(0)$ we have, by Theorem 7.1, that the function

$$(x, y) \to f(x, y) - v(y) + \alpha(x_0)g_+(x, y)$$

attains a local minimum at $(x_0, 0)$ and hence $(x_0, 0, v(0))$ is a local minimum of the function

$$(x, y, t) \to f(x, y) - t + \alpha(x_0)g_+(x, y)$$

over the set $\{(x, y, t) : \quad v(y) = t\}$, and the result follows by using Fritz–John necessary optimality conditions for approximate subdifferential. □

The previous result may be applied to produce estimates of the subdifferential of the marginal function

$$m(y, z) = \inf\{f(x(0), x(1)) : \quad x(1) + y \in C_1,\ x(0) + z \in C_0,\\ x'(t) \in F(t, x(t)),\ \text{a.e.}\}$$

Let $S(y, z)$ be the set of feasible points x satisfying $m(y, z) = f(x(0), x(1))$.

COROLLARY 7.3. *Suppose that for all $\overline{x} \in S(0, 0)$, the assumptions of Theorem 6.7 hold, f is locally Lipschitzian around $(\overline{x}(0), \overline{x}(1))$ with constant $k_f(\overline{x})$, and suppose that the multivalued mapping $(y, z) \to S(y, z)$ satisfies some compactness assumption like (SA). Then*

- *m is locally Lipschitzian around $(0, 0)$ (with constant k_m),*
- *$\forall \overline{x} \in S(0, 0)$, $\partial^A m(0, 0) \cap N(\overline{x}) \neq \emptyset$,*
- *$\partial^A m(0, 0) \subset N(0, 0)$.*

Where $K_1(\overline{x}) = (k_f(\overline{x}) + k_m)a(\overline{x})$, $a(\overline{x})$ is as in Theorem 6.7,

$$N(0, 0) = \bigcup_{\overline{x} \in S(0,0)} N(\overline{x}),$$

and $N(\overline{x})$ is the set of elements $(a, b) \in K_1(\overline{x})\partial^A d((\overline{x}(0), \overline{x}(1)), C_0 \times C_1)$ for which there exist $(c, d) \in \partial^A f(\overline{x}(0), \overline{x}(1))$ and v such that

$$(v'(t), v(t)) \in K_1(\overline{x})\partial^C \rho(t, \overline{x}(t), \overline{x}'(t)),\ \text{a.e.}$$

with

$$a + c = v(0) \quad \text{and} \quad b + d = -v(1).$$

Proof. Apply Theorem 7.2 with $g(x, y, z) = d(x(1)+y, C_1)+d(x(0)+z, C_0)+\rho_F(x)$ and use the formulae on generalized gradients of integral functionals [6]. □

8. Further results: Farkas lemma for quadratic inequality systems and weak sharp minima (or linear conditioning). We will use the previous results and the result by Mangasarian [37] to obtain a Farkas lemma for quadratic inequality system

$$x \in C, \quad \frac{1}{2}\langle x, Bx\rangle + \langle b, x\rangle + c \leq 0, \tag{8.1}$$

where C is polyhedral in R^n and $B \in R^{n\times m}$ is symmetric and positive semidefinite. We let S be a solution set to this system.

We use the description of the solution set of a convex program established in [37]. See [33] for a simple proof of the following proposition in the case where f is locally Lipschitz.

PROPOSITION 8.1. *Let D be the solution set to the problem*

$$\min_{x\in C} f(x),$$

where both $f : R^n \to R$ and $C \subset R^n$ are convex with f differentiable. Let $u \in D$. Then

$$D = \{x \in C : \quad \nabla f(x) = \nabla f(u),\ \langle \nabla f(u), x - u\rangle = 0\}.$$

Now we are able to present a Farkas lemma for (8.1).

PROPOSITION 8.2. *Let u be a boundary point in S and $B_i, i = 1, \dots, m$, be vector lines of B. Set $g(x) = \frac{1}{2}\langle x, Bx\rangle + \langle b, x\rangle + c$. Then*

$$N(S,u) \subset N(C,u) + R_+\partial g_+(u) \text{ or } N(S,u) \subset N(C,u) + \operatorname{span}\{B_1, \dots, B_m, b\}.$$

Proof. First case. If there exists $u_0 \in C$ such that $g(u_0) < 0$, then, by Proposition 3.2, we have

$$d(x,S) \le \frac{f_+(x)}{-g(u_0)}\|x - u_0\| \quad \forall x \in R^n,$$

where $f(x) = g(x) + \Psi_C(x)$, or, equivalently,

$$d(x,S) \le \frac{g_+(x)}{-g(u_0)}\|x - u_0\| \quad \forall x \in C.$$

This ensures that

$$N(S,u) \subset N(C,u) + R_+\partial g_+(u).$$

Second case. For all $x \in C$, $g(x) \ge 0$. In this case S is the set of solutions of the problem

$$\min_{x\in C} g(x).$$

Thus, by Proposition 8.1, $S = \{x \in C : \quad A(x - u) = 0\}$, where $A = (B, b)$ and Theorem 4.3 implies the existence of $a > 0$ such that

$$d(x,S) \le a\|A(x - u)\| \quad \forall x \in C.$$

Thus

$$N(S,u) \subset N(C,u) + \operatorname{span}\{B_1, \dots, B_m, b\}.$$

This completes the proof. □

Remark. Proposition 8.2 may be obtained as a consequence of the well-known results in convex analysis (see [51]).

Consider a function $g : X \to R \cup \{\infty\}$ and a subset C of X. We say that $S := \arg\min_C g$ is a set of weak sharp minima for g relative to C with modulus $b > 0$ if

$$g(x) \geq g(u) + bd(x, S) \quad \forall x \in C, \quad \forall u \in S.$$

As we can see that this is equivalent to Hoffman's error bound

$$d(x, S) \leq \frac{1}{b} f_+(x) \quad \forall x \in X,$$

where $f(x) = g(x) - g(u) + \Psi_C(x)$ for some $u \in S$. To simplify we assume that $C = X$.

THEOREM 8.3. *Let f be lower semicontinuous, $b > 0$, and $S := \arg\min f$ be nonempty. Consider the following statements.*

(i) $(\partial f)^{-1}(x^*) \subset S \ \forall x^* \in X^*$ *with* $\|x^*\| < b$.

(ii) $\forall x \in \mathrm{dom} f$, $x \notin S \quad d(0, \partial f(x)) \geq b$.

(iii) S *is a set of weak sharp minima for f on X with modulus b.*

Then (i) $\Longleftrightarrow$ (ii) $\Longrightarrow$ (iii). *If in addition f is proper and convex, statements* (i), (ii), *and* (iii) *are equivalent. Here* $(\partial f)^{-1}(x^*) := \{x \in X : \quad x^* \in \partial f(x)\}$.

Proof. The equivalence (i) $\Longleftrightarrow$ (ii) is obvious. The implication (ii) $\Longrightarrow$ (iii) follows from Theorem 2.4. Theorem 5.1 in [8] ensures the implication (iii) $\Longrightarrow$ (i). □

Remark. One of the referees pointed out that the nonparametric version of Theorem 2.4 and the equivalence of (ii) and (iii) were obtained independently and presented at a meeting in Dallas in 1997 by J. V. Burke and S. Deng.

Acknowledgment. The author wishes to acknowledge the comments of the referees, which helped to correct and improve the presentation of the paper.

REFERENCES

[1] J.-P. AUBIN, *Lipschitz behavior of solutions to convex minimization problems*, Math. Oper. Res., 9 (1984), pp. 87–111.

[2] J.-P. AUBIN AND H. FRANKOWSKA, *On inverse function theorems for set-valued maps*, J. Math. Pures Appl., 66 (1987), pp. 71–89.

[3] J. V. BURKE AND C. FERRIS, *Weak sharp minima in mathematical programming*, SIAM J. Control Optim., 31 (1993), pp. 1340–1359.

[4] J. V. BURKE AND P. TSENG, *A unified analysis of Hoffman's bound via Fenchel duality*, SIAM J. Optim., 6 (1996), pp. 265–282.

[5] F. H. CLARKE, *Solving equations by decrease principle*, CRM Proc. Lecture Notes 11, Amer. Math. Soc., Providence, RI, 1997, pp. 29–39.

[6] F. H. CLARKE, *Optimization and Nonsmooth Analysis*, John Wiley, New York, 1983.

[7] F. H. CLARKE, YU. S. LEDYAEV, R. T. STERN, AND P. R. WOLENSKI, *Nonsmooth Analysis and Control Theory*, Grad. Texts in Math. 178, Springer, New York, 1998.

[8] O. CORNEJO, A. JOURANI, AND C. ZALINESCU, *Conditioning and upper Lipschitz inverse subdifferentials in nonsmooth optimization problems*, J. Optim. Theory Appl., 95 (1998), pp. 127–148.

[9] S. DENG, *Computable error bounds for convex inequality systems in reflexive Banach spaces*, SIAM J. Optim., 7 (1997), pp. 274–279.

[10] H. FRANKOWSKA, *Local controllability and infinitesimal generators of semigroups of set-valued maps*, SIAM J. Contol Optim., 25 (1987), pp. 412–432.

[11] A. J. HOFFMAN, *On approximate solutions of systems of linear inequalities*, J. Research Nat. Bur. Standards, 49 (1952), pp. 263–265.

[12] H. HU AND Q. WANG, *On approximate solutions of infinite systems of linear inequalities*, Linear Algebra Appl., 114/115 (1989), pp. 429–438.

[13] A. D. IOFFE, *Regular points of Lipschitz functions*, Trans. Amer. Math. Soc., 251 (1979), pp. 61–69.

[14] A. D. Ioffe, *Approximate subdifferentials and applications* 2*: Functions on locally convex spaces*, Mathematika, 33 (1986), pp. 111–128.
[15] A. D. Ioffe, *Approximate subdifferentials and applications* 3*: Metric theory*, Mathematika, 36 (1989), pp. 1–38.
[16] A. D. Ioffe, *Approximate subdifferential and applications* 1*: The finite dimensional theory*, Trans. Amer. Math. Soc., 281 (1984), pp. 389–416.
[17] A. D. Ioffe, *Codirectional compactness, metric regularity and subdifferential calculus*, Nonlinear Anal., to appear.
[18] A. Jourani, *Qualification conditions for multivalued functions in Banach spaces with applications to nonsmooth vector optimization problems*, Math. Programming, 66 (1994), pp. 1–23.
[19] A. Jourani, *Intersection formulae and the marginal function in Banach spaces*, J. Math. Anal. Appl., 192 (1995), pp. 867–891.
[20] A. Jourani, *Refinements of Necessary Conditions for the Generalized Problem of Bolza and Applications*, preprint, Université de Bourgogne, France, 1999.
[21] A. Jourani, *Metric regularity and second order necessary optimality conditions for minimization problems under inclusion constraints*, J. Optim. Theory Appl., 81 (1994), pp. 97–120.
[22] A. Jourani and L. Thibault, *Approximate subdifferential and metric regularity: Finite dimensional case*, Math. Programming, 47 (1990), pp. 203–218.
[23] A. Jourani and L. Thibault, *Verifiable conditions for openness and metric regularity of multivalued mappings in Banach spaces*, Trans. Amer. Math. Soc., 347 (1995), pp. 1255–1268.
[24] A. Jourani and L. Thibault, *Coderivatives of multivalued mappings, locally compact cones and metric regularity*, Nonlinear Anal., 35 (1999), pp. 925–945.
[25] A. Jourani and L. Thibault, *Metric regularity for strongly compactly Lipschitzian mappings*, Nonlinear Anal., 24 (1994), pp. 229–240.
[26] A. Jourani and L. Thibault, *Qualification conditions for calculus rules of coderivatives of multivalued mappings*, J. Math. Anal. Appl., 218 (1998), pp. 66–81.
[27] A. Jourani and L. Thibault, *Chain rules for coderivatives of multivalued mappings in Banach spaces*, Proc. Amer. Math. Soc., 126 (1998), pp. 1479–1485.
[28] D. Klatte, *Hoffman's error bound for systems of convex inequalities*, Mathematical Programming with Data Perturbations, A. V. Fiacco, ed., Lecture Notes in Pure and Appl. Math. 195, Marcel Dekker, New York, Basel, 1998, pp. 185–199.
[29] A. Ya. Kruger, *Properties of generalized differentials*, Siberian Math. J., 26 (1985), pp. 822–832.
[30] A. Ya. Kruger and B. S. Mordukhovich, *Extremal points and the Euler equations in nondifferentiable optimization problems*, Dokl. Akad. Nauk USSR, 24 (1980), pp. 684–687.
[31] A. Ya. Kruger and B. S. Mordukhovich, *Generalized normals and derivatives, and necessary optimality conditions in nondifferentiable programming*, *Part* I: Depon. VINITI 408-80; *Part* II: Depon. VINITI 494-80, Moscow, 1980.
[32] Yu. S. Ledyaev and Q. J. Zhu, *Implicit Multifunction Theorems*, CECM Research Report 98:105, Simon Fraser University, Vancouver, BC, Canada, 1998.
[33] O. Lefebvre, private communication, Université de Bourgogne, France, 1998.
[34] X.-D. Luo and Z.-Q. Luo, *Extension of Hoffman's error bound to polynomial systems*, SIAM J. Optim., 4 (1994), pp. 383–392.
[35] Z. Q. Luo and J. S. Pang, *Error bounds for analytic systems and their applications*, Math. Programming, 67 (1995), pp. 1–28.
[36] O. L. Mangasarian, *Error bounds for nondegenerate monotone linear complementarity problems*, Math. Programming, 48 (1990), pp. 437–446.
[37] O. L. Mangasarian, *A simple characterization of solution sets of convex programs*, Oper. Res. Lett., 7 (1998), pp. 21–26.
[38] R. Mathias and J. S. Pang, *Error bounds for the linear complemetarity problem with a P-matrix*, Linear Algebra Appl., 132 (1990), pp. 123–136.
[39] B. S. Mordukhovich, *Maximum principle in problems of time optimal control with nonsmooth constraints*, J. Appl. Math. Mech., 40 (1976), pp. 960–969.
[40] B. S. Mordukhovich, *Metric approximations and necessary optimality conditions for general class of nonsmooth extremal problems*, Soviet Math. Dokl., 22 (1980), pp. 526–530.
[41] B. S. Mordukhovich, *Complete characterization of openness, metric regularity, and Lipschitzian properties of multifunctions*, Trans. Amer. Math. Soc., 340 (1993), pp. 1–35.
[42] B. S. Mordukhovich, *Stability theory for parametric generalized equations and variational inequalities via nonsmooth analysis*, Trans. Amer. Math. Soc., 343 (1994), pp. 609–657.
[43] B. S. Mordukhovich, *Lipschitzian stability of constraint systems and generalized equations*, Nonlinear Anal., 22 (1994), pp. 173–206.

[44] B. S. Mordukhovich, *Approximation Methods in Problems of Optimization and Control*, Nauka, Moscow, 1988. Wiley-Interscience English translation to appear.
[45] B. S. Mordukhovich and Y. Shao, *Stability of set-valued mappings in infinite dimensions: Point criteria and applications*, SIAM J. Control Optim., 35 (1997), pp. 285–314.
[46] B. S. Mordukhovich and Y. Shao, *Differential characterizations of covering, metric regularity, and Lipschitzian properties of multifunctions between Banach spaces*, Nonlinear Anal., 25 (1995), pp. 1401–1424.
[47] J.-P. Penot, *Well-Behavior, Well-Posedness and Nonsmooth Analysis*, preprint, University of Pau, France, 1998.
[48] R. R. Phelps, *Convex Functions, Monotone Operators and Differentiability*, 2nd ed., Lecture Notes in Math. 1364, Springer, Berlin, 1993.
[49] S. Robinson, *Generalized equations and their solutions, part* I: *Basic theory*, Math. Program. Study, 10 (1979), pp. 128–141.
[50] S. M. Robinson, *An application of error bounds for convex programming in a linear space*, SIAM J. Control, 13 (1975), pp. 271–273.
[51] R. T. Rockafellar, *Convex Analysis*, Princeton University Press, Princeton, NJ, 1970.
[52] R. T. Rockafellar, *Lipschitzian properties of multifunctions*, Nonlinear Anal., 9 (1985), pp. 867–885.
[53] L. Thibault, *Equivalence between metric regularity and graphical metric regularity*, Université de Montpellier II, France, to appear.
[54] L. Thibault, *Calcul sous-differentiel et calcul des variations en dimension infinie*, Bull. Soc. Math. France, 60 (1979), pp. 161–175.
[55] L. Thibault and D. Zagrodny, *Integration of subdifferentials of lower semicontinuous functions in Banach spaces*, J. Math. Anal. Appl., 189, (1995), pp. 33–58.
[56] T. Wang and J. S. Pang, *Global error bounds for convex quadratic inequality systems*, Optimization, 31 (1994), pp. 1–12.

SIAM J. CONTROL OPTIM.
Vol. 38, No. 3, pp. 971–994

ANY DOMAIN OF ATTRACTION FOR A LINEAR CONSTRAINED SYSTEM IS A TRACKING DOMAIN OF ATTRACTION*

FRANCO BLANCHINI† AND STEFANO MIANI‡

Abstract. In the stabilization problem for systems with control and state constraints a domain of attraction is a set of initial states that can be driven to the origin by a feedback control without incurring constraints violations. If the problem is that of tracking a reference signal, that converges to a constant constraint-admissible value, a tracking domain of attraction is a set of initial states from which the reference signal can be asymptotically approached without constraints violation during the transient. Clearly, since the zero signal is an admissible reference signal, any tracking domain of attraction is a domain of attraction. We show that the opposite is also true. For constant reference signals we establish a connection between the convergence speed of the stabilization problem and tracking convergence which turns out to be independent of the reference signal. We also show that the tracking controller can be inferred from the stabilizing (possibly nonlinear) controller associated with the domain of attraction.

Key words. constrained control, domain of attraction, Lyapunov functions, tracking, convex analysis

AMS subject classifications. 93B52, 93D30, 93B51, 93B50

PII. S036301299834661X

1. Introduction.

In this paper we consider linear systems with state and control constraints. For this class of systems we say that a certain convex and compact set including the origin in its interior is a domain of attraction to the origin if there exists a feedback control such that for any initial state in this set the state is driven asymptotically to the origin without constraints violation. Several previous references have dealt with the construction of such domains (see for instance [3], [6], [11], [12], [17], [18], [19], [20], [21], [26], [28]).

Here we consider the problem of tracking a reference signal including the zero one. We assume that the system is square (i.e., it has as many inputs as outputs) and that it is free of invariant zeros at one in the discrete-time case (or it is free from zeros at the origin in the continuous-time case). We define a set of reference vectors which are constraints-compatible in the sense that they are associated with state vectors which are in the interior of the domain of attraction to the origin and to feasible input vectors. The only condition required for the signal to be tracked is that it asymptotically converges to a value which is in this set. A signal satisfying the above requirement will be said to be admissible. We stress that the signal, during the transient, may assume values outside this set and it is a goal of the control to avoid constraints violations. For the zero reference signal, such a goal can be achieved only if the initial state is inside a domain of attraction to the origin. Any initial state from which we can track asymptotically an admissible reference signal is said to belong to a tracking domain of attraction. The question which hence arises is whether there are initial states inside the domain of attraction to the origin from which we cannot

*Received by the editors October 29, 1998; accepted for publication (in revised form) August 3, 1999; published electronically March 15, 2000.
http://www.siam.org/journals/sicon/38-3/34661.html
†Dipartimento di Matematica ed Informatica, Università di Udine, via delle Scienze 208, 33100 Udine, Italy (blanchini@uniud.it).
‡Dipartimento di Ingegneria Elettrica, Gestionale e Meccanica, Università di Udine, via delle Scienze 208, 33100 Udine, Italy (miani.stefano@uniud.it).

solve the problem of tracking an admissible reference signal. Put in other words this means to verify whether or not the largest tracking domain of attraction is a proper subset of the assigned domain of attraction. We show that the answer is negative.

The basic ideas we develop here are related to previously published results [2], [15], [14], [16]. The main difference relies on the fact that in those references it is assumed that a stabilizing linear nominal compensator (often referred to as precompensator) is applied to the system. To this nominal compensator the so called reference governor is added which is a device that possibly attenuates the effect of the reference signal in order to avoid violations. This is basically equivalent to considering a stable system (i.e., a system to which a precompensator is applied) and to managing the reference for this system. Although this assumption is quite reasonable, it turns out that the resulting constructed invariant sets depend on the precompensator. Therefore, an unsuitable choice of the compensator can produce a very small domain of attraction.

Conversely, here *we do not assume the existence of any precompensator* and thus we may take as a domain of attraction the largest one in absolute in the sense that, if the initial state is outside that region, there is no guarantee that constraints violation can be avoided, even for zero reference signal. We show that this is a tracking domain of attraction, in the sense that there is a (possibly nonlinear) compensator, which is not a priori fixed, that solves the tracking problem for any initial state in this region. Furthermore, we show that the control strategy can be inferred from the original stabilizing controller associated with the given domain of attraction. We use as a technical support a Lyapunov function suitably constructed by "reshaping" the one associated with such a domain.

Finally we show that, for constant reference signals and symmetric domains, the speed of convergence can be estimated a priori from the speed of convergence achieved by the stabilizing controller, and it does not depend on the particular reference.

The paper is organized as follows. In section 2 we provide the basic definitions. In section 3 we state the main results which will be proven later in section 4 and section 5. We particularize the results to domains of attraction of special shapes in section 6. We show how to extend the results to nonsquare systems in section 7. We finally present two examples in section 8 to illustrate the method, and we derive our conclusions in section 9.

2. Definitions and problem statement. In what follows we denote by $int\mathcal{P}$ the interior of a set $\mathcal{P}$ and by $\partial\mathcal{P}$ its boundary. With the term C-set we denote a convex and compact set containing the origin as an interior point. It is known that any C-set $\mathcal{P}$ induces a positively homogeneous convex function which is known as Minkowski functional (see [22]):

$$\psi_{\mathcal{P}}(x) = \inf\{\xi \geq 0 : x \in \xi\mathcal{P}\}. \tag{2.1}$$

The function $\psi_{\mathcal{P}}(x) \geq 0$ is such that $\psi_{\mathcal{P}}(x) = 0$ iff $x = 0$, and $\psi_{\mathcal{P}}(\lambda x) = \lambda\psi_{\mathcal{P}}(x)$ for any $\lambda \geq 0$. If $\mathcal{P}$ is 0-symmetric (i.e., $x \in \mathcal{P}$ implies $-x \in \mathcal{P}$), then $\psi_{\mathcal{P}}$ is a norm. The index $\mathcal{P}$ will be omitted for brevity when the inducing set $\mathcal{P}$ of $\psi_{\mathcal{P}}$ is clear from the context.

In the following we consider both discrete and continuous-time square systems

$$\begin{array}{rcl} \delta x(t) & = & Ax(t) + Bu(t) \\ y(t) & = & Cx(t) + Du(t), \end{array} \tag{2.2}$$

where δ represents the shift operator $\delta x(t) = x(t+1)$ in the discrete-time case and the derivative operator in the continuous-time case. The vector $y(t) \in \mathbb{R}^q$ is the system

output, the vector $x(t) \in \mathbb{R}^n$ is the system state, and $u(t) \in \mathbb{R}^q$ is the control input. We assume that (A, B) is stabilizable and that x and u are subject to the constraints

$$x(t) \in \mathcal{X}, \tag{2.3}$$

$$u(t) \in \mathcal{U}, \tag{2.4}$$

where $\mathcal{X} \subset \mathbb{R}^n$ and $\mathcal{U} \subset \mathbb{R}^q$ are assigned C-sets.

Due to the presence of the constraints, the solution of the stabilization problem implies restrictions on the admissible initial conditions.

DEFINITION 2.1. *The C-set $\mathcal{S} \subset \mathcal{X}$ is a domain of attraction for the system* (2.2) *if there exists a continuous feedback control function $u(t) = \phi(x(t))$ such that any trajectory $x(t)$ with initial condition $x(0) \in \mathcal{S}$ is such that $x(t) \in \mathcal{S}$, $u(t) \in \mathcal{U}$, and*

$$\lim_{t \to \infty} x(t) = 0.$$

It is well known that the above definition is fundamental for the stabilization problem under constraints. An initial condition can be driven to the origin without constraints violations iff it belongs to a domain of attraction. However, in practice, simple convergence is not enough, but it is important to provide an index of the speed of convergence to the origin. Thus we introduce the following two definitions.

DEFINITION 2.2. *The C-set $\mathcal{S} \subset \mathcal{X}$ is a domain of attraction with speed of convergence λ for the* discrete-time *system* (2.2) *if there exists $0 \leq \lambda < 1$ and a continuous feedback control function $u(t) = \phi(x(t))$ such that any trajectory $x(t)$ with initial condition $x(0) \in \mathcal{S}$ is such that*

$$\psi_{\mathcal{S}}(x(t)) \leq \lambda^t \psi_{\mathcal{S}}(x(0)).$$

If we take $\lambda = 1$, the set $\mathcal{S}$ is simply said to be positively invariant (often referred to as controlled–invariant). We say that $\mathcal{S}_\lambda \subset \mathcal{X}$ is the largest domain of attraction if for any domain of attraction $\mathcal{S}$ in $\mathcal{X}$ with speed of convergence λ we have $\mathcal{S} \subset \mathcal{S}_\lambda$.

DEFINITION 2.3. *The C-set $\mathcal{S} \subset \mathcal{X}$ is a domain of attraction with speed of convergence β for the* continuous-time *system* (2.2) *if there exists $\beta > 0$ and a continuous feedback control function $u(t) = \phi(x(t))$ such that any trajectory $x(t)$ with initial condition $x(0) \in \mathcal{S}$ is such that*

$$\psi_{\mathcal{S}}(x(t)) \leq e^{-\beta t} \psi_{\mathcal{S}}(x(0))$$

If we take $\beta = 0$, the set $\mathcal{S}$ is simply said to be positively invariant (often referred to as controlled-invariant). We say that $\mathcal{S}_\beta \subset \mathcal{X}$ is the largest domain of attraction if for any domain of attraction $\mathcal{S}$ in $\mathcal{X}$ with speed of convergence β we have $\mathcal{S} \subset \mathcal{S}_\beta$.

A stabilizable system always admits a domain of attraction $\mathcal{P} \subset \mathcal{X}$. The knowledge of such a domain is fundamental in the stabilization problem under constraints, since if $x(0) \in \mathcal{P}$, then the conditions $x(t) \in \mathcal{P} \subset \mathcal{X}$ for $t \geq 0$, and $x(t) \to 0$ as $t \to \infty$ can be assured. Thus, once we have computed a domain of attraction $\mathcal{P}$ to solve the problem (possibly the largest one [4], [6], [7]), we can replace the constraint $x(t) \in \mathcal{X}$ by the new constraint

$$x(t) \in \mathcal{P}.$$

In this paper we deal with the problem of tracking a certain class of reference signals. To this aim we make the assumption that the system is free from zeros at one (zeros at the origin).

ASSUMPTION 2.1. *The square matrix*

$$M_d = \begin{bmatrix} A-I & B \\ C & D \end{bmatrix} \left(\textit{resp.}, \; M_c = \begin{bmatrix} A & B \\ C & D \end{bmatrix} \right)$$

is invertible.

The general case in which M_d (M_c) is not invertible or even nonsquare will be considered in section 7. Under Assumption 2.1 the system has the property that for any constant reference $r \in \mathbb{R}^q$ there is an unique state-input equilibrium pair $(\bar{x}, \bar{u})$ such that the corresponding equilibrium output is r. Such a pair is the unique solution of the equation

$$M_d \begin{bmatrix} \bar{x}(r) \\ \bar{u}(r) \end{bmatrix} = \begin{bmatrix} 0 \\ r \end{bmatrix} \left(\text{resp.}, \; M_c \begin{bmatrix} \bar{x}(r) \\ \bar{u}(r) \end{bmatrix} = \begin{bmatrix} 0 \\ r \end{bmatrix} \right).$$

Thus we can define the set of admissible constant reference vectors

$$\mathcal{R} = \{ r \in \mathbb{R}^q : \; \bar{u}(r) \in \mathcal{U}, \; \bar{x}(r) \in \mathcal{P} \}.$$

Being that $\mathcal{P}$ and $\mathcal{U}$ are both bounded, $\mathcal{R}$ is also bounded. The set $\mathcal{R}$ is of fundamental importance. It is the set of all reference vectors for which the corresponding input and state equilibrium pairs do not violate the constraints $\bar{u} \in \mathcal{U}$ and $\bar{x} \in \mathcal{P}$. While the reason for imposing the former is obvious, the second deserves some explanation. Indeed we have imposed $x \in \mathcal{X}$. However, note that if $\bar{x}(r)$ is not included in a domain of attraction, then the condition $x(t) \to \bar{x}(r)$ can cause a violation of the constraints if, for instance, after a sufficiently long period, the reference $r(t)$ switches to zero.

We are now going to introduce the set of all the admissible signals to be tracked, formed by the signals $r(t)$ having a finite limit r_∞, with the condition that r_∞ has some admissibility condition with respect to the constraints.

DEFINITION 2.4. *Assume that a small* $0 < \epsilon < 1$ *is given. A reference signal* $r(t)$ *is admissible if it is continuous*[1] *and if it is such that*

$$\lim_{t \to \infty} r(t) = r_\infty \in (1-\epsilon)\mathcal{R} \doteq \mathcal{R}_\epsilon.$$

The parameter ϵ, as we will see later, is introduced to avoid singularities in the control. Such ϵ may be arbitrarily small, and thus it does not practically affect the problem. We stress that an admissible reference signal *does not need to assume its values in* $\mathcal{R}_\epsilon$, but only its limit r_∞ needs to do this. Now we can state the following basic definition.

DEFINITION 2.5. *The set* $\mathcal{P} \subset \mathcal{X}$ *is a tracking domain of attraction if there exists a (possibly nonlinear) feedback control*

$$u(t) = \Phi(x(t), r(t))$$

such that for any $x(0) \in \mathcal{P}$ *and for every admissible reference signal* $r(t)$

(i) $x(t) \in \mathcal{P}$ *and* $u(t) \in \mathcal{U}$,

(ii) $y(t) \to r_\infty$ *as* $t \to \infty$.

[1]The continuity requirement for the function $r(t)$ is limited to the continuous-time case only. It is not essential but avoids unnecessary complications.

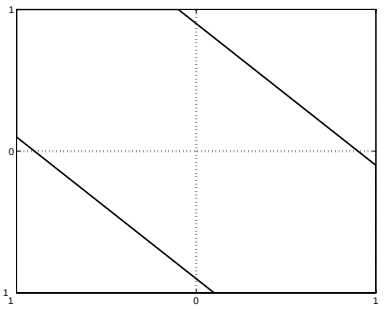


FIG. 1. *The largest domain of attraction when* $\lambda = .9$.

Since $r(t) = 0$ is an admissible reference signal, any tracking domain of attraction is a domain of attraction. The main result of this paper is that of showing that every domain of attraction $\mathcal{P}$ is also a tracking domain of attraction. The importance of this assertion lies in the fact that the tracking problem can be solved once one has found a domain of attraction. Since the latter operation is a well-established topic, this allows for the solution of a more general problem.

REMARK 2.1. *Note that under Assumption* 2.1 *the condition* $y(t) \to r_\infty$ *as* $t \to \infty$ *is equivalent to the two conditions* $x(t) \to \bar{x}_\infty \doteq \bar{x}(r_\infty)$ *and* $u(t) \to \bar{u}_\infty \doteq \bar{u}(r_\infty)$.

2.1. A simple example. Consider the discrete-time system

$$A = \begin{bmatrix} 1 & -1 \\ 1 & 1 \end{bmatrix} \quad B = \begin{bmatrix} 1 \\ 0 \end{bmatrix} \quad C = \begin{bmatrix} 0 & 1 \end{bmatrix},$$

together with the constraints sets $\mathcal{X} = \{x : \|x\|_\infty \leq 1\}$ and $\mathcal{U} = \{u : |u| \leq 1.2\}$. The largest domain of attraction with speed of convergence λ for $\lambda \geq 0.9$ is given by

$$\mathcal{P} = \{x : |x_1| \leq 1,\ |x_2| \leq 1,\ |x_1 + x_2| \leq \lambda\}$$

and is reported in Figure 1 for $\lambda = .9$. Such a domain is associated with the polyhedral function $\psi(x) = \max\{|x_1|, |x_2|, |x_1 + x_2|/\lambda\}$.

If we take $\lambda < 1$ we assure that $\mathcal{P}$ is a domain of attraction. Note that for λ approaching 1 the set $\mathcal{P}$ approaches the largest invariant set. Therefore there is a tradeoff between the choice of the largest invariant set and the convergence speed. Once this tradeoff is fixed by a choice of λ, we have that there exists a stabilizing control which avoids constraints violation for all initial states inside $\mathcal{P}$. Now if we want to track a reference signal $r(t)$ the situation is different. For a given fixed $\bar{r}$, the corresponding state input pair is

$$\begin{bmatrix} \bar{x} \\ \bar{u} \end{bmatrix} = \begin{bmatrix} A - I & B \\ C & 0 \end{bmatrix}^{-1} \begin{bmatrix} 0 \\ \bar{r} \end{bmatrix}.$$

If our goal were asymptotic tracking only, in view of the linearity of the system we could just use the translated Lyapunov function

$$\psi(x - \bar{x}) = \max\{|x_1 - \bar{x}_1|, |x_2 - \bar{x}_2|, |x_1 - \bar{x}_1 + x_2 - \bar{x}_2|/\lambda\}.$$

But in this way the constraints could not be satisfied. The basic idea we pursue is to "deform" the function ψ in such a way that the surface of level one is unchanged and the minimum point (the zero) is assumed on $\bar{x}$.

The set $\mathcal{R}_\epsilon$ is the interval

$$\mathcal{R}_\epsilon = \{r : \ |r| \leq \lambda - \epsilon\}.$$

For every reference signal $r(t) \to \bar{r} \in \mathcal{R}_\epsilon$ and $x(0) \in \mathcal{P}$, our goal is that of letting $y(t) \to \bar{r}$ and $x(t) \in \mathcal{P}$. According to Remark 2.1, for the first condition it is necessary and sufficient that $x(t) \to \bar{x} = [0 \ \ r]^T$ and $u(t) \to r$ (in this case $\bar{u}(r) = r$).

3. Main results. We will now introduce the functions which will subsequently be used for tracking every admissible reference signal. Suppose that a C-set $\mathcal{P}$, which is a domain of attraction, is given and consider its Minkowski functional which, from (2.1), can be rewritten as

$$\psi_{\mathcal{P}}(x) = \inf\{\alpha > 0 \ : \ \frac{1}{\alpha} x \in \mathcal{P}\}. \tag{3.1}$$

For every $\bar{x} \in \text{int}\mathcal{P}$ and $x \in \mathcal{P}$, we introduce the following function:

$$\Psi_{\mathcal{P}}(x, \bar{x}) = \inf\{\alpha > 0 \ : \ \bar{x} + \frac{1}{\alpha}(x - \bar{x}) \in \mathcal{P}\}. \tag{3.2}$$

It is immediate that the just introduced function Ψ recovers the values of ψ when $\bar{x} = 0$, say $\Psi_{\mathcal{P}}(x, 0) = \psi_{\mathcal{P}}(x)$. For fixed $\bar{x}$, $\Psi_{\mathcal{P}}(x, \bar{x})$ is convex. Furthermore, the function $\Psi_{\mathcal{P}}(x, \bar{x})$ for $(x, \bar{x}) \in \mathcal{P} \times \text{int}\mathcal{P}$ is such that

$$\Psi_{\mathcal{P}}(x, \bar{x}) = 0 \quad \text{iff} \quad x = \bar{x}, \tag{3.3}$$

$$\Psi_{\mathcal{P}}(x, \bar{x}) < 1 \quad \text{iff} \quad x \in \mathcal{P}, \tag{3.4}$$

$$\Psi_{\mathcal{P}}(x, \bar{x}) = 1 \quad \text{iff} \quad x \in \partial\mathcal{P}. \tag{3.5}$$

A sketch of the function $\Psi_{\mathcal{P}}(x, \bar{x})$ for fixed $\bar{x}$ is in Fig. 2. One further relevant property of this function is that $\Psi_{\mathcal{P}}$ is Lipschitz in x and positively homogeneous of order one with respect to the variable $z = x - \bar{x} \in \mathbb{R}^n$, $\bar{x} \in \text{int}\mathcal{P}$, i.e.,

$$\Psi_{\mathcal{P}}(\xi z + \bar{x}, \bar{x}) = \xi \Psi_{\mathcal{P}}(z + \bar{x}, \bar{x}). \tag{3.6}$$

In view of property (3.3) we have that the function is a suitable Lyapunov candidate for tracking, and from (3.4)–(3.5) we have that this function is suitable to prevent constraints violations, as we will show later.

Let us now consider the function $\Psi_{\mathcal{P}}(x, \bar{x})$ and for every $x \in \mathcal{P}$ and $\bar{x} \in \text{int}\mathcal{P}$, with $x \neq \bar{x}$, set

$$\tilde{x} \doteq \bar{x} + (x - \bar{x}) \frac{1}{\Psi_{\mathcal{P}}(x, \bar{x})} \in \partial\mathcal{P}.$$

The vector $\tilde{x}$ is the intersection of $\partial\mathcal{P}$ with the half line starting from $\bar{x}$ and passing through x (see Fig. 2). Assume that $\mathcal{P}$ is a given domain of attraction and that $\phi(x)$ is the control law associated with this set. As shown in [4], [7], there exists always a positively homogeneous control of order 1 which can be associated with such a domain. Therefore, without restriction, we introduce the following assumption.

ASSUMPTION 3.1. *The stabilizing control law $\phi(x)$ associated with the domain of attraction $\mathcal{P}$ is Lipschitz and positively homogeneous of order* 1*, i.e., $\phi(\alpha x) = \alpha\phi(x)$ for $\alpha \geq 0$ and $x \in \mathcal{P}$.*

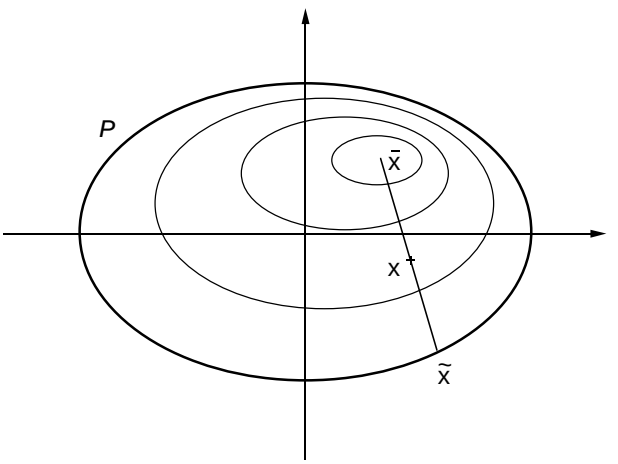


FIG. 2. *The function $\Psi_{\mathcal{P}}(x,\bar{x})$ for fixed $\bar{x}$.*

The next step in the derivation of a feedback control law $\Phi(x,r)$ is the definition of a saturation map $\Gamma : \mathbb{R}^q \to \mathcal{R}_\epsilon$ as follows:

$$\Gamma(r) = \begin{cases} r\psi_{\mathcal{R}_\epsilon}(r)^{-1} & \text{when} \quad \psi_{\mathcal{R}_\epsilon}(r) > 1 \\ r & \text{otherwise} \end{cases}$$

say

$$\Gamma(r) = r * sat\left(\frac{1}{\psi_{\mathcal{R}_\epsilon}(r)}\right),$$

where $sat(\cdot)$ is the saturation function. Note that $\Gamma(r)$ is the identity if we restrict $r \in \mathcal{R}_\epsilon$. Conversely, for $r \notin \mathcal{R}_\epsilon$, $\Gamma(r)$ is the intersection of $\partial\mathcal{R}_\epsilon$ and the segment having extrema 0 and r.

The control we propose has the form

$$\Phi(x,r) = \phi(\tilde{x})\Psi_{\mathcal{P}}(x,\bar{x}) + (1 - \Psi_{\mathcal{P}}(x,\bar{x}))\bar{u}, \tag{3.7}$$

where

$$\begin{bmatrix} \bar{x}(\bar{r}) \\ \bar{u}(\bar{r}) \end{bmatrix} \doteq M_d^{-1} \begin{bmatrix} 0 \\ \bar{r} \end{bmatrix} \left(\text{resp.,} \begin{bmatrix} \bar{x}(\bar{r}) \\ \bar{u}(\bar{r}) \end{bmatrix} \doteq M_c^{-1} \begin{bmatrix} 0 \\ \bar{r} \end{bmatrix}\right) \tag{3.8}$$

and

$$\bar{r} = \Gamma(r). \tag{3.9}$$

Note that, for $r \in \mathcal{R}_\epsilon$, (3.9) does not play any role. Note also that, since $\bar{r} = \Gamma(r) \in \mathcal{R}_\epsilon$, then $\bar{x} \in int\mathcal{P}$, thus the term $\Psi_{\mathcal{P}}(x,\bar{x})$ in (3.7) is defined. However, the expression (3.7) is not defined for $x = \bar{x}$ because of the critical term $\phi(\tilde{x})\Psi_{\mathcal{P}}(x,\bar{x})$. Nevertheless, in view of the homogeneity of ϕ and from the expression of $\tilde{x}$, we have that

$$\phi(\tilde{x})\Psi_{\mathcal{P}}(x,\bar{x}) = \phi(\Psi_{\mathcal{P}}(x,\bar{x})\tilde{x}) = \phi\left(x + (\Psi_{\mathcal{P}}(x,\bar{x}) - 1)\bar{x}\right). \tag{3.10}$$

Then $\phi(\tilde{x})\Psi_{\mathcal{P}}(x,\bar{x}) \to 0$ as $x \to \bar{x}$ so that we can extend the function by continuity by assuming

$$\phi(\tilde{x})\Psi_{\mathcal{P}}(\bar{x},\bar{x}) = 0.$$

The introduced control law inherits most of the properties from $\phi(x)$ according to the next proposition which assures existence and uniqueness of the solution of (2.2) when the control $\Phi(x,r)$ is used, provided that the admissible reference signal $r(t)$ is measurable.

PROPOSITION 3.1. *Suppose $\phi(x)$ is Lipschitz and homogeneous of order 1. Then $\Phi(x,r):\mathcal{P}\times\mathbb{R}^q\to\mathcal{U}$, defined as in (3.7)–(3.9), is continuous and it is Lipschitz with respect to x.*

To have an idea of how this control works, note that as long as the condition

$$\Psi_{\mathcal{P}}(x(t),\bar{x}(t))\le 1 \tag{3.11}$$

is satisfied, we have that $x(t)\in\mathcal{P}$. Moreover, for $\bar{x}\in int\mathcal{P}$, the control is just a *convex combination* of the control $\bar{u}(\bar{r})$ and $\phi(\tilde{x})$. By construction, $\bar{u}(\bar{r})\in\mathcal{U}$ and $\phi(\tilde{x})\in\mathcal{U}$; thus $\Phi(x,r)\in\mathcal{U}$.

Our effort will be devoted to proving that the proposed control law guarantees (3.11) as well as the limit condition

$$\Psi_{\mathcal{P}}(x(t),\bar{x}(r_\infty))\to 0, \tag{3.12}$$

where $\bar{x}(r_\infty)$ is the steady state associated with $r_\infty\in\mathcal{R}_\epsilon$ (note that $\Gamma(r_\infty)=r_\infty$). Indeed such a limit condition implies $x(t)\to\bar{x}(r_\infty)$ and, from (3.7) and (3.10), $\Phi(x(t),r(t))\to\Phi(\bar{x}(r_\infty),r_\infty)=\bar{u}(r_\infty)$. Therefore if (3.12) holds, we have that $y(t)\to r_\infty$.

For evident practical reasons, we introduce an extended concept of speed of convergence, appropriate for the condition (3.12). In the continuous-time case, given a *fixed* $\bar{x}\in int\mathcal{P}$ we say that the speed of convergence is $\beta>0$ if the Lyapunov derivative $\dot{\Psi}_{\mathcal{P}}(x(t),\bar{x})$ of $\Psi_{\mathcal{P}}(x(t),\bar{x})$ is such that

$$\dot{\Psi}_{\mathcal{P}}(x,\bar{x},u)\doteq\lim_{h\to 0^+}\frac{\Psi_{\mathcal{P}}(x+h(Ax+Bu),\bar{x})-\Psi_{\mathcal{P}}(x,\bar{x})}{h}\le-\beta\Psi_{\mathcal{P}}(x,\bar{x}),$$

where the existence of the limit is assured by the convexity of $\Psi_{\mathcal{P}}$ with respect to x. In the discrete-time case we say that the speed of convergence is $\lambda<1$ if

$$\Psi_{\mathcal{P}}(Ax+Bu,\bar{x})\le\lambda\Psi_{\mathcal{P}}(x,\bar{x}).$$

We start by considering a special case, namely the one in which the reference signal is constant. In this case we can show that, if we have a domain of attraction to the origin with a certain speed of convergence, we can achieve the tracking goal without constraints violation for all the initial states in such a domain. Furthermore, for symmetric domains, we can guarantee a speed of convergence which is *independent of the reference input.*

THEOREM 3.2. *Let $\mathcal{P}$ be a domain of attraction with speed of convergence λ for the discrete-time dynamic system (2.2) associated with the control $\phi(x)$ satisfying Assumption 3.1. Then, for every admissible constant reference signal $r(t)=\bar{r}$, the control law (3.7)–(3.8) is such that for every initial condition $x(0)\in\mathcal{P}$ we have that $x(t)\in\mathcal{P}$ and $u(t)\in\mathcal{U}$ for every $t\ge 0$ and $\lim_{t\to\infty}y(t)=\bar{r}$. Moreover, if $\mathcal{P}$ is 0-symmetric, the speed of convergence $\lambda_{TR}=\frac{\lambda+1}{2}$ is guaranteed.*

The next theorem is the continuous-time version of the above.

THEOREM 3.3. *Let $\mathcal{P}$ be a domain of attraction with speed of convergence β for the continuous-time dynamic system (2.2) associated with the control $\phi(x)$ satisfying Assumption 3.1. Then, for every admissible constant reference signal $r(t)=\bar{r}$, the*

control law (3.7)–(3.8) *is such that for every initial condition* $x(0) \in \mathcal{P}$, *we have that* $x(t) \in \mathcal{P}$ *and* $u(t) \in \mathcal{U}$ *for every* $t \geq 0$ *and* $\lim_{t\to\infty} y(t) = \bar{r}$. *Moreover, if* $\mathcal{P}$ *is* 0*-symmetric, the speed of convergence is at least* $\beta_{TR} = \frac{\beta}{2}$.

The proposed control law can be successfully used even when the reference $r(t)$ is allowed to vary provided that it is admissible according to Definition 2.4.

THEOREM 3.4. *Let* $r(t)$ *be admissible as in Definition* 2.4. *Any domain of attraction* $\mathcal{P}$, *with speed of convergence* $\beta > 0$ $(0 \leq \lambda < 1)$, *for system* (2.2) *is a tracking domain of attraction. Moreover, the control law in* (3.7)–(3.9) *assures the conditions* (i) *and* (ii) *of Definition* 2.5.

REMARK 3.1. *If we consider a constant reference* $r \in \mathcal{R}_\epsilon$ *and the corresponding steady state vectors derived from* $\bar{x}$ *and* $\bar{u}$ *by means of* (3.8), *then we can apply a state and control translation by considering the system* $\delta\hat{x} = A\hat{x} + B\hat{u}$ *with the new constraints* $\hat{u} = u - \bar{u} \in \mathcal{U} - \bar{u} = \hat{\mathcal{U}}$ *and* $\hat{x} = x - \bar{x} \in \mathcal{X} - \bar{x} = \hat{\mathcal{X}}$. *From this algebraic point of view, our result amounts to proving that the largest domain of attraction of the translated problem is just achieved by translating the original largest domain of attraction as* $\hat{\mathcal{P}} = \mathcal{P} - \bar{x}$.

4. Properties of the function $\Psi(x, \bar{x})$. To prove the results of the previous section we first need to present some properties of the function $\Psi(x, \bar{x})$. First we prove its continuity. For the simple notation we will drop the index $\mathcal{P}$.

LEMMA 4.1. *For* $(x, \bar{x}) \in \mathcal{P} \times int\mathcal{P}$, *the function* $\Psi(x, \bar{x})$ *is continuous.*

Proof. Consider the case $x \neq \bar{x}$. Take any point $(x, \bar{x})$ and a close pair $(x', \bar{x}')$. Assume that $x' \in V_x$ and $\bar{x}' \in V_{\bar{x}} \subset int\mathcal{P}$ where V_x and $V_{\bar{x}}$ are close neighborhoods of x and $\bar{x}$ such that $V_x \bigcap V_{\bar{x}} = \emptyset$, the empty set. This implies that $\|x' - \bar{x}'\| \geq M$ for some positive M. Assume now $x' \in \mathcal{P}$ and take $\alpha > 0$ (which exists because $\bar{x}' \in int\mathcal{P}$) such that $\bar{x}' + (x' - \bar{x}')/\alpha \in \mathcal{P}$. Then $\|(x' - \bar{x}')/\alpha\| \leq m$, for some positive m. This means that $\alpha \geq M/m$; thus it is lower bounded. From (3.2) we have

$$\Psi(x', \bar{x}') = \inf\left\{\alpha > 0 : \ \bar{x}' + \frac{1}{\alpha}(x' - \bar{x}') \in \mathcal{P}\right\} \tag{4.1}$$

$$= \inf\left\{\alpha > 0 : \ \bar{x} + \frac{1}{\alpha}(x - \bar{x}) \in \mathcal{P} + \underbrace{(\bar{x} - \bar{x}')\left(1 - \frac{1}{\alpha}\right) + \frac{1}{\alpha}(x - x')}_{=z}\right\},$$

where $\mathcal{P} + z \doteq \{y = x + z, \quad x \in \mathcal{P}\}$. Since α as in (4.1) is lower bounded by M/m, then $\|z\| \to 0$ as $(x', \bar{x}') \to (x, \bar{x})$.

Moreover, $\mathcal{P}$ is a C-set; thus for every $0 < \hat{\epsilon} < 1$ there exists δ such that if $\|x - x'\| \leq \delta$ and $\|\bar{x} - \bar{x}'\| \leq \delta$, then

$$(1 - \hat{\epsilon})\mathcal{P} \subset \mathcal{P} + z \subset (1 + \hat{\epsilon})\mathcal{P}. \tag{4.2}$$

Consider the following function:

$$\alpha(\epsilon) = \inf\left\{\alpha > 0 : \quad \bar{x} + \frac{1}{\alpha}(x - \bar{x}) \in (1 + \epsilon)\mathcal{P}\right\}.$$

The inclusion (4.2) assures that for $\hat{\epsilon}$ and δ as above

$$\alpha(\hat{\epsilon}) \leq \Psi(x', \bar{x}') \leq \alpha(-\hat{\epsilon}),$$

and from (3.2) we have

$$\alpha(\hat{\epsilon}) \leq \Psi(x, \bar{x}) \leq \alpha(-\hat{\epsilon}).$$

Thus, if we show that $\alpha(\epsilon)$ is continuous at 0, we have $|\alpha(\hat{\epsilon}) - \alpha(-\hat{\epsilon})| \to 0$, as $\hat{\epsilon} \to 0$, and we have completed the proof.

Since $\bar{x} \in int\mathcal{P}$, there exists $0 < \bar{\epsilon} < 1$ such that $\bar{x} \in int\{(1+\epsilon)\mathcal{P}\}$ for all $|\epsilon| < \bar{\epsilon}$.

The set of all points of the form $\bar{x} + (x - \bar{x})/\alpha$, $\alpha > 0$ is a ray originating from $\bar{x}$ which, for $|\epsilon| \leq \bar{\epsilon}$, intersects the boundary of $(1+\epsilon)\mathcal{P}$ at the unique point $\bar{x} + (x - \bar{x})/\alpha(\epsilon)$. We can write this condition as

$$\psi\left(\bar{x} + (x - \bar{x})/\alpha(\epsilon)\right) = 1 + \epsilon. \tag{4.3}$$

Note also that for $\alpha > \alpha(\epsilon)$ the point $\bar{x} + (x - \bar{x})/\alpha \in int\{(1+\epsilon)\mathcal{P}\}$; thus the function $\alpha(\epsilon)$ is strictly decreasing on $[-\bar{\epsilon}, \bar{\epsilon}]$, and so it is invertible. From (4.3) its inverse is the function

$$\epsilon(\alpha) \doteq \psi\left(\bar{x} + (x - \bar{x})/\alpha\right) - 1$$

which is continuous (actually convex). Then $\alpha(\epsilon)$ is continuous.

In the case $x = \bar{x} \in int\mathcal{P}$, say $\Psi(x, \bar{x}) = 0$ it is very easy to show from (3.2) that $\Psi(x', \bar{x}') \to 0$ as $(x', \bar{x}') \to (\bar{x}, \bar{x})$, and we skip the proof for brevity. □

The function $\Psi(x, \bar{x})$ is not necessarily differentiable and then, to use it as a candidate Lyapunov function, we invoke the theory of subdifferentials. Given a convex function $F(x)$, the subdifferential of F in x is defined as the set [23]

$$\nabla F(x) = \{w : \langle w, z - x\rangle \leq F(z) - F(x) \quad \text{for all} \quad z \in \mathbb{R}^n\}.$$

Each element w of ∇F is called subgradient. In what follows we denote by $\nabla\Psi(x, \bar{x})$ the subdifferential of the function $\Psi(x, \bar{x})$ with respect to x (i.e., $\bar{x}$ is assumed constant):

$$\nabla\Psi(x, \bar{x}) = \{w : \langle w, z - x\rangle \leq \Psi(z, \bar{x}) - \Psi(x, \bar{x})\}.$$

If $\Psi(x, \bar{x})$ is differentiable with respect to x, then $\nabla\Psi(x, \bar{x})$ is a singleton including the gradient vector

$$\nabla\Psi(x, \bar{x}) \doteq \left[\frac{\partial\Psi(x, \bar{x})}{\partial x_1}, \ldots, \frac{\partial\Psi(x, \bar{x})}{\partial x_n}\right]^T.$$

For constant $\bar{x}$, the Lyapunov derivative of the function $\Psi(x, \bar{x})$, along the system trajectory of the system (2.2), is given by

$$\dot{\Psi}(x, \bar{x}, u) = \lim_{\tau \to 0^+} \frac{\Psi(x + \tau(Ax + Bu), \bar{x}) - \Psi(x, \bar{x})}{\tau}. \tag{4.4}$$

From [23, Th. 23.4], we have that

$$\dot{\Psi}(x, \bar{x}, u) = \sup_{w \in \nabla\Psi(x, \bar{x})} \langle w, Ax + Bu\rangle. \tag{4.5}$$

This fact implies that the condition

$$\dot{\Psi}(x, \bar{x}, u) \leq -\beta\Psi(x, \bar{x})$$

is equivalent to

$$\langle w, Ax + Bu\rangle \;\leq\; -\beta\Psi(x, \bar{x}) \quad \text{for all} \quad w \in \nabla\Psi(x, \bar{x}).$$

The next lemma will show some fundamental properties concerning the subdifferentiability of Ψ. Its importance will be evident later when we deal with continuous-time systems.

LEMMA 4.2. *Assume that $\bar{x} \in int\mathcal{P}$ is fixed. Then*

(i) *Its subdifferential is constant over any ray emanating from* $\bar{x}$

$$\nabla\Psi(\bar{x}+\mu y,\bar{x}) = \nabla\Psi(\bar{x}+y,\bar{x}) \quad \textit{for all} \;\; \mu \geq 0; \tag{4.6}$$

(ii) *The equality*

$$\Psi(x,\bar{x}) = \langle \hat{w}, x-\bar{x}\rangle$$

holds for all $\hat{w} \in \nabla\Psi(x,\bar{x})$.

(iii) *For every* $x \in \partial\mathcal{P}$ *the positive cones generated by the subdifferentials* $\nabla\psi(x)$ *and* $\nabla\Psi(x,\bar{x})$ *coincide. This means that each* $w \in \nabla\psi(x)$ *is aligned with some* $\hat{w} \in \nabla\Psi(x,\bar{x})$ *and vice-versa. Furthermore, if* $\mathcal{P}$ *is 0-symmetric (i.e.,* $\mathcal{P} = -\mathcal{P}$*) for every* $\bar{x} \in (1-\epsilon)\mathcal{P}$ *and* $x \in \partial\mathcal{P}$*, the aligning factor* $\gamma(\hat{w},w)$ *such that*

$$\hat{w} = \gamma(\hat{w},w)w, \qquad \hat{w} \in \nabla\Psi(x,\bar{x}), \quad w \in \nabla\psi(x)$$

satisfies the inequality

$$\gamma(\hat{w},w) \geq \frac{1}{2-\epsilon} \geq \frac{1}{2}. \tag{4.7}$$

Proof. Property (i) is a straightforward derivation from the definition of the subdifferential once we have established that $\tilde{\psi}(y) = \Psi(\bar{x}+y,\bar{x})$ is a positively homogeneous function of order 1 according to (3.6). Property (ii) follows from the following properties of the positively homogeneous functions of order 1 (see the appendix)

$$\tilde{\psi}(y) = \langle w, y\rangle \quad \text{for all} \;\; w \in \nabla\tilde{\psi}(y). \tag{4.8}$$

To prove (iii), let $\bar{x} \in int\mathcal{P}$ be fixed. From (3.4)–(3.5) the unit balls of the functions $\Psi(x,\bar{x})$ and $\psi(x)$ are both equal to $\mathcal{P}$. For $x \in \partial\mathcal{P}$, the normal cone [13] to this set is given by

$$\mathcal{N} = \{z \in \mathbb{R}^n : \langle z, x'-x\rangle \leq 0 \;\; \text{for all} \;\; x' \in \mathcal{P}\}.$$

Since the normal cone for $x \in \partial\mathcal{P}$ is the convex positive cone generated by any of the two subdifferentials $\nabla\Psi(x,\bar{x})$ and $\nabla\psi(x)$, we have the alignment of any element of the former with an element of the latter and the existence of a positive real $\gamma(\hat{w},w)$ such that $\hat{w} = \gamma w$, with $\hat{w} \in \nabla\Psi(x,\bar{x})$ and $w \in \nabla\psi(x)$.

To prove (4.7) consider $x \in \partial\mathcal{P}$ so that $\Psi(x,\bar{x}) = 1$ with $\bar{x} \in (1-\epsilon)\mathcal{P}$. Let $\hat{w} \in \nabla\Psi(x,\bar{x})$ and $w \in \nabla\psi(x)$ be as above. From property (ii)

$$\Psi(x,\bar{x}) = 1 = \langle \hat{w}, x-\bar{x}\rangle = \gamma\langle w, x-\bar{x}\rangle.$$

Then, since $w \in \nabla\psi(x)$, $\psi(x) = \langle w,x\rangle = 1$ and by the symmetry of $\mathcal{P}$, $\psi(-\bar{x}) = \psi(\bar{x}) \leq 1-\epsilon$ we get

$$\frac{1}{\gamma} = \langle w, x-\bar{x}\rangle = \langle w, -\bar{x}-x\rangle + 2\langle w,x\rangle \tag{4.9}$$

$$\leq \psi(-\bar{x}) - \psi(x) + 2 = \psi(-\bar{x}) + 1 \leq 2-\epsilon \tag{4.10}$$

which is equivalent to (4.7). ∎

REMARK 4.1. *If symmetry is removed, it is possible to show that there exists* $m_\epsilon > 0$ *such that* $\gamma(\hat{w}, w) \geq m_\epsilon > 0$, *although no lower bounds can be provided a priori.*

We recall that if $\mathcal{P}$ is a convex set in a real normed space X, then it is possible to define on the dual space X^* the support functional $h(x^*)$ of $\mathcal{P}$

$$h(x^*) = \sup_{x \in \mathcal{P}} \langle x, x^* \rangle. \tag{4.11}$$

If the set contains the origin as an interior point, then $h(x^*)$ is positive definite, and if the set is compact, there exists a finite constant K such that $h(x^*) \leq K\|x^*\|$ for every x^*. Once $h(\cdot)$ is known, it is possible to describe a closed convex set as the intersection of all the half-spaces that contain it:

$$\mathcal{P} = \bigcap_{x^*} \{x : \langle x, x^* \rangle \leq h(x^*)\}. \tag{4.12}$$

For every x belonging to a C-set $\mathcal{P}$ we have that $-h(-x^*) \leq \langle x, x^* \rangle \leq h(x^*)$ for every $x^* \in X^*$. If $\mathcal{P}$ is 0-symmetric, then $h(-x^*) = h(x^*)$ for all x^*. The next lemma characterizes $\Psi(x, \bar{x})$ in terms of h.

LEMMA 4.3. *Let* $h(\cdot)$ *be the support functional of* $\mathcal{P}$. *Then*

$$\Psi(x, \bar{x}) = \sup_{\|x^*\|=1} \frac{\langle x^*, x - \bar{x} \rangle}{h(x^*) - \langle x^*, \bar{x} \rangle}. \tag{4.13}$$

Proof. By a proper scaling of the vectors x^* in (4.12), we know that $x \in \mathcal{P}$ iff $\langle x, x^* \rangle \leq h(x^*)$ for every x^* such that $\|x^*\| = 1$. Then from (3.2)

$$\begin{aligned} \Psi(x, \bar{x}) &= \inf \left\{ \alpha \ : \langle \bar{x} + \frac{1}{\alpha}(x - \bar{x}), x^* \rangle \ \leq \ h(x^*), \ \text{ for all } \ \|x^*\| = 1 \right\} \\ &= \inf \left\{ \alpha \ : \frac{1}{\alpha} \langle (x - \bar{x}), x^* \rangle \ \leq \ h(x^*) - \langle \bar{x}, x^* \rangle, \ \text{ for all } \ \|x^*\| = 1 \right\}. \end{aligned}$$

from which follows (4.13). ∎

The next lemma we present gives an upper bound to the values of the function $\Psi(x, \bar{x})$ when $(x, \bar{x}) \in \alpha\mathcal{P} \times int\mathcal{P}$. It will be useful to guarantee a convergence speed in the discrete-time case.

LEMMA 4.4. *Suppose* $\mathcal{P}$ *is* 0*-symmetric and let* $x \in \alpha\mathcal{P}$ *and* $\bar{x} \in int\mathcal{P}$. *Then*

$$\Psi(x, \bar{x}) \leq \frac{\alpha + 1}{2}. \tag{4.14}$$

Proof. From Lemma 4.3 we need to prove that

$$\frac{\langle x^*, x - \bar{x} \rangle}{h(x^*) - \langle x^*, \bar{x} \rangle} \leq \frac{1 + \alpha}{2} \tag{4.15}$$

for every x^* such that $\|x^*\| = 1$. For every such x^* we have that $\langle x, x^* \rangle \leq \alpha h(x^*)$ whenever $x \in \alpha\mathcal{P}$. Thus

$$\begin{aligned} \frac{\langle x^*, x - \bar{x} \rangle}{h(x^*) - \langle x^*, \bar{x} \rangle} &\leq \frac{\alpha h(x^*) - \langle x^*, \bar{x} \rangle}{h(x^*) - \langle x^*, \bar{x} \rangle} \\ &= \frac{\alpha - \frac{\langle x^*, \bar{x} \rangle}{h(x^*)}}{1 - \frac{\langle x^*, \bar{x} \rangle}{h(x^*)}} \end{aligned} \tag{4.16}$$

for every $\|x^*\| = 1$ (note that $h(x^*) \neq 0$). Now we notice that, since $\mathcal{P}$ is 0-symmetric,

$$-1 \leq \frac{\langle x^*, \bar{x} \rangle}{h(x^*)} \leq 1.$$

Thus expression (4.16) is bounded by the extrema of the function

$$f(\xi) = \frac{\alpha - \xi}{1 - \xi} \quad \text{with} \quad |\xi| \leq 1.$$

Simple algebra shows that $f(\xi)$ is continuous and decreasing thus attaining its maximum at $\xi = -1$. This proves (4.15). □

REMARK 4.2. *The previous lemma gives a specific upper bound for $\Psi(x, \bar{x})$ for $\bar{x} \in int\mathcal{P}$ and $x \in \alpha\mathcal{P}$. If symmetry is removed, we have that there exists $m_\alpha < 1$ such that $\Psi(x, \bar{x}) \leq m_\alpha < 1$, although no upper bounds can be guaranteed a priori.*

5. Proofs of the main results. For brevity, the next proofs will consider the case of a symmetric $\mathcal{P}$ only. The general proofs follow immediately by considering Remarks 4.1 and 4.2.

Proof of Proposition 3.1. In view of the homogeneity of $\phi(x)$ we have that

$$\begin{aligned} \Phi(x, r) &= \phi(\tilde{x})\Psi(x, \bar{x}) + (1 - \Psi(x, \bar{x}))\bar{u} \\ &= \phi(\Psi(x, \bar{x})\bar{x} + x - \bar{x}) + (1 - \Psi(x, \bar{x}))\bar{u}. \end{aligned} \tag{5.1}$$

By exploiting the Lipschitz condition on $\phi(x)$ and the triangular inequality we get

$$\begin{aligned} \|\Phi(x_1, r) - \Phi(x_2, r)\| \leq L\|\Psi(x_1, \bar{x})\bar{x} + x_1 - \bar{x} - \Psi(x_2, \bar{x})\bar{x} - x_2 + \bar{x}\| \\ + \|(1 - \Psi(x_1, \bar{x}))\bar{u} - (1 - \Psi(x_2, \bar{x}))\bar{u}\|; \end{aligned}$$

therefore, after proper rearrangement,

$$\|\Phi(x_1, r) - \Phi(x_2, r)\| \leq L\|x_1 - x_2\| + (L\|\bar{x}\| + \|\bar{u}\|)\|\Psi(x_1, \bar{x}) - \Psi(x_2, \bar{x})\|.$$

Let $L_\mathcal{P}$ be the Lipschitz constant of the function $\Psi(x, \bar{x})$ and let K be a constant such that $L\|\bar{x}\| + \|\bar{u}\| \leq K$ so that we can finally obtain

$$\|\Phi(x_1, r) - \Phi(x_2, r)\| \leq (L + L_\mathcal{P} K)\|x_1 - x_2\|.$$

The Lipschitz constant of $\Phi(x, r)$ is then $L_\Phi = (L + L_\mathcal{P} K)$.

The continuity of $\Phi(x, r)$ can be inferred easily by considering expression (5.1) which, in view of Lemma 4.1 and the Lipschitz condition of ϕ, is continuous with respect to x, $\bar{x}$, and $\bar{u}$. Now $\bar{x}$ and $\bar{u}$ are linear functions of $\bar{r} = \Gamma(r)$, and Γ is continuous. This completes the proof. □

Proof of Theorem 3.2. We have already noticed that $\Phi(x, r) \in \mathcal{U}$, which is an immediate consequence of the convexity of $\mathcal{U}$ and the fact that the control $\Phi(x, r)$ as in (3.7)–(3.9) is the convex combination of $\phi(\tilde{x})$ and $\bar{u}$, both belonging to $\mathcal{U}$. The condition $x \in \mathcal{P}$ for every $t \geq 0$ will be proved by showing that $\Psi(x, \bar{x})$ is a Lyapunov function for system (2.2). Consider $x \in \mathcal{P}$, $x \neq \bar{x}$ and let

$$\begin{aligned} y &= Ax + B\Phi(x, r) \\ &= A(\tilde{x}\Psi(x, \bar{x}) + (1 - \Psi(x, \bar{x}))\bar{x}) + B(\phi(\tilde{x})\Psi(x, \bar{x}) + (1 - \Psi(x, \bar{x})\bar{u}) \\ &= \Psi(x, \bar{x})(A\tilde{x} + B\phi(\tilde{x})) + (1 - \Psi(x, \bar{x})(A\bar{x} + B\bar{u}) \\ &= \Psi(x, \bar{x})(A\tilde{x} + B\phi(\tilde{x})) + (1 - \Psi(x, \bar{x}))\bar{x}, \end{aligned}$$

where we have replaced $A\bar{x} + B\bar{u} = \bar{x}$ due to the fact that $(\bar{x}, \bar{u})$ is an equilibrium pair. Define now $\tilde{y} = A\tilde{x} + B\phi(\tilde{x})$, since $\mathcal{P}$ is domain of attraction as in definition 2.2. Then $\psi(A\tilde{x} + B\phi(\tilde{x})) \leq \lambda\psi(\tilde{x}) = \lambda$, and thus $\tilde{y} \in \lambda\mathcal{P}$. By virtue of (3.6),

$$\begin{aligned}\Psi(y, \bar{x}) &= \Psi(\tilde{y}\Psi(x, \bar{x}) + (1 - \Psi(x, \bar{x}))\bar{x}, \bar{x}) \\ &= \Psi((\tilde{y} - \bar{x})\Psi(x, \bar{x}) + \bar{x}, \bar{x}) \\ &= \Psi(x, \bar{x})\Psi(\tilde{y}, \bar{x}) \leq \frac{1 + \lambda}{2}\Psi(x, \bar{x}),\end{aligned}$$

where the inequality comes from Lemma 4.4. Thus

$$\Psi(Ax + B\Phi(x, r), \bar{x}) \leq \lambda_{TR}\Psi(x, \bar{x}). \qquad \square$$

Proof of Theorem 3.3 Let $x \in \mathcal{P}$ and $\bar{x} \in (1 - \epsilon)\mathcal{P}$. From (4.5) we need to prove that $\dot{\Psi}(x, \bar{x}) \leq -\beta/2\Psi(x, \bar{x})$, which is true if

$$\langle \hat{w}, Ax + B\Phi(x, r)\rangle \leq -\frac{\beta}{2}\Psi(x, \bar{x}) < 0 \tag{5.2}$$

is satisfied for all $\hat{w} \in \nabla\Psi(x, \bar{x})$ and $x \neq \bar{x}$. Let, then, $x \neq \bar{x}$. Note that

$$\begin{aligned}&\langle \hat{w}, Ax + B\Phi(x, r)\rangle \\ &= \langle\ \hat{w}\ ,\ A(\Psi(x, \bar{x})\tilde{x} + (1 - \Psi(x, \bar{x}))\bar{x}) + B(\Psi(x, \bar{x})\phi(\tilde{x}) + (1 - \Psi(x, \bar{x}))\bar{u}\ \rangle \\ &= \langle\ \hat{w}\ ,\ [A\tilde{x} + B\phi(\tilde{x})]\,\Psi(x, \bar{x}) + (1 - \Psi(x, \bar{x}))\,[A\bar{x} + B\bar{u}]\ \rangle.\end{aligned}$$

We remind the reader that in the continuous-time case we have that $A\bar{x} + B\bar{u} = 0$ and that (i) of Lemma 4.2 holds so that $\hat{w} \in \nabla\Psi(x, \bar{x})$, iff $\hat{w} \in \nabla\Psi(\tilde{x}, \bar{x})$. Since $\tilde{x} \in \partial\mathcal{P}$, then, by virtue of (iii) of Lemma 4.2, $\hat{w} = \gamma(\hat{w}, w)w$, with $w \in \nabla\psi(\tilde{x})$ and $\gamma(\hat{w}, w) \geq 1/(2 - \epsilon)$. Therefore

$$\begin{aligned}\langle \hat{w}, Ax + B\Phi(x, r)\rangle &= \Psi(x, \bar{x})\langle \hat{w}, A\tilde{x} + B\phi(\tilde{x})\rangle \\ &= \gamma(\hat{w}, w)\ \langle w, A\tilde{x} + B\phi(\tilde{x})\rangle\ \Psi(x, \bar{x}).\end{aligned}$$

Due to the fact that $\mathcal{P}$ is a domain of attraction, we have that

$$\langle w, A\tilde{x} + B\phi(\tilde{x})\rangle \leq -\beta \quad \text{for all} \quad w \in \nabla\psi(x).$$

From (iii) of Lemma 4.2 we have that, if $\mathcal{P}$ is symmetric, $\gamma(\hat{w}, w) \geq 1/(2 - \epsilon) \geq 1/2$, and thus we have (5.2). $\square$

REMARK 5.1. *According to Remarks* 4.1 *and* 4.2*, for nonsymmetric domains we have that* $\lambda_{TR} = m_\alpha < 1$ *and* $\beta_{TR} = \beta m_\epsilon > 0$.

Proof of Theorem 3.4 We will consider the control law (3.7)–(3.9) and we will prove the result for the continuous-time case only, as the discrete-time version can be obtained by following a similar argument.

To show that the proposed control law maintains the state inside $\mathcal{P}$ for every $x(0) \in \mathcal{P}$, we need to show its positive invariance for the closed-loop system. But this is immediate because we can just consider the original Lyapunov function $\psi(x)$. For $x \in \partial\mathcal{P}$ we have that $\Phi(x, r) = \phi(x)$, and thus the Lyapunov derivative of the system with the tracking control is

$$\dot{\psi}(x(t)) = \sup_{w \in \nabla\psi(x)} \langle w, Ax + B\phi(x)\rangle,$$

which is the same as the Lyapunov derivative of the system with original stabilizing control associated with $\mathcal{P}$. By Definition 2.3, the derivative for $x \in \partial\mathcal{P}$ is such that

$$\dot{\psi}(x) \leq -\beta,$$

which obviously prevents the state $x(t)$ from leaving $\mathcal{P}$ [1].

The fact that the control $\Phi(x, r) \in \mathcal{U}$ has been shown at the beginning of the proof of Theorem 3.2 (we remind the reader that $\bar{u}$ is computed with respect to the "saturated" reference $\bar{r} = \Gamma(r)$).

To complete the proof we need now to show that $y(t)$ approaches r_∞ when $t \to \infty$. In view of Remark 2.1, this amounts to showing that $x(t) \to \bar{x}_\infty$, say $\Psi(x, \bar{x}_\infty) \to 0$.

Let $\Phi(x(t), r(t))$ and $\Phi(x(t), r_\infty)$ be the control actions associated with the reference values $r(t)$ and r_∞, resp., so that we can write

$$\dot{x} = Ax + B\Phi(x, r) = Ax + B\Phi(x, r_\infty) + B[\Phi(x, r) - \Phi(x, r_\infty)].$$

To show that $x(t) \to \bar{x}_\infty$ we consider the candidate Lyapunov function $\Psi(x(t), \bar{x}_\infty)$. Take any vector $w \in \nabla\Psi(x(t), \bar{x}_\infty)$. Then

$$\begin{aligned}\langle w, Ax + B\Phi(x, r)\rangle &= \langle w, Ax + B\Phi(x, r_\infty)\rangle + \underbrace{\langle w, B[\Phi(x, r) - \Phi(x, r_\infty)]\rangle}_{z} \\ &\leq -\frac{\beta}{2}\Psi(x(t), \bar{x}_\infty) + |z(t)|,\end{aligned}$$

where we have exploited the condition $\langle w, Ax + B\Phi(x, r_\infty)\rangle \leq -\frac{\beta}{2}\Psi(x(t), \bar{x}_\infty)$ proven in Theorem 3.3. As the above inequality holds for all $w \in \nabla\Psi(x(t), \bar{x}_\infty)$, from (4.5) we get

$$\dot{\Psi}(x(t), \bar{x}_\infty, \Phi(x(t), r(t))) \leq -\beta/2\Psi(x(t), \bar{x}_\infty) + |z(t)|.$$

The vector $w \in \nabla\Psi(x(t), \bar{x}_\infty)$ is bounded due to the fact that $\Psi(x, \bar{x}_\infty)$ is Lipschitz with respect to the first argument x. Since $r(t) \to r_\infty \in \mathcal{R}_\epsilon$, there exists T_1 such that $r(t) \in \mathcal{R}_{\frac{\epsilon}{2}}$ for $t \geq T_1$. $\Phi(x, r)$ is continuous, thus it is uniformly continuous on $\mathcal{P} \times \mathcal{R}_\epsilon$. This means that $z(t) \to 0$ as $\bar{r}(t) \to r_\infty$. Standard Lyapunov arguments lead to the conclusion that $\Psi(x(t), \bar{x}_\infty) \to 0$, say $x(t) \to \bar{x}_\infty$. □

6. Special domains of attraction. The proofs of the preceding sections have allowed us to show that every domain of attraction is indeed a tracking domain. The actual implementation of the proposed control law requires the computation of the functions $\Psi(x, \bar{x})$ and $\psi_{\mathcal{R}_\epsilon}(r)$. In the next subsections we will see how the above arguments apply when we consider different families of domains.

6.1. Ellipsoidal domains. Ellipsoids are very popular candidate domains of attraction. They have been deeply investigated in literature, see for instance [18], [21]. A detailed exposition on the properties of such sets can be found in [10]. Assume that an ellipsoidal domain of attraction

$$\mathcal{P} = \{x : x^T Q x \leq 1\}, \quad Q > 0,$$

is available and this domain is associated with the linear control law $\phi(x) = Kx$.

Since by Assumption 2.1 M_c (resp., M_d) is invertible we denote with K_{xr} and K_{ur}, resp., the first n lines and the last m lines of M_c^{-1} (resp., M_d) so that $\bar{x}(r) = K_{xr}r$ and $\bar{u} = K_{ur}r$. The set $\mathcal{R}$ is the intersection of the sets

$$\mathcal{R}_x = \{r : r^T K_{xr}^T Q K_{xr} r \leq 1\}$$

and

$$\mathcal{R}_u = \{r : K_{ur} r \in \mathcal{U}\}.$$

The Minkowski functional of $\mathcal{R}_\epsilon = (1-\epsilon)\mathcal{R}$ is easily computable. For instance, if $\mathcal{U}$ is the unit ball of some norm $\|\cdot\|_{\mathcal{U}}$, then $\mathcal{R}_u = \{r : \ \|K_{ur} r\|_{\mathcal{U}} \le 1\}$ so that

$$\psi_{\mathcal{R}_\epsilon} = (1-\epsilon)^{-1} \max\{(r^T K_{xr}^T Q K_{xr} r)^{1/2}, \|K_{ur} r\|_{\mathcal{U}}\}.$$

With the choice of an ellipsoidal tracking domain, denoting by $y = x - \bar{x}$, the expression of $\Psi(x, \bar{x})$ reduces to the following:

$$\Psi(x, \bar{x}) = \frac{\bar{x}^T Q y + [(\bar{x}^T Q y)^2 + y^T Q y (1 - \bar{x}^T Q \bar{x})]^{\frac{1}{2}}}{1 - \bar{x}^T Q \bar{x}};$$

thus its computation essentially involves the computation of the three quantities $x^T Q \bar{x}$, $x^T Q y$ and $y^T Q y$.

The operations involved on-line in the implementation of the proposed control law are hence the following:

1. Compute the "constrained" reference value $r_c = \Gamma(r)$;
2. Compute $\bar{x} = K_{xr} r_c$ and $\bar{u} = K_{ur} r_c$;
3. Compute $\Psi(x, \bar{x})$;
4. Set

$$\Phi(x, r) = Kx + (1 - \Psi(x, \bar{x}))(\bar{u} - K\bar{x}).$$

Although we have explicitly considered the case in which the original controller is linear, we can easily consider the case in which such a controller is nonlinear. The price we pay is a less neat expression of the controller.

6.2. Polyhedral domains. Polyhedral invariant sets have been deeply investigated for constrained stabilization problems [3], [11], [12], [17], [19], [20], [26], [28]. The basic reason is that if linear state and control constraints are considered, the largest domain of attraction can be arbitrarily closely approximated by a polyhedral set which is a domain of attraction. Such a result holds for both discrete-time [17], [19] and continuous-time systems [7]. It is also known that once such a domain of attraction is available, a feedback stabilizing control law can be inferred. Therefore, in the constrained stabilization problem, considering polyhedral sets as candidate domains of attraction is not restrictive.

We consider now the implementation of a tracking control strategy for a polyhedral domain of attraction. A polyhedral C-set $\mathcal{P}$ can be represented as

$$\mathcal{P} = \{x : \ \ F_i x \le 1, \ \ i = 1, 2, \ldots, s\} = \{x : \ \ Fx \le \bar{1}\},$$

where F is the matrix whose rows are F_i, while $\bar{1} \doteq [1 \ 1 \ \ldots \ 1]^T$. The Minkowski functional associated with $\mathcal{P}$ is given by [5], [28]

$$\psi(x) = \max_i \ F_i x.$$

From the above expression it is evident that in this particular case the set is the intersection of a finite number of half-spaces, thus if we consider the function $\Psi(x, \bar{x})$, its expression is given by

$$\Psi(x, \bar{x}) \doteq \max_i \frac{F_i(x - \bar{x})}{1 - F_i \bar{x}}, \quad x \in (1-\epsilon)\mathcal{P},$$

where $\epsilon > 0$ is a small number. Note that $1 - F_i\bar{x} \geq \epsilon$, so singularities are avoided.

The construction of a polyhedral domain of attraction is a well-established topic and allows the designer to derive a Lipschitz piecewise linear control law of the form $u = K(x)x$ according to [4], [17], which is positively homogeneous of order 1. Once such a domain is known, the expression of the control can be derived by means of (5.1) with the provided expression of Ψ.

After the determination of the value of Ψ, as for the other families considered here, the first step necessary for the implementation of the tracking control (3.7)–(3.9) concerns the evaluation of the function

$$\psi_{\mathcal{R}_\epsilon}(r) = (1-\epsilon)^{-1} \max\{\psi(K_{xr}r), \|K_{ur}r\|_{\mathcal{U}}\}$$

(we remind the reader that $\bar{x} = K_{xr}r$ and $\bar{u} = K_{ur}r$), whose computation is needed for the determination of the "constrained" reference value $r_c = \Gamma(r)$.

6.3. Smoothed domains. In this section we consider the family of domains of attraction whose Minkowski functional is given by

$$\psi(x) = \left(\sum_{i=1}^{s} \sigma_{2p}(F_i x) \right)^{\frac{1}{2p}}$$

where

$$\sigma_r(x) = \begin{cases} x^r \text{ if } x \geq 0, \\ 0 \text{ otherwise.} \end{cases}$$

This family has been analyzed in [9] for the constrained stabilization problem. Similar functions have been used for tracking constant signals for continuous-time systems in the presence of state constraints only in [8]. It is helpful to recall that, although the functions used in [8] are similar to those proposed here, they are essentially different. Their definition is based on an analytic expression and such definition does not apply to the Minkowski function of a generic C-set. In particular the unitary level surface is not invariant when the reference is allowed to vary. They are unsuitable for the cases in which there are control constraints and variable reference signals.

Here we reconsider this class for tracking purposes for continuous-time systems and we assume that the input constraints are of the form $\mathcal{U} = \{u \ : \|u\|_\infty \leq 1\}$. The reason for considering this class of functions comes from the fact that the differentiability of this class of functions together with the assumption on $\mathcal{U}$ to be a box allows for the determination of an explicit control law (as opposed to the control law which can be associated with a polyhedral domain which is general implicit). Indeed it can be shown that to every such domain of attraction $\mathcal{P}$ can be associated the Lipschitz control function

$$\phi(x) = -sat(k(B^T F^T \Gamma(x)^T))\psi(x)$$

where

$$\Gamma(x) \doteq \psi(x)^{(1-2p)} \begin{bmatrix} (F_1 x)^{2p-1} & \dots & (F_s x)^{2p-1} \end{bmatrix}$$

and k is a computable positive constant (see [9]). The drawback of the capability of determining an explicit control law is given anyway by the fact that it is not possible to give an explicit expression for $\Psi(x, \bar{x})$, but this has to be computed by bisection on the parameter α in (3.2).

Again, the control action requires the computation on-line of $\Psi(x(t), \bar{x}(t))$ and the application of (5.1).

7. Nonsquare systems. So far we have considered the case in which the system has as many inputs as outputs. It is not difficult to extend the results to nonsquare systems. Consider the nonsquare matrix M_c and the equation

$$\begin{bmatrix} A & B \\ C & D \end{bmatrix} \begin{bmatrix} \bar{x}(r) \\ \bar{u}(r) \end{bmatrix} = \begin{bmatrix} 0 \\ \bar{r} \end{bmatrix}.$$

It is well known that, if the above equation does not admit an unique solution, two cases are possible: for a given $\bar{r}$ there may be infinitely many solutions, or there may be none. In the first case one has to choose one vector $[\bar{x}(r)^T \ \bar{u}(r)^T]$ among the solutions. One obvious choice is to consider the minimum (possibly weighted) norm solution given by

$$\begin{bmatrix} \bar{x}(r) \\ \bar{u}(r) \end{bmatrix} = \begin{bmatrix} A & B \\ C & D \end{bmatrix}^{\dagger} \begin{bmatrix} 0 \\ \bar{r} \end{bmatrix}, \tag{7.1}$$

where $M_c^{\dagger}$ is the Moore–Penrose inverse of M_c.

In the case in which there are no solutions, the obvious extension, the least square solution, is not suitable since for tracking purpose we need the steady-state condition

$$\begin{bmatrix} A & B \end{bmatrix} \begin{bmatrix} \bar{x}(r) \\ \bar{u}(r) \end{bmatrix} = 0 \tag{7.2}$$

to be satisfied exactly, a condition which is in general not satisfied by the solution (7.1). In this case we may take, among all the equilibrium points satisfying (7.2), those which minimize $\|\bar{y} - \bar{r}\|$. This is obtained by first parameterizing all the elements of the kernel of $[A\ B]$ as $x = Nz$, $u = Mz$, where $AN + BM = 0$ and then choosing for a given reference r the least square solution of

$$\min_z \ \|[CN + DM]z - r\|.$$

Since the minimum is obtained for $z = [CN + DM]^{\dagger} r$ we finally obtain

$$\begin{bmatrix} \bar{x}(r) \\ \bar{u}(r) \end{bmatrix} = \begin{bmatrix} N[CN + DM]^{\dagger} r \\ M[CN + DM]^{\dagger} r \end{bmatrix}.$$

Once the linear relation above is established, all the theory developed in the previous chapters remains valid. Clearly in the over-determined (i.e., no exact solution) case, asymptotic tracking, even for a constant signal, is not possible in general.

8. Examples.

8.1. Continuous-time example: linear state feedback. As a first example we consider the following continuous-time system:

$$\dot{x}(t) = \begin{bmatrix} 1 & .4 \\ .8 & .5 \end{bmatrix} x(t) + \begin{bmatrix} 0 \\ 1 \end{bmatrix} u(t)$$
$$y(t) = \begin{bmatrix} .2 & .1 \end{bmatrix} x(t).$$

We derived for this system a linear state feedback compensator $u = Kx$ which is such to stabilize the closed loop system while guaranteeing that for every initial state $\|x(0)\| \leq 1$ we have $\|Kx\| \leq U_{max}$ along the system trajectory. By solving the

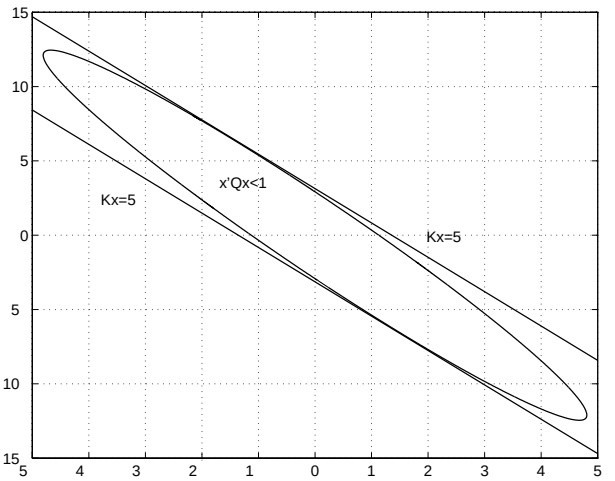


FIG. 3. *Domain of attraction.*

corresponding set of linear matrix inequalities (LMI) [10] for $U_{max} = 5$, we obtained the gain

$$K = \begin{bmatrix} -3.6845 & -1.5943 \end{bmatrix}$$

which corresponds to the ellipsoidal domain of attraction (with speed of convergence $\beta = .0001$) $\mathcal{P} = \{x \ : x^T Q x \leq 1\}$, where

$$Q = \begin{bmatrix} .7859 & .2950 \\ .2950 & .1172 \end{bmatrix}.$$

The set $\mathcal{P}$ is the ellipse depicted in Figure 3 and is by construction contained in the region where $|Kx| \leq U_{max}$ (also reported in Figure 3); thus the constraints on the reference value $\bar{r}$ can be derived by the set $\mathcal{P}$ alone and result in $|\bar{r}| \leq .2403$. We slightly reduced this value to .24 to guarantee that $\bar{x}(\bar{r})$ belongs to the interior of $\mathcal{P}$. This means that $\bar{r} = \Gamma(r) = 0.24\ sat(r/0.24)$.

We report in Figure 4 the zero initial state time evolution of the output corresponding to the reference signal (dashed line)

$$r(t) = \begin{cases} 0.40 & \text{for } \ 0 \leq t \leq 100, \\ 0.20 + 0.20e^{-(t-100)/10} & \text{for } \ 100 < t \end{cases}$$

and we report in Figure 5 the state space evolution.

8.2. Discrete-time example: linear variable structure state feedback. As a second example we consider the discrete-time system

$$x(k+1) = \begin{bmatrix} 1 & .3 \\ -1 & 1 \end{bmatrix} x(k) + \begin{bmatrix} .5 \\ 1 \end{bmatrix} u(k)$$
$$y(k) = \begin{bmatrix} -.1 & .3 \end{bmatrix} x(k)$$

with the state and control constraint sets, resp., given by

$$\mathcal{X} = \{x \ : \|x\|_\infty \leq 1\}$$

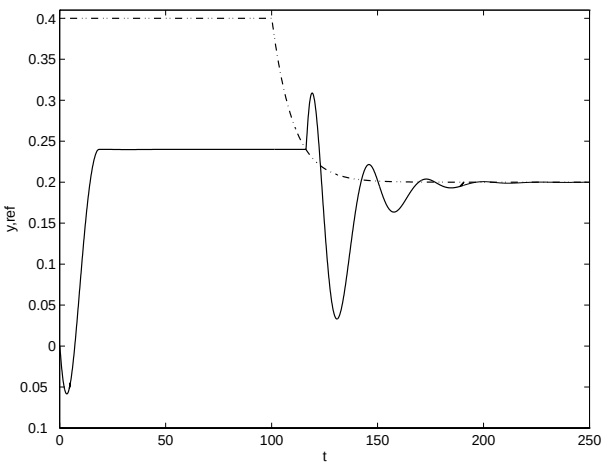


FIG. 4. *Output and reference (dashed) time evolution.*

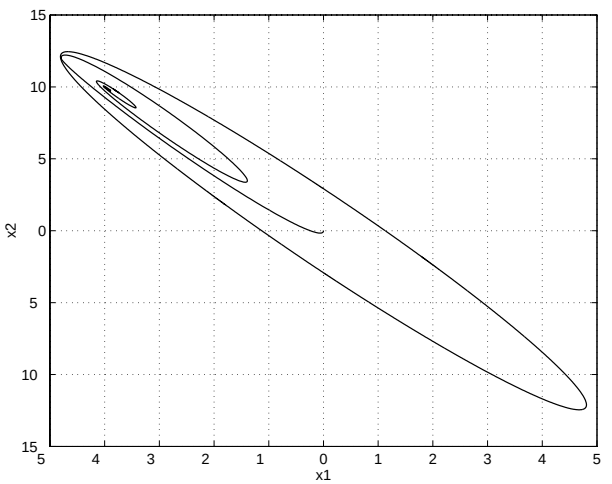


FIG. 5. *Domain of attraction and state space evolution.*

and

$$\mathcal{U} = \{u \ : |u| \leq 1\}.$$

For this system we computed a symmetric polyhedral domain of attraction $\mathcal{P} = \{x \ : \ \|Fx\|_\infty \leq 1\}$ with speed of convergence $\lambda = .9$ (which is the polyhedral region in Figure 6), where

$$F = \begin{bmatrix} 0 & 1 \\ 1.8160 & -0.2421 \\ 1.3140 & 0.1932 \end{bmatrix}.$$

In this case the constraints on the reference value derive from the constraint that $\bar{x} \in \mathcal{P}$ and translate in $|\bar{r}| \leq .27$. The linear variable structure controller associated with $\mathcal{P}$ is given by $u(x) = k_i x$ with $i = \arg\max_j |f_j x|$, where f_i and k_i are the ith

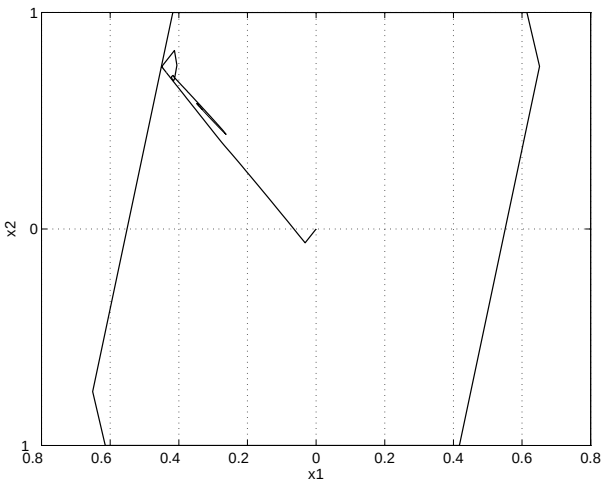


FIG. 6. *Polyhedral domain of attraction and state space evolution.*

rows of F and K, resp., where

$$K = \begin{bmatrix} -0.4690 & -0.7112 \\ -0.6355 & -0.7806 \\ -1.3140 & -0.1931 \end{bmatrix}.$$

We applied the control law proposed in section 6.2 with $\epsilon = .01$, starting from zero initial state, for the reference signal

$$r(t) = 0.2 + 0.4 * \sin(0.01 * t) * e^{-0.005 * t}.$$

Figure 7 depicts the time evolution of the output and the reference value, and Figure 6 shows the corresponding state space trajectory.

9. Conclusions. In this paper it was shown that every set of initial states which can be brought to the origin while assuring that no control and state constraints are violated is indeed a tracking domain. This amounts to saying that it is possible to track a given reference value only if its corresponding state and input equilibrium pair belongs to a domain of attraction, whereas if this condition is not met the state is not "trackable." We have provided this result based on the system structure, without assuming the existence of a stabilizing compensator, by furnishing a Lyapunov function which is derived by the domain of attraction and which allows us to derive a tracking strategy that avoids state and input constraints yet guarantees an a priori speed of convergence. Future directions in this area concern the possibility of deriving stabilizing tracking strategies capable of maintaining the tracking error within prespecified bounds and of improving the speed of convergence.

The presented approach can be applied as well to systems with control constraints only. In this case the domain of attraction can be computed by considering a fictitious (sufficiently large) state constraints region and by applying the techniques suggested in [9], [17], [20]. We finally remind the reader that for systems with no state constraints and free from eigenvalues in the open right half plane (or with no eigenvalues outside the closed unit disk) a completely different approach can be used for global

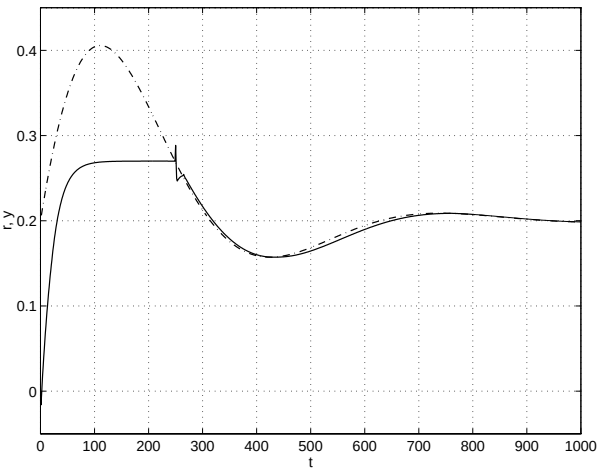


FIG. 7. *Output and reference time-evolution.*

stabilization [25] or semiglobal stabilization [24], [27]. In this case those approaches can be successfully applied to the solution of the tracking problem with an arbitrarily large initial condition set.

Appendix. Given a differentiable positively homogeneous function of order 1 $F : \mathbb{R}^n \to \mathbb{R}$ it is well known that $\langle \nabla F(y), y \rangle = F(y)$. This property can be extended easily to our case of nondifferentiable functions under the convexity assumption.

Proof of the equality (4.8). Let $F : \mathbb{R}^n \mapsto \mathbb{R}$ be a convex and positively homogeneous function of order 1. Then for any y and any $w \in \nabla F(y)$, we have $\langle w, y \rangle = F(y)$.

For $y = 0$ the property is obvious. So let $y \neq 0$. By definition any $w \in \nabla F(y)$ is such that

$$F(z) - F(y) \geq \langle w, z - y \rangle$$

for all $z \in \mathbb{R}^n$, that is

$$F(z) - \langle w, z \rangle \geq F(y) - \langle w, y \rangle.$$

Now let $z = \lambda y$ with $\lambda > 0$. Then

$$\lambda \left[F(y) - \langle w, y \rangle \right] \geq F(y) - \langle w, y \rangle \quad \text{for all} \quad \lambda > 0,$$

which necessarily implies

$$F(y) - \langle w, y \rangle = 0. \qquad \square$$

Acknowledgments. The authors are grateful to Professor Zanolin for helpful discussions on the paper.

REFERENCES

[1] J. P. Aubin and A. Cellina, *Differential Inclusions*, Springer-Verlag, Berlin, 1984.
[2] A. Bemporad, A. Casavola, and E. Mosca, *Nonlinear control of constrained linear systems with predictive reference management*, IEEE Trans. Automat. Control, 42 (1997), pp. 340–349.
[3] A. Benzaouia and C. Burgat, *The regulator problem for a class of linear systems with constrained control*, Systems Control Lett., 10 (1988), pp. 357–363.
[4] F. Blanchini, *Ultimate boundedness control for discrete-time uncertain system via set-induced Lyapunov functions*, IEEE Trans. Automat. Control, 39 (1994), pp. 428–433.
[5] F. Blanchini, *Non-quadratic Lyapunov functions for robust control*, Automatica, 31 (1995), pp. 451–461.
[6] F. Blanchini, *Set invariance in control—a survey*, Automatica, 35 (1999), pp. 1747–1767.
[7] F. Blanchini and S. Miani, *Constrained stabilization for continuous-time systems*, Systems Control Lett., 28 (1996), pp. 95–102.
[8] F. Blanchini and S. Miani, *Set based constant reference tracking for continuous-time constrained systems*, in Advances of Stability Theory at the End of the XXth Century, A.A. Martynyuk, ed., Stability Control Theory Methods Appl. 13, Gordon and Breach, London, UK, 2000, pp. 267–278.
[9] F. Blanchini and S. Miani, *Constrained stabilization via smooth Lyapunov functions*, Systems Control Lett., 35 (1998), pp. 155–163.
[10] S. Boyd, L. El Ghaoui, E. Feron, and V. Balakrishnan, *Linear Matrix Inequalities in System and Control Theory*, SIAM Studies in Applied Mathematics 15, SIAM, Philadelphia, PA, 1994.
[11] C. Burgat and S. Tarbouriech, *Positively invariant sets for constrained continuous-time systems with cone properties*, IEEE Trans. Automat. Control, 39 (1994), pp. 401–405.
[12] E. B. Castelan and J. C. Hennet, *Eigenstructure assignment for state constrained linear continuous time systems*, Automatica, 28 (1992), pp. 605–611.
[13] F. Clarke, *Optimization and Nonsmooth Analysis*, John Wiley & Sons, New York, 1983.
[14] E. G. Gilbert and K. T. Tan, *Linear systems with state and control constraints: The theory and the applications of the maximal output admissible sets*, IEEE Trans. Automat. Control, 36 (1991), pp. 1008-1020.
[15] E. G. Gilbert, I. Kolmanowsky, and K. T. Tan, *Discrete–time reference governors and the nonlinear control of systems with state and control constraints*, Internat. J. Robust Nonlinear Control, 5 (1995), pp. 487-504.
[16] T. J. Graettinger and B. H. Krogh, *On the computation of reference signal constraints for guaranteed tracking performance*, Automatica, 28 (1992), pp. 1125–1141.
[17] P. O. Gutman and M. Cwikel, *Admissible sets and feedback control for discrete-time linear systems with bounded control and states*, IEEE Trans. Automat. Control, 31 (1986), pp. 373–376.
[18] P. O. Gutman and P. Hagander, *A new design of constrained controllers for linear systems*, IEEE Trans. Automat. Control, 30 (1985), pp. 22–33.
[19] S. S. Keerthy and E. G. Gilbert, *Computation of minimum-time feedback control laws for discrete-time systems with state and control constraints*, IEEE Trans. Automat. Control, 32 (1987), pp. 432–435.
[20] J. B. Lasserre, *Reachable, controllable sets and stabilizing control of constrained systems*, Automatica, 29 (1993), pp. 531–536.
[21] Z. Lin, A. A. Stoorvogel, and A. Saberi, *Output regulation for linear systems subject to input saturation*, Automatica, 32 (1996), pp. 29–47.
[22] D. G. Luenberger, *Optimization by Vector Space Methods*, John Wiley & Sons, New York, 1969.
[23] R. T. Rockafellar, *Convex Analysis*, Princeton University Press, Princeton, NJ, 1972.
[24] A. Saberi, Z. Lin, and A. R. Teel, *Control of linear systems with saturating actuators*, IEEE Trans. Automat. Control, 41 (1996), pp. 368–378.
[25] H. J. Sussmann, E. D. Sontag, and Y. D. Yang, *A general result on the stabilization of linear systems using bounded controls*, IEEE Trans. Automat. Control, 39 (1994), pp. 2411–2425.

[26] M. Sznaier and M. J. Damborg, *Heuristically enhanced feedback control of constrained discrete-time systems*, Automatica, 26 (1990), pp. 521–532.
[27] A. R. Teel, *Global stabilization and restricted tracking for multiple integrators with bounded controls*, Systems Control Lett., 18 (1992), pp. 165–171.
[28] M. Vassilaki and G. Bitsoris, *Constrained regulation of linear continuous-time dynamical systems*, Systems Control Lett., 47 (1989), pp. 247-252.

SIAM J. CONTROL OPTIM.
Vol. 38, No. 4, pp. 995–1032

SELECTION OF BEST ORTHONORMAL RATIONAL BASIS*

PER BODIN†, LARS F. VILLEMOES‡, AND BO WAHLBERG§

Abstract. This contribution deals with the problem of structure determination for generalized orthonormal basis models used in system identification. The model structure is parameterized by a prespecified set of poles representing a finite-dimensional subspace of $\mathcal{H}^2$. Given this structure and experimental data, a model can be estimated using linear regression techniques. Since the variance of the estimated model increases with the number of estimated parameters, one objective is to find coordinates, or a basis, for the finite-dimensional subspace giving as compact or parsimonious a system representation as possible. In this paper, a best basis algorithm and a coefficient decomposition scheme are derived for the generalized orthonormal rational bases. Combined with linear regression and thresholding this leads to compact transfer function representations. The methods are demonstrated with several examples.

Key words. system identification, model structures, orthonormal basis functions, best basis

AMS subject classifications. 39A30, 41A20, 42C20, 05A05

PII. S036301299732818X

1. Introduction.

Over the last few years, the use of orthonormal rational basis functions in system identification has received considerable attention. The idea is to approximate the transfer function of a linear time-invariant system with a linear combination of orthonormal basis functions having fixed poles. Given a finite collection of asymptotically stable filters, the basis functions can be constructed from a Gram–Schmidt orthogonalization of the given filters using all-pass filters with balanced state space realizations. Orthogonality in $\mathcal{H}^2$ is defined with respect to the inner product (2.2). The resulting family of orthonormal functions then forms a basis for the finite-dimensional subspace of $\mathcal{H}^2$ consisting of all possible transfer functions having the poles given at hand. Depending on the order in which the orthogonalization is performed, different bases are obtained. These represent different selections of coordinate systems for the finite-dimensional subspace.

Given input/output data from a linear time invariant system, the system transfer function can be approximated in the given finite-dimensional subspace by estimating the expansion coefficients associated with the selected coordinate system. Supposing that the output data is corrupted by noise, the estimated coefficients also can be considered noisy. In particular, small coefficients will be more or less hidden in noise and thus the system estimate actually benefits from discarding them. In this way, selection of the coordinate system becomes an important issue. Since the bases are orthogonal, the energy of the coefficients will always be exactly the same independent of the selection of coordinates. If all coefficients below some magnitude are discarded, a coordinate system concentrating the energy into a few large and many small coef-

*Received by the editors October 6, 1997; accepted for publication (in revised form) June 15, 1999; published electronically April 4, 2000. This research was supported by the Swedish Research Council for Engineering Sciences (TFR).

http://www.siam.org/journals/sicon/38-4/32818.html

†Swedish Space Corporation, Space Vehicle Design, Science Systems Division, P.O. Box 4207, S-171 04 Solna, Sweden (per.bodin@ssc.se).

‡Department of Mathematics, The Royal Institute of Technology, S-100 44 Stockholm, Sweden (larsv@math.kth.se).

§S3-Automatic Control, The Royal Institute of Technology, S-100 44 Stockholm, Sweden (bo@s3.kth.se).

ficients results in a smaller error than a selection having a more evenly distributed energy. This can be viewed as selecting a coordinate system where some of the basis functions are close to, or resemble, the system being identified. Coefficient concentration can, for example, be measured by the entropy or the ℓ^1-norm of the coefficients. The following example demonstrates the consequence of orthogonalization order. Let a transfer function be given by

$$G(z) = \frac{3z + 1/\sqrt{2}}{5z(z - 1/\sqrt{2})}. \tag{1.1}$$

Using a Gram–Schmidt orthogonalization of the two functions $\frac{1}{z}$ and $\frac{1}{z-1/\sqrt{2}}$, the orthonormal pair of functions

$$\psi_1(z) = \frac{1}{z} \quad \text{and} \quad \psi_2(z) = \frac{1/\sqrt{2}}{z(z - 1/\sqrt{2})} \tag{1.2}$$

is constructed. The transfer function $G(z)$ can then be written $G(z) = g_1\psi_1(z) + g_2\psi_2(z)$, where

$$g_1 = \frac{3}{5} \quad \text{and} \quad g_2 = \frac{4}{5}. \tag{1.3}$$

On the other hand, if the orthogonalization is performed the other way around, the orthonormal pair

$$\tilde{\psi}_1(z) = \frac{1/\sqrt{2}}{z - 1/\sqrt{2}} \quad \text{and} \quad \tilde{\psi}_2(z) = \frac{1 - z/\sqrt{2}}{z(z - 1/\sqrt{2})} \tag{1.4}$$

is obtained so that $G(z)$ can be written $G(z) = \tilde{g}_1\tilde{\psi}_1(z) + \tilde{g}_2\tilde{\psi}_2(z)$, with

$$\tilde{g}_1 = \sqrt{\frac{49}{50}} \quad \text{and} \quad \tilde{g}_2 = \frac{1}{\sqrt{50}}. \tag{1.5}$$

In the first case, the two expansion coefficients g_1 and g_2 have almost equal size, while in the second case $\tilde{g}_1$ is much larger than $\tilde{g}_2$. In this way, almost all of the "energy" is concentrated in the coefficient $\tilde{g}_1$. The ℓ^1-norms in these two cases satisfy

$$|g_1| + |g_2| = \frac{7}{5} > \sqrt{2}\frac{4}{5} = |\tilde{g}_1| + |\tilde{g}_2|. \tag{1.6}$$

In general, under the constraint that $g_1^2 + g_2^2$ is constant, $|g_1| + |g_2|$ is minimized when one of the coefficients is zero so that the other term becomes the system $G(z)$ itself. The ℓ^1-norm is maximized when the two coefficients are equal so that the energy is maximally spread out. The example illustrates the importance of selecting coordinates. The coefficient energy becomes concentrated into few coefficients if the coordinates are selected as close as possible to the expanded function.

The contribution of this paper deals with the selection of the best coordinates, or basis. The orthogonalization procedure results in many different possible bases. However, a specific basis function can belong to many different bases. In fact, the overlap can be tremendous. If for example the dimension of the subspace is n, and if all the initially given filters have different poles, there will be $n!$ different orthogonal bases, each one resulting from a different orthogonalization order. However, all of these different bases are constructed from only $n2^{n-1}$ different basis functions. For example,

$n = 10$ gives 3,628,800 bases constructed from only 5,120 basis functions. This observation leads to a scheme for selecting the very best coordinates, where "best" is in terms of concentration of the expansion coefficients, measured by entropy, ℓ^1-norm, or some other cost. Moreover, if expansion coefficients are estimated in one single basis, the coefficients belonging to all the other basis functions can be calculated using local orthogonal transformations. Thus, a total framework is developed for decomposition into all possible expansion coefficients and then selection of the best basis. All of this is performed in the same complexity as the number of basis functions. In this meaning, the scheme is "fast." Several examples of applications of the best-basis selection scheme are given in section 5.

Construction of orthonormal rational bases is by no means new. In [33], it was investigated how the Laguerre basis could be used for system identification. This basis can be generated from identical first order all-pass filters, and the results were generalized to the second order Kautz functions in [34]. Further generalizations were made in [16, 15, 14] and [21]. In [5], it is demonstrated how general orthonormal rational bases can be generated in a simple manner from all-pass filters with balanced state space realizations. This result is mainly due to [20, 27], but similar basis constructions appear in the literature as early as 1925 in [31]. The historical background of these constructions is investigated in [2].

In order to have an efficient representation of a system using an orthonormal rational expansion, it is important that the basis functions be calibrated to the system being identified. This corresponds to selecting an approximating subspace of $\mathcal{H}^2$ and determines the rate of convergence of the expansion coefficients; see [36]. Calibration can be made by incorporating a priori knowledge about the poles of the system in the construction of the basis. As already mentioned, the mean squared error of the estimate is further reduced if coefficients with magnitudes less than the corresponding variance error are set to zero. This way of thresholding was investigated in [5], where ideas from [9] were applied to rational bases. Model sets based on orthonormal functions have nice properties in terms of least squares estimates of the coefficients since if many coefficients are discarded in the final estimate, the least squares estimate of the kept coefficients remains the same; see [13].

In the best basis selection scheme developed in this paper, the observation that the number of basis functions is much smaller than the number of different bases is crucial for the existence of an efficient best basis algorithm. This situation is similar to the wavelet packet best-basis selection methods suggested in [7], and the same best basis criteria can be used in the case of the best rational basis as well. Another best basis criterion was suggested in [10]. In this paper the cost function is related to the entropy cost of [8].

More precisely, the parallel to the wavelet packet transform can briefly be described in the following way. In the wavelet packet case, a signal is made finite dimensional through sampling. The finite-dimensional space to work with then becomes $\mathbb{R}^n$ or $\mathbb{C}^n$, where n is the number of samples. The basis in which the signal is described is the canonical Dirac basis. The signal is then transformed into all the wavelet packet coefficients using a low-complexity implementation of the transform. The best wavelet packet basis can then be selected. This is illustrated in Figure 1.1. The number of different bases is approximately 1.5^n, while the number of basis functions is only $n(\log_2(n) + 1)$.

In the case of choosing the best rational basis, a collection of filters is given. The finite-dimensional space is then identified with the span of the poles of these filters

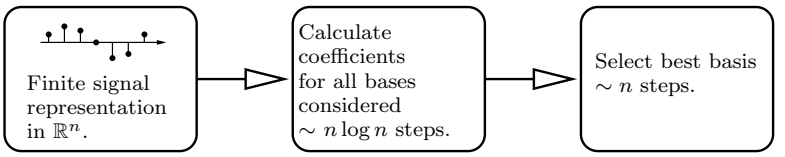


FIG. 1.1. *Wavelet packet best basis decomposition.*

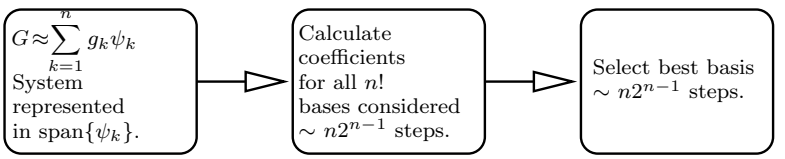


FIG. 1.2. *Best rational basis decomposition.*

and the system is approximated in this space by estimating the coefficients for a basis given by some ordering of the filters. This basis can then be called the canonical basis. All the coefficients needed for all the other orderings of the filters can then be calculated from the canonical coefficients. As mentioned above, if all n filters are different, the number of bases is $n!$ while the number of different coefficients and steps for calculating them is $n2^{n-1}$. As in the case of wavelet packet best basis selection, the best rational basis can now be selected. This can be done in $n2^{n-1}$ steps as well. Figure 1.2 illustrates the scheme.

Connections between wavelets and orthonormal rational basis functions have previously been studied in [11] and frames of rational wavelets for system identification were examined in [25, 24] with promising results, but where one conclusion is that bases of rational wavelets can never be orthogonal.

The paper is organized as follows. In section 2, the background for rational basis construction and system identification is given. Section 3 presents the best basis selection scheme while the procedure for calculating all possible expansion coefficients is derived in section 4. Different examples are given in section 5 and finally in section 6, some conclusions are made. The two main results of the paper are the best basis selection scheme presented in section 3.3 and the coefficient generation of section 4.2.

2. Background. This section presents the background for construction of orthonormal rational basis functions and for their use in system identification. Section 2.1 gives a convenient way to construct rational bases from a cascade of all-pass filters with balanced state space realizations. In section 2.2, it is shown how thresholding can be applied when using these bases in system identification.

2.1. An orthonormal family in $\mathcal{H}^2$. Assume that a strictly proper asymptotically stable linear time-invariant dynamical system is described by

$$y(t) = G(q)u(t), \tag{2.1}$$

where $\{u(t)\}$ is the input sequence, $\{y(t)\}$ is the output sequence, and q is the forward shift operator. The transfer function $G(z)$ is assumed to be asymptotically stable so that the system will belong to $\mathcal{H}^2$, being the Hardy space of functions analytical outside, and square integrable on the unit circle. See, e.g., [1, 28]. Since $G(z)$ is

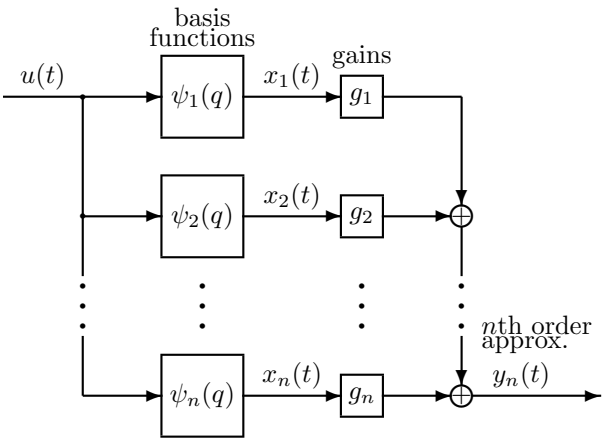


FIG. 2.1. *The approximate system* $G_n(z)$.

assumed to be strictly proper, only the part of $\mathcal{H}^2$ with functions fulfilling $G(\infty) = 0$ is considered.

Let $\{\psi_k(z)\}_{k=1}^{\infty}$ be an orthonormal family of functions in $\mathcal{H}^2$ such that $\langle \psi_k, \psi_l \rangle = \delta_{k,l}$, where the inner product is defined as

$$\langle f, g \rangle = \frac{1}{2\pi} \int_{-\pi}^{\pi} f(e^{i\omega}) \overline{g(e^{i\omega})} d\omega \tag{2.2}$$

for any pair of functions $f, g \in \mathcal{H}^2$, and where $\delta_{k,l} = 1$ if $k = l$ and 0 otherwise. Suppose now that $G(z)$ can be represented as

$$G(z) = \sum_{k=1}^{\infty} g_k \psi_k(z). \tag{2.3}$$

The simplest example of such a family of orthonormal functions is the standard basis

$$\psi_k(z) = z^{-k}, \qquad k = 1, 2, \ldots . \tag{2.4}$$

Since the transfer function $G(z)$ belongs to $\mathcal{H}^2$, the expansion coefficient sequence $\{g_k\}$ will be square summable so that $g_k \to 0$ as $k \to \infty$. Therefore, the sequence can be truncated in order to obtain an approximation $G_n(z)$ of $G(z)$, where

$$G_n(z) = \sum_{k=1}^{n} g_k \psi_k(z). \tag{2.5}$$

Due to orthonormality, the expansion coefficients for both (2.3) and (2.5) will be given by the least squares generalized Fourier coefficients $g_k = \langle G, \psi_k \rangle$. Assume now that the functions $\psi_k(z)$ are state transfer functions of the approximating system $G_n(z)$. This is illustrated in Figure 2.1. A state space realization of $G_n(z)$ is written

$$\begin{aligned} \mathbf{x}(t+1) &= \mathbf{A}_n \mathbf{x}(t) + \mathbf{B}_n u(t), \\ y(t) &= \mathbf{C}_n \mathbf{x}(t), \qquad \mathbf{C}_n = (g_1 \ldots g_n), \end{aligned} \tag{2.6}$$

for which the functions $\{\psi_k\}_{k=1}^n$ correspond to

$$\psi_k(z) = \mathbf{e}_k^T \left(z\mathbf{I} - \mathbf{A}_n\right)^{-1} \mathbf{B}, \qquad k = 1, \ldots, n, \tag{2.7}$$

where $\mathbf{e}_k = (0 \ldots 1 \ldots 0)^T$ is a unit vector with 1 in position k and where $\mathbf{I}$ is the identity matrix. Assume that the system matrix $\mathbf{A}_n$ has all its eigenvalues strictly inside the unit circle so that the system (2.6) is asymptotically stable. Then, the family of functions (2.7) has mutual inner products $\langle \psi_k, \psi_l \rangle = \mathbf{e}_k^T \mathbf{P} \mathbf{e}_l$, where $\mathbf{P}$ is the unique solution of the discrete Lyapunov equation

$$\mathbf{A}_n \mathbf{P} \mathbf{A}_n^T + \mathbf{B}_n \mathbf{B}_n^T = \mathbf{P}. \tag{2.8}$$

This follows from the fact that the inner products can be written

$$\begin{aligned} \langle \psi_k, \psi_l \rangle &= \frac{1}{2\pi} \int_{-\pi}^{\pi} \left[\mathbf{e}_k^T \left(\mathbf{I}e^{i\omega} - \mathbf{A}_n\right)^{-1} \mathbf{B}_n\right] \left[\mathbf{e}_l^T \left(\mathbf{I}e^{-i\omega} - \mathbf{A}_n\right)^{-1} \mathbf{B}_n\right] d\omega \\ &= \mathbf{e}_k^T \left\{ \frac{1}{2\pi} \int_{-\pi}^{\pi} \left[\left(\mathbf{I}e^{i\omega} - \mathbf{A}_n\right)^{-1} \mathbf{B}_n\right] \left[\left(\mathbf{I}e^{-i\omega} - \mathbf{A}_n\right)^{-1} \mathbf{B}_n\right]^T d\omega \right\} \mathbf{e}_l =: \mathbf{e}_k^T \mathbf{P} \mathbf{e}_l . \end{aligned} \tag{2.9}$$

It is then well known that the controllability Gramian $\mathbf{P}$ satisfies the Lyapunov equation (2.8) and that $\mathbf{P}$ is unique and positive definite if $\mathbf{A}_n$ is stable. It is easy to conclude that the family $\{\psi_k\}_{k=1}^n$ becomes orthonormal if and only if

$$\mathbf{A}_n \mathbf{A}_n^T + \mathbf{B}_n \mathbf{B}_n^T = \mathbf{I}. \tag{2.10}$$

In this way, the parameters available for constructing an orthonormal family are $(\mathbf{A}_n, \mathbf{B}_n)$ under the constraint (2.10), and $\mathbf{A}_n$ is asymptotically stable. However, if n is large (2.10) might be complicated to solve and if the family is to be expanded to, say, order $n+1$, it would be convenient to do this in some structured way.

A state space realization satisfying (2.10) is said to be input balanced. In [20, 27], a way of constructing orthonormal families from serial connections of all-pass filters with such realizations was given. This kind of construction has previously been presented in [5, 2] and similar ideas have also been discussed in [39, 6, 21].

Crucial for all of these constructions is the concept of orthogonal all-pass filters defined as follows.

DEFINITION 2.1 (orthogonal all-pass filter). *A scalar transfer function $H(z)$ is said to be all-pass if*

$$H(z)H(1/z) \equiv 1, \tag{2.11}$$

and it is said to be orthogonal if its state-space realization $(\mathbf{A}, \mathbf{B}, \mathbf{C}, D)$ is input balanced, i.e., if it fulfills

$$\mathbf{A}\mathbf{A}^T + \mathbf{B}\mathbf{B}^T = \mathbf{I}. \tag{2.12}$$

It is straightforward to verify [27] that a filter $H(z)$ with state space realization $(\mathbf{A}, \mathbf{B}, \mathbf{C}, D)$ is orthogonal and all-pass if and only if

$$\mathbf{F} := \begin{bmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{C} & D \end{bmatrix} \tag{2.13}$$

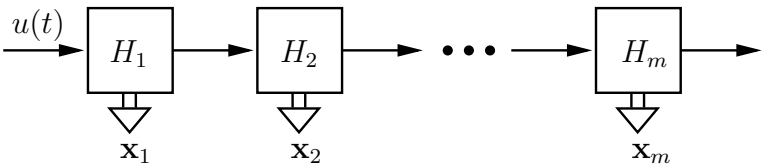


FIG. 2.2. *A cascade of all-pass filters with input balanced realizations generating an orthonormal family.*

is an orthogonal matrix, $\mathbf{F}\mathbf{F}^T = \mathbf{I}$.

The key result in [27, sect. 10.4] is that the orthogonality and all-pass properties are preserved through several different connections of such filters. One such operation is the serial, or cascade, connection. More formally, let $H_1(z)$ and $H_2(z)$ be two orthogonal all-pass filters with states $\mathbf{x}_1$ and $\mathbf{x}_2$, respectively. Then, the serial connection $H_1(z)H_2(z)$ of the two filters with the state vector $(\mathbf{x}_1^T \ \mathbf{x}_2^T)^T$ is also orthogonal and all-pass.

This follows from the fact that if $H_1(z)$ and $H_2(z)$ have the state space realizations $(\mathbf{A}_1, \mathbf{B}_1, \mathbf{C}_1, D_1)$ and $(\mathbf{A}_2, \mathbf{B}_2, \mathbf{C}_2, D_2)$ respectively, and if the matrices

$$\mathbf{F}_1 := \begin{bmatrix} \mathbf{A}_1 & \mathbf{B}_1 \\ \mathbf{C}_1 & D_1 \end{bmatrix} \quad \text{and} \quad \mathbf{F}_2 := \begin{bmatrix} \mathbf{A}_2 & \mathbf{B}_2 \\ \mathbf{C}_2 & D_2 \end{bmatrix} \tag{2.14}$$

are defined, the serial connection will have the corresponding matrix

$$\mathbf{F}_3 := \left[\begin{array}{cc|c} \mathbf{A}_1 & \mathbf{0} & \mathbf{B}_1 \\ \mathbf{B}_2\mathbf{C}_1 & \mathbf{A}_2 & \mathbf{B}_2 D_1 \\ \hline D_2\mathbf{C}_1 & \mathbf{C}_2 & D_2 D_1 \end{array}\right] = \begin{bmatrix} \mathbf{I} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{A}_2 & \mathbf{B}_2 \\ \mathbf{0} & \mathbf{C}_2 & D_2 \end{bmatrix} \begin{bmatrix} \mathbf{A}_1 & \mathbf{0} & \mathbf{B}_1 \\ \mathbf{0} & \mathbf{I} & \mathbf{0} \\ \mathbf{C}_1 & \mathbf{0} & D_1 \end{bmatrix} \tag{2.15}$$

for the state $(\mathbf{x}_1^T \ \mathbf{x}_2^T)^T$. It is then trivial to verify that if $\mathbf{F}_1$ and $\mathbf{F}_2$ are orthogonal matrices, the factors of $\mathbf{F}_3$ are also orthogonal, making their product orthogonal as well.

From this result it is easy to see that, given a family of stable all-pass filters $\{H_k\}_{k=1}^m$ with McMillan degrees $\{n_k\}_{k=1}^m$, an orthonormal family in $\mathcal{H}^2$ can be generated by serially connecting the filters in an array as shown in Figure 2.2.

The orthonormal functions $\{\psi_k\}_{k=1}^n$ are then simply given as the state transfer functions from the input to the cascade to each of the components of the states of the filters, making $n = \sum_{k=1}^m n_k$. More specifically let $\mathbf{x}_k$, defined as the states of $H_k(z)$ and shown in Figure 2.2, have components

$$\mathbf{x}_k(t) = \begin{bmatrix} x_{k,1}(t) \\ x_{k,2}(t) \\ \vdots \\ x_{k,n_k}(t) \end{bmatrix}, \tag{2.16}$$

where n_k is the McMillan degree of the filter $H_k(z)$. The transfer functions $\psi_{k,l}(z)$, defined by $x_{k,l}(t) = \psi_{k,l}(q)u(t)$, then satisfy $\langle \psi_{i,j}, \psi_{k,l} \rangle = \delta_{i,k}\delta_{j,l}$.

Define

$$\Psi_k(z) = \begin{bmatrix} \psi_{k,1}(z) \\ \psi_{k,2}(z) \\ \vdots \\ \psi_{k,n_k}(z) \end{bmatrix}, \qquad k = 1, \ldots, m, \tag{2.17}$$

and let $H_k(z)$ have the state transfer function $\Phi_k(z)$. Then the $\Psi_k(z)$'s obtained from the m ordered filters in the cascade are given by

$$\Psi_k(z) = \Phi_k(z) \prod_{l=1}^{k-1} H_l(z), \qquad k = 1, \ldots, m. \tag{2.18}$$

Some special cases should be mentioned here. If all the all-pass filters fulfill $H_k(z) = H(z)$ with McMillan degree n and have state transfer functions $\Phi_k(z) = \Phi(z)$, the construction of [15, 14] is obtained. In this case, the basis functions become

$$\Psi_k(z) = \Phi(z) \left[H(z)\right]^{k-1}, \qquad k = 1, 2, \ldots . \tag{2.19}$$

Moreover, equal first order filters result in the Laguerre basis suggested for system identification already in [17]. The Laguerre basis was also suggested for time series modeling in [35]. Identical second order filters give the Kautz basis. None of these constructions provides any flexibility in coordinate selection since they are constructed from only one kind of all-pass filter. Explicit expressions for balanced state space realizations for the Laguerre and Kautz bases can be found in, e.g., [14]. The approximation properties of both the first order filter-based Laguerre basis and the second order Kautz basis are thoroughly investigated in [36]. These results are generalized to the Hambo construction of [14] in [23].

Historical remarks on the construction of orthonormal rational bases can be found in [2]. The first such constructions were made almost simultaneously by Takenaka in [31] and by Malmquist; see [37].

The completeness condition for the family (2.18) with $k = 1, \ldots, \infty$ is very mild. Basically, it spans the subspace of all possible strictly proper transfer functions in $\mathcal{H}^2$ if the poles of $\Psi_k(z)$ do not (if at all) converge too fast to the unit circle when $k \to \infty$. This was shown in [30] in a paper that only considers the case of poles with single multiplicity. The more general result can be found in [37, chap. 10]. The condition automatically provides completeness for the Laguerre, Kautz, and Hambo constructions.

2.2. System identification and thresholding. The general orthonormal basis functions (2.18) are now possible to use for system identification. As mentioned earlier, one of the major advantages with such a model structure is that it is linear in the parameters. If the measurements are noisy, the estimated expansion coefficients will be contaminated with noise as well. One way to regularize the system estimate is then to discard small coefficients. This way of thresholding the estimate was suggested in [9] and is efficient when the basis is able to give a sparse representation of the system, but not the noise. The motivation is that if this is the case, small coefficients will mostly consist of noise so that the estimation error is reduced if these are replaced with zeros.

Let the system identification problem be defined as

$$y(t) = G(q)u(t) + e(t), \qquad t = 1, \ldots, N, \tag{2.20}$$

where $\{y(t)\}_{t=1}^N$ is the output sequence, $\{u(t)\}_{t=1}^N$ is the input sequence, and $\{e(t)\}_{t=1}^N$ is a white Gaussian measurement noise with zero mean and variance σ^2. The system is assumed to be asymptotically stable and strictly proper so that it belongs to $\mathcal{H}^2$. Define the vector valued least squares expansion coefficients $\mathbf{g}_k := \langle G, \Psi_k \rangle$, $k = 1, \ldots, m$, so that

$$\mathbf{g}_k = \begin{bmatrix} g_{k,1} \\ g_{k,2} \\ \vdots \\ g_{k,n_k} \end{bmatrix} = \begin{bmatrix} \langle G, \psi_{k,1} \rangle \\ \langle G, \psi_{k,2} \rangle \\ \vdots \\ \langle G, \psi_{k,n_k} \rangle \end{bmatrix}. \tag{2.21}$$

Suppose that the functions $\Psi_k(z)$ are generated by the cascade shown in Figure 2.2, where all-pass filter k has McMillan degree n_k. Let $n = \sum_{k=1}^m n_k$ and define the nth order approximation of $G(z)$ as

$$G_n(z) := \sum_{k=1}^m \mathbf{g}_k^T \Psi_k(z), \tag{2.22}$$

which minimizes the approximation error $\|G - G_n\|_2$. Let the parameter vector Θ be defined by

$$\Theta := \begin{bmatrix} \mathbf{g}_1 \\ \vdots \\ \mathbf{g}_m \end{bmatrix}. \tag{2.23}$$

The least squares estimate of Θ resulting from N measurements then becomes

$$\hat{\Theta} := \mathbf{R}_N^{-1} \mathbf{f}_N, \tag{2.24}$$

where

$$\mathbf{R}_N = \frac{1}{N} \sum_{t=1}^N \mathbf{X}(t) \mathbf{X}^T(t), \qquad \mathbf{f}_N = \frac{1}{N} \sum_{t=1}^N \mathbf{X}(t) y(t) \tag{2.25}$$

with

$$\mathbf{X}(t) = \begin{bmatrix} \mathbf{x}_1(t) \\ \mathbf{x}_2(t) \\ \vdots \\ \mathbf{x}_m(t) \end{bmatrix} \tag{2.26}$$

and where $\mathbf{x}_k$ is defined by (2.16). It is well known (see, e.g., [19, chap. 8]) that $\hat{\Theta} \to \Theta^*$ with probability one (w.p.1) as $N \to \infty$, where

$$\Theta^* := \arg\min_{\Theta} \frac{1}{2\pi} \int_{-\pi}^{\pi} \left| G(e^{i\omega}) - G_n(e^{i\omega}) \right|^2 \phi_u(e^{i\omega}) d\omega, \tag{2.27}$$

where ϕ_u is the power spectral density of the input sequence $\{u(t)\}_{t=1}^N$. In this way, if the input has a constant mean spectrum such as the realizations of white noise, $\hat{\Theta}$

will converge to the true least-squares gains Θ. The asymptotic distribution of the gains becomes

$$\sqrt{N}(\hat{\Theta} - \Theta^*) \sim \mathrm{AsN}(\mathbf{0}, \mathbf{R}_N^{-1}\sigma^2), \tag{2.28}$$

where σ^2 is the variance of the measurement noise. From the orthonormality of the basis functions, $\mathbf{R}_N$ will asymptotically become the identity matrix w.p.1 so that the asymptotic distribution of the coefficients $\hat{\Theta}$ will be independently and identically distributed (i.i.d.) Define the nth order approximation of $G(z)$ using the estimated least squares gains $\hat{\Theta}$ as

$$\hat{G}_n(z) := \sum_{k=1}^{m} \hat{\mathbf{g}}^T \Psi_k(z). \tag{2.29}$$

Let $\{u(t)\}_{t=1}^N$ be a realization of a white noise. The least-squares error then satisfies

$$\|G - \hat{G}_n\|_2 \to \|G - G_n\|_2 \quad \text{w.p.1} \quad \text{as } N \to \infty. \tag{2.30}$$

From the orthonormality of the basis functions, the squared error can now be written

$$\|G - \hat{G}_n\|_2^2 = \sum_{k=1}^{m} \|\mathbf{g}_k - \hat{\mathbf{g}}_k\|_2^2 + \sum_{k=m+1}^{\infty} \|\mathbf{g}_k\|_2^2 \tag{2.31}$$

so that the error is divided into two terms. The first part, $\sum_{k=1}^m \|\mathbf{g}_k - \hat{\mathbf{g}}_k\|_2^2$, is minimized given the first m all-pass filters by the least-squares estimate $\hat{\Theta}$ of the coefficients Θ, while the second term is completely determined by the choice of the first m filters. It would be desirable to make this part as small as possible for fixed m, but in general this leads to a difficult nonlinear optimization problem. However, a priori knowledge about the true system can often be used when selecting the filters $\{H_k\}_{k=1}^m$.

A systematic iterative scheme for selecting the all-pass filter for the construction in [14] is suggested in [16, 15]. In this case, when the basis is generated from one single repeated all-pass filter, a model is identified using a basis generated from some initial all-pass filter. The model is then reduced by means of balanced model reduction to some lower-order system, which is used to construct a new all-pass filter. The steps above are then repeated until a refined low order model is obtained. The method seems to converge in many cases and reduces the tail contribution of the error to almost zero after only three or four iterations. If the model order for the all-pass filter is selected as the true order of $G(z)$, the poles of the all-pass filter converge to the poles of $G(z)$ making only the first basis function necessary. However, in some cases the method diverges, especially in the presence of substantial measurement noise.

In [5], it was suggested that the thresholding rule of [9] should be used to regularize the system estimate if measurement noise is present. The suggestion builds on the observation that if, for some k and l,

$$(\hat{g}_{k,l} - g_{k,l})^2 > g_{k,l}^2, \tag{2.32}$$

a smaller error in the estimate is obtained by simply replacing $\hat{g}_{k,l}$ with 0. In this case the true coefficient $g_{k,l}$ is totally "drowned" in noise. Since we hope the system is well described with only a few basis functions while the noise is difficult to express in the basis, the threshold algorithm described in section 3.2 could be applied to this kind of basis function as well. Then, the problem is reduced to finding the proper

threshold level in order to detect when (2.32) happens. From (2.28), the asymptotic distribution of the coefficients is known. Suppose that the input sequence $\{u(t)\}$ has zero mean and constant power spectral density σ_u^2 and that the measurement noise $\{v(t)\}$ is white stationary zero mean with variance σ^2. Then

$$\sqrt{N}(\hat{g}_{k,l} - g_{k,l}) \sim \mathrm{AsN}(0, \sigma^2/\sigma_u^2). \tag{2.33}$$

Furthermore they are asymptotically independent for different (k, l). The threshold rule suggested in [9] then relies on the fact that

$$\mathrm{Prob}\left(\max_{\substack{1\le k\le m\\ 1\le l\le n_k}} |\hat{g}_{k,l} - g_{k,l}| > \frac{\sigma}{\sigma_u}\sqrt{\frac{2\log n}{N}} \right) \to 0 \tag{2.34}$$

as $m \to \infty$. This follows from a result in [18] which was used in [9].

Assume the problem formulation of (2.20), where the input sequence $\{u(t)\}_{t=1}^N$ has constant power spectral density σ_u^2 and zero mean. Let the measurement noise be defined as above so that (2.34) results in the following regularization scheme:

1. Choose a rational orthonormal basis constructed from a cascade of orthonormal all-pass filters $\{H_k\}_{k=1}^m$ with McMillan degrees n_k, respectively.
2. Calculate the least squares estimates $\hat{\Theta}$ of the least squares coefficients Θ as in (2.24).
3. Let

$$\hat{g}_{k,l}^s = \begin{cases} \hat{g}_{k,l} & \text{if } |\hat{g}_{k,l}| \ge \tau\frac{\sigma}{\sigma_u}, \\ 0 & \text{if } |\hat{g}_{k,l}| < \tau\frac{\sigma}{\sigma_u}, \end{cases} \qquad \begin{matrix} k = 1, \dots, m, \\ l = 1, \dots, n_k, \end{matrix} \tag{2.35}$$

where

$$\tau = \sqrt{\frac{2\log n}{N}}. \tag{2.36}$$

4. Form the thresholded transfer function estimate

$$\hat{G}_n^{(s)}(z) = \sum_{k=1}^{m}\sum_{l=1}^{n_k} \hat{g}_{k,l}^s \psi_{k,l}(z). \tag{2.37}$$

The algorithm results in only significant coefficients being kept. In order for it to work well, the system $G(z)$ should be as compressible as possible in the basis chosen. It is therefore important to incorporate as much a priori information as possible when selecting the all-pass filters $H_k(z)$. However, even when the all-pass filters have been selected, their ordering in the cascade has to be selected as well. The compressibility of a collection of all-pass filters can differ significantly depending on the order in which the filters appear. In parallel to the best basis selection of wavelet packets, given a collection of all-pass filters, a similar problem of best basis, or ordering, can be stated for rational orthonormal functions. In the following section, a scheme for selecting a suitable such basis is presented.

3. Best basis selection. As mentioned in the previous section, a collection of orthonormal all-pass filters can be used to construct an orthonormal rational basis. Different orderings of these filters will then correspond to different coordinate selections for the same subspace of $\mathcal{H}^2$. In this section, a scheme is given for selecting

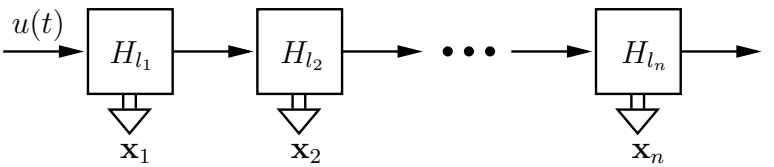


FIG. 3.1. *Cascade of orthogonal all-pass filters.*

the best among all possible coordinates. The outline is as follows. In section 3.1, the number of different bases and different basis functions is given. Section 3.2 considers different criteria for optimality. The selection scheme is derived in section 3.3 and finally, in section 3.4, some implementation issues are discussed.

3.1. Combinatorial observations. Using notation similar to that in section 2.1, a basis is associated with a collection of n balanced all-pass filters placed in an array as shown in Figure 3.1. Let the all-pass filter $H_k(z)$ have state transfer function $\Phi_k(z)$ and McMillan degree n_k. Suppose that there is a total of p *different* filters, represented by the family $\{H_l(z)\}_{l=1}^p$, and that the filter $H_l(z)$ appears exactly κ_l times for $l = 1, \ldots, p$. The number of filters n in the array then becomes

$$n = \sum_{l=1}^{p} \kappa_l. \tag{3.1}$$

Given the array in Figure 3.1, the vector valued basis functions will be

$$\Psi_k(z) = \Phi_{l_k}(z) \prod_{i=1}^{k-1} H_{l_i}(z), \qquad k = 1, \ldots, n, \tag{3.2}$$

for some ordering $\{l_k\}_{k=1}^n$. Define the finite-dimensional space

$$\mathcal{H}_n := \operatorname{span} \{\Psi_k\}_{k=1}^n \subset \mathcal{H}^2, \tag{3.3}$$

which is independent of the ordering of the filters since they will all have the same finite fractional expansions. Thus, each ordering of the functions will correspond to a selection of coordinates for $\mathcal{H}_n$. The number of different such choices is given by the following proposition, first presented in [4].

PROPOSITION 3.1 (number of bases). *Given the different orthogonal all-pass filters* $\{H_l(z)\}_{l=1}^p$ *and multiplicities* $\{\kappa\}_{l=1}^p$ *with* $n = \sum_{l=1}^p \kappa_l$, *the number of arrangements of these filters is*

$$\binom{n}{\kappa_1, \kappa_2, \ldots, \kappa_p} := \frac{n!}{\kappa_1! \cdot \kappa_2! \cdots \kappa_p!}. \tag{3.4}$$

Proof. The result is obtained from simple counting; see for example [12]. ∎

The number of bases can be described combinatorially as the number of "words" that can be written using the "letters" $\{H_l\}_{l=1}^p$, where each letter $H_l(z)$ appears κ_l times. A number of special cases can now be examined.

All filters equal. This results in the construction of [14] with the Laguerre or Kautz bases as special cases. Of course, there is only one way to arrange a number of equal filters and consequently, with $p = 1$ and $\kappa_1 = n$, Proposition 3.1 gives only $\binom{n}{n} = 1$ basis.

All filters different. This problem is identical to the number of arrangements of n distinct objects and is equal to $n!$, which is exactly what Proposition 3.1 gives with $p = n$ and $\kappa_l = 1$ for $l = 1, \ldots, n$.

Each filter appears κ times. This case can be viewed as starting with the construction of [14] with κ equal filters, where each of the filters has p factors. Then, Proposition 3.1 gives the number of bases that can be constructed by splitting up each filter into its factors and rearranging them. With $\kappa_l = \kappa$ for $l = 1, \ldots, p$ this results in $n!/(\kappa!)^p$ bases. For example, if each filter appears twice, the number of bases will be $n!/2^{n/2}$.

Two different filters. With only two different all-pass filters, the number of bases becomes $\binom{n}{\kappa}$, where κ is the number of filters $H_1(z)$ so that $H_2(z)$ appears $n-\kappa$ times.

Observe that a vector valued basis function, defined as in (3.2), depends on the state transfer function of its output filter $H_{l_k}(z)$ and on which filters precede it in the cascade, but not on their ordering. Obviously, there will be an overlap in which *different* basis functions there are in all the bases given by Proposition 3.1. The number of different basis functions is determined by the following proposition.

PROPOSITION 3.2 (number of basis functions). *Given different orthogonal all-pass filters $\{H_l(z)\}_{l=1}^p$ with multiplicities $\{\kappa_l\}_{l=1}^p$ and $n = \sum_{k=1}^p \kappa_l$, the number of vector valued basis functions as defined in* (3.2) *that can be constructed from these filters is*

$$\sum_{l=1}^{p} \left[\kappa_l \prod_{\substack{m \neq l \\ m=1}}^{p} (\kappa_m + 1) \right]. \tag{3.5}$$

Proof. A basis function is uniquely defined by the output filter $H_l(z)$ (or actually its state transfer function $\Phi_l(z)$) together with the p-tuple

$$(k_1, k_2, \ldots, k_p), \qquad \text{where} \quad \begin{cases} 0 \le k_m \le \kappa_m, & m \neq l, \\ 0 \le k_l \le \kappa_l - 1 & \text{otherwise,} \end{cases} \tag{3.6}$$

denoting that the array of filters preceding the output filter contains $H_m(z)$ k_m times for $m = 1, \ldots, p$. Each filter $H_m(z)$ can appear from 0 up to κ_m times in this array except for $H_l(z)$, which can appear only from 0 up to $\kappa_l - 1$ times, since the filter $H_l(z)$ is used as an output filter. Simple counting then gives a total of

$$(\kappa_1 + 1)(\kappa_2 + 1) \cdots (\kappa_{l-1} + 1)\, \kappa_l\, (\kappa_{l+1} + 1) \cdots (\kappa_p + 1) = \kappa_l \prod_{\substack{m \neq l \\ m=1}}^{p} (\kappa_m + 1) \tag{3.7}$$

basis functions having $H_l(z)$ as their output filter. The total number of functions (3.5) is then obtained as the sum over all possible output filters. ∎

The number of basis functions can now be calculated for the special cases considered above.

All filters equal. With $p = 1$ and $\kappa_1 = n$, (3.5) gives n. This is of course expected since there is only one basis.

All filters different. With $p = n$ and $\kappa_l = 1$ for $l = 1, \ldots, p$, the number of basis functions becomes $n2^{n-1}$. This can be checked since the filters preceding the output filter can be represented by an $(n-1)$-position binary number, which can take 2^{n-1} values, and since there are n possible output filters.

Each filter appears κ times. In this case, with $\kappa_l = \kappa$ for $l = 1, \ldots, p$, the number of functions will be $p\kappa(\kappa+1)^{n-1}$, according to (3.5).

Two different filters. With $H_1(z)$ appearing κ times, the number of basis functions will be $2\kappa(n-\kappa)+n$. This expression degenerates to n for $\kappa = 0$ or $\kappa = n$, which corresponds to the case when all the filters are equal. It is maximized for $\kappa = n/2$, where there are equally many of the two different filters. In this case, the number of functions is $n(n+2)/2$. This should be compared to the number of bases, which is $\binom{n}{n/2} \approx 2^{n+1/2}/\sqrt{\pi n}$ according to Stirling's formula. In this way there is an exponentially growing number of bases containing only a quadratically growing number of basis functions.

Propositions 3.1 and 3.2 show that in many cases, the number of basis functions is dramatically smaller than the number of orthonormal basis functions that can be constructed from these, while in other cases the difference is more modest. If, e.g., all the filters are equal and $n = 10$, the total number of bases is $n! = 3,628,800$ while the number of basis functions becomes only $n2^{n-1} = 5,120$. If there are 5 of each of 2 different all-pass filters, the number of bases is 252 while the number of functions is 60. In both these cases, the number of basis functions is smaller than the number of bases that can be constructed from them. In section 3.3, it will be shown that in the case of rational basis functions, a fast algorithm also exists that permits selection of the best basis.

3.2. Best basis criteria. In order for the thresholding of section 2.2 to work as an efficient regularization method in system identification, it is necessary to have a basis that provides a sparse system description. This situation is similar to wavelet-packet best basis selection. In the case of rational basis functions, the best ordering of the all-pass filters has to be found. This section discusses different optimality criteria or cost functions for determining what is meant by the best basis. The same cost functions were used for wavelet packet best basis selection in [7, 8, 10].

Let $\{g_k\}_{k\in\mathcal{B}}$ be the expansion coefficients of a finite orthonormal rational basis expansion, where $\mathcal{B}$ are the indices of n basis functions formed from all the state outputs from an array of all-pass filters. Let all possible $\mathcal{B}$ denote the bases that can be formed by permuting the ordering of the all-pass filters. In this way, a finite library of bases is obtained. In [8], it was suggested that the entropy of the expansion coefficients should be used as a cost function for best basis selection. This is defined by

$$V_e(\mathcal{B}) = -\sum_{k\in\mathcal{B}} |g_k|^2 \log |g_k| \tag{3.8}$$

and is a well-known data compression measure; see [38]. Another possible criterion for best basis selection is the ℓ^α-cost

$$V_\alpha(\mathcal{B}) = \sum_{k\in\mathcal{B}} |g_p|^\alpha \tag{3.9}$$

with $0 < \alpha < 2$. In particular, $\alpha = 1$ is often considered. Both these cost functions have the property that for free minimization subject only to the constraint

$$\sum_{k\in\mathcal{B}} |g_k|^2 = \text{Constant}, \tag{3.10}$$

they result in all g_k but one equal to zero. The corresponding maximization problem gives all $|g_k|$ equal. This indicates the ability to measure compression. However, for

best basis selection, the minimization is also constrained by the requirement that the coefficients of all different bases also must describe the same system.

A third cost function,

$$V(\mathcal{B},\tau)=\sum_{k\in\mathcal{B}}\min(|g_k|^2,\tau^2\sigma^2), \tag{3.11}$$

was suggested in [10]. This function is strongly connected to denoising by thresholding using the threshold level $\tau\sigma$, thus involving this parameter also in the best basis selection process. A more general form of (3.11) was suggested for speech denoising using wavelet packets in [3], where colored noise was taken into account.

The cost function (3.11) relates to the entropy cost and the ℓ^1-cost via the following result.

THEOREM 3.3. *Let $V_e(\mathcal{B})$, $V_1(\mathcal{B})$, and $V(\mathcal{B},\tau)$ be defined by* (3.8), (3.9), *and* (3.11), *respectively. Let*

$$C:=\sqrt{\sum_{k\in\mathcal{B}}|g_k|^2}, \tag{3.12}$$

which is constant for any choice of orthonormal basis. Then

$$V_e(\mathcal{B})=\int_0^{C/\sigma}V(\mathcal{B},\tau)\frac{d\tau}{\tau}-\frac{1}{2}C^2-C^2\log C \tag{3.13}$$

and

$$V_1(\mathcal{B})=\frac{1}{2\sigma}\int_0^{\infty}V(\mathcal{B},\tau)\frac{d\tau}{\tau^2}. \tag{3.14}$$

Proof. Integration of one term of $V(\mathcal{B})$ results in

$$\begin{aligned}
&\int_0^{C/\sigma}\min(|g_k|^2,\tau^2\sigma^2)\frac{d\tau}{\tau}=\sigma^2\int_0^{C/\sigma}\min\left(\frac{|g_k|^2}{\sigma^2},\tau^2\right)\frac{d\tau}{\tau}\\
&=\sigma^2\left\{\int_0^{|g_k|/\sigma}\tau^2\frac{d\tau}{\tau}+\int_{|g_k|/\sigma}^{C/\sigma}\frac{|g_k|^2}{\sigma^2}\frac{d\tau}{\tau}\right\}\\
&=\sigma^2\left\{\frac{1}{2}\frac{|g_k|^2}{\sigma^2}+\frac{|g_k|^2}{\sigma^2}\log\frac{C}{\sigma}-\frac{|g_k|^2}{\sigma^2}\log\frac{|g_k|}{\sigma}\right\}\\
&=\frac{1}{2}|g_k|^2+|g_k|^2\log C-|g_k|^2\log|g_k|.
\end{aligned} \tag{3.15}$$

Summation over $k\in\mathcal{B}$ of the left- and right-hand sides gives (3.13). Similarly,

$$\begin{aligned}
&\int_0^{\infty}\min(|g_k|^2,\tau^2\sigma^2)\frac{d\tau}{\tau^2}=\sigma^2\int_0^{\infty}\min\left(\frac{|g_k|^2}{\sigma^2},\tau^2\right)\frac{d\tau}{\tau^2}\\
&=\sigma^2\left\{\int_0^{|g_k|/\sigma}\tau^2\frac{d\tau}{\tau^2}+\int_{|g_k|/\sigma}^{\infty}\frac{|g_k|^2}{\sigma^2}\frac{d\tau}{\tau^2}\right\}=2\sigma|g_k|,
\end{aligned} \tag{3.16}$$

which via summation over $k\in\mathcal{B}$ results in (3.14). ∎

The theorem shows that the entropy and ℓ^1-cost in fact are weighted versions of (3.11), where the weight is taken (integrated) over many threshold levels.

The cost function (3.11) also has an interpretation in terms of rate distortion. If a signal is decomposed in the best basis according to this cost and reconstructed using only coefficients with a magnitude greater than τ, then the ℓ^2 reconstruction error is minimal over all bases in the library for any construction using the selected number of coefficients. See [26, 32].

The cost functions discussed above can now be used for selecting the best rational basis.

3.3. Selection scheme. In section 3.1, it was concluded that the number of bases that can be constructed from an array of all-pass filters in most cases is much larger than the number of basis functions forming these bases. Using wavelet packet terminology, there is a library of bases containing only a few different basis functions. This section provides a scheme for selecting the best among all possible bases in the library. The search is made in a complexity no larger than the number of different basis functions so that in this sense, the search can be called "fast."

In section 3.2, different criteria for best-basis selection were discussed. Which criterion to use is up to the user, but important for all these methods is that they describe compression so that a basis with the energy concentrated in few coefficients is selected. Supposing that the true system is well described by only a few coefficients in the finite-dimensional subspace $\mathcal{H}_n$ but that noise is present, a denoising scheme such as the one described in section 2.2 can be applied. This was suggested for orthonormal rational bases in [5]. Since only coefficients above the threshold level are kept, as few such coefficients as possible would give a smaller variance for the system estimate. This is a result of fewer parameters being estimated; see [22].

The selection criteria from wavelet packet basis selection presented in section 3.2 are directly applicable to the selection of the best orthonormal rational basis as well. The selection algorithm, however, becomes different. As for wavelet packets, it is crucial that the cost function be additive for the algorithm derived here. The following terminology will be used.

DEFINITION 3.4 (collection and word). *An m-collection is a set of m all-pass filters, where the filter $H_l(z)$ appears k_l times for $l = 1, \dots, p$. The m-collection is represented by the p-tuple*

$$(k_1, k_2, \dots, k_p), \qquad \text{where} \quad \sum_{l=1}^{p} k_l = m. \tag{3.17}$$

An m-word is an ordered m-collection.

The vector valued basis functions (3.2) are thus fully defined by the output filter together with a *collection* of preceding filters. A basis for $\mathcal{H}_n$, on the other hand, is defined by the *word* of all-pass filters constituting the array. With this terminology an algorithm for selection of the best basis can now be formulated. The algorithm is recursive: If the best m-word and its cost are known for each different m-collection, the best $(m+1)$-word and its cost can be calculated for all different $(m+1)$-collections. Formally, this can be written as follows.

ALGORITHM 3.5 (best basis selection). *The algorithm selects the best basis constructed from an array of n orthogonal all-pass filters. The given filters are $\{H_l(z)\}_{l=1}^{p}$ with multiplicities $\{\kappa_l\}_{l=1}^{p}$, where $n = \sum_{l=1}^{p} \kappa_l$. In other words, the output of the algorithm is the best n-word constructed from these filters.*

1. *If $m = 1$, all the best 1-words and their costs are trivial.*
2. *For all $(m+1)$-collections:*

(a) *For every l where $k_l \neq 0$ in the current $(m+1)$-collection: Construct the basis function having $H_l(z)$ as output filter preceded by the remaining m-collection. Calculate the cost of the expansion coefficient corresponding to each such basis function.*

(b) *For each basis function: Add its cost to the known best cost of the corresponding m-collection and select the best. The corresponding best $(m+1)$-word is given by the best m-word for the current m-collection with the selected output filter appended.*

3. *Terminate if $m = n-1$. Otherwise, increment m by one and return to step* 2.

The single n-word resulting from the algorithm will correspond to the globally optimal basis with respect to the chosen cost function. From the way the algorithm is constructed, it is important that the cost function be additive in the sense that the sum of the costs for two disjoint subsets of the coefficients is equal to the cost of the union of these two coefficient sets. This is fulfilled for all the cost functions considered in section 3.2. Additivity of the cost then implies that a suboptimal m-word cannot result in an optimal $(m+1)$-word, and in this way the resulting n-word has to be globally optimal. The number of steps for Algorithm 3.5 is given by the following theorem.

THEOREM 3.6 (algorithmic complexity). *The order of complexity of Algorithm* 3.5 *with respect to n and $\{\kappa_l\}_{l=1}^{p}$ is*

$$\sum_{l=1}^{p} \left[\kappa_l \prod_{\substack{m \neq l \\ m=1}}^{p} (\kappa_m + 1) \right], \tag{3.18}$$

which is the same as the number of basis functions given by Proposition 3.2.

Proof. The steps 2(a) and 2(b) of Algorithm 3.5 are performed once for each l having $k_l \neq 0$ in the current m-collection. Step 3.3 is performed once for each $(m+1)$-collection. These steps are repeated for $m = 1, \ldots, n-1$. The initial step takes p operations. In total this gives

$$\begin{aligned} p + \sum_{m=1}^{n-1} \left(\sum_{\substack{\forall (k_1,\ldots,k_p) \\ \sum k_l = m+1 \\ 0 \leq k_l \leq \kappa_l}} \left[\sum_{\substack{i=1 \\ k_i \neq 0}}^{p} 1 \right] \right) &= \sum_{\substack{i=1 \\ k_i \neq 0}}^{p} \left(\sum_{m=0}^{n-1} \left[\sum_{\substack{\forall (k_1,\ldots,k_p) \\ \sum k_l = m+1 \\ 0 \leq k_l \leq \kappa_l}} 1 \right] \right) \\ = \sum_{\substack{i=1 \\ k_i \neq 0}}^{p} \#\{(k_1, \ldots, k_p) | k_i \neq 0\} &= \sum_{l=1}^{p} \kappa_l \prod_{\substack{i=1 \\ i \neq l}}^{p} (1 + \kappa_i). \end{aligned} \tag{3.19}$$

∎

REMARK 3.7. *In the proof, the change of order of summation avoids using the number of $(m+1)$-collections. Due to the constraints imposed by the κ_l's, this number cannot be explicitly calculated. However, for given values of the κ_l's, the number of m-collections is the coefficient in front of x^m in the generating function $\prod_{l=1}^{p} \sum_{k=0}^{\kappa_l} x^k$.*

As will be shown in the next section, the $(m+1)$-collections can recursively be calculated from the m-collections without knowing in advance how many these are for a certain $m+1$. In this way, no search for the $(m+1)$-collections has to be made.

Theorem 3.6 shows that, as in the case of wavelet packet best basis selection, the number of algorithmic steps for selecting the best orthonormal rational basis is

the same as the number of different basis functions in all bases considered. The next section demonstrates how to recursively calculate m-collections and how to implement the algorithm.

3.4. Implementation. Algorithm 3.5, defined in section 3.3, performs the inner iteration once for each $(m+1)$-collection with respect to the multiplicities $\{\kappa_l\}_{l=1}^{p}$. The constraints imposed by the κ_l's result in difficulties in explicitly calculating all $(m+1)$-collections for a given m without having to search over all the collections with all different lengths. However, the following algorithm shows how all $(m+1)$-collections can be calculated recursively given all the m-collections.

ALGORITHM 3.8 (recursive collection generation). *Given all m-collections $(k_1, \ldots, k_p)$ with $0 \le k_l \le \kappa_l$ for $l = 1, \ldots, p$ so that $\sum_{l=1}^{p} k_l = m$, the following steps generate all $(m+1)$-collections.*

For each m-collection $(k_1, \ldots, k_p)$:

1. $l \leftarrow 0$
2. $l \leftarrow l + 1$
   (a) *If $k_l = 0$, increment k_l by 1 to generate an $(m+1)$-collection and repeat step 2 if $l < p$. Stop otherwise.*
   (b) *If $0 < k_l < \kappa_l$, increment k_l by 1 to generate an $(m+1)$-collection and stop.*
   (c) *Stop if $k_l = \kappa_l$.*

Proof. The following arguments show that the mapping performed by Algorithm 3.8 from all m-collections to the $(m+1)$-collections is onto: First of all, it is evident that if there is any output from the algorithm, it will be an $(m+1)$-collection, and the only time when there is no output is when $k_l = \kappa_l$ for all $l = 1, \ldots, p$. It then remains to show that any $(m+1)$-collection always comes from the algorithm: Take an arbitrary $(m+1)$-collection $(k_1, \ldots, k_p)$. If $k_1 > 0$, it comes from $(k_1 - 1, k_2, \ldots, k_p)$ via Algorithm 3.8. If $k_1 = 0$ and $k_2 > 0$, it comes from the m-collection $(0, k_2 - 1, k_3, \ldots, k_p)$ via the algorithm. If $k_1 = k_2 = 0$ and $k_3 > 0$ it comes from $(0, 0, k_3 - 1, \ldots, k_p)$, and so on. Thus, any $(m+1)$-collection was generated from the algorithm and the mapping is onto. This inversion also shows that no $(m+1)$-collection is produced more than once by the algorithm. Note here that, in step 2(a), the $l < p$ condition is only needed if $m = 0$. ∎

The algorithm above provides a fast way to recursively calculate collections with different number sums. For implementation of the best basis selection algorithm, it is also convenient to have some systematic addressing of the collections. This is to say that an integer must be assigned to each of the collections in a unique way. The total number of collections, or p-tuples $(k_1, \ldots, k_p)$, with $0 \le k_l \le \kappa_l$ for $l = 1, \ldots, p$ becomes $\prod_{l=1}^{p}(\kappa_l+1)$. Thus, the collections can be numbered from 0 to $\prod_{l=1}^{p}(\kappa_l+1)-1$ in the following way.

DEFINITION 3.9 (addressing). *Let $0 \le k \le \kappa_l$ for $l = 1, \ldots, p$ and denote a collection by*

$$\mathbf{c}_q = (k_1, \ldots, k_p), \tag{3.20}$$

where q is an integer in $\{0, 1, \ldots, \prod_{l=1}^{p}(\kappa_l + 1) - 1\}$, calculated as

$$q := \sum_{l=1}^{p} k_l \prod_{j=1}^{l-1}(1 + \kappa_j), \tag{3.21}$$

and where the product over no elements is defined as 1.

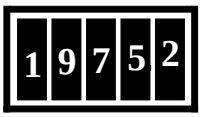


FIG. 3.2. *Counter.*

This way of numbering the collections is in fact a multibase representation with respect to the κ_l's. One interpretation of the numbering is that the collection $\mathbf{c}_q$ can be viewed as the qth number on the counter shown in Figure 3.2, where the lth wheel is numbered from 0 to κ_l. This is also quite the same as the numbering of time, where for example hours, days, and weeks all have different numbers of their respective subquantities minutes, hours, and days. The numbering and recursive collection calculation is demonstrated with the following example.

EXAMPLE 3.10. *Let* $p = 3$ *with* $\kappa_1 = 3$, $\kappa_2 = 1$, *and* $\kappa_3 = 2$. *The total number of collections is then* 24 *and with the addressing of Definition* 3.9, *they become*

$$
\begin{array}{lll}
\mathbf{c}_0 = (0,0,0), & \mathbf{c}_8 = (0,0,1), & \mathbf{c}_{16} = (0,0,2),\\
\mathbf{c}_1 = (1,0,0), & \mathbf{c}_9 = (1,0,1), & \mathbf{c}_{17} = (1,0,2),\\
\mathbf{c}_2 = (2,0,0), & \mathbf{c}_{10} = (2,0,1), & \mathbf{c}_{18} = (2,0,2),\\
\mathbf{c}_3 = (3,0,0), & \mathbf{c}_{11} = (3,0,1), & \mathbf{c}_{19} = (3,0,2),\\
\mathbf{c}_4 = (0,1,0), & \mathbf{c}_{12} = (0,1,1), & \mathbf{c}_{20} = (0,1,2),\\
\mathbf{c}_5 = (1,1,0), & \mathbf{c}_{13} = (1,1,1), & \mathbf{c}_{21} = (1,1,2),\\
\mathbf{c}_6 = (2,1,0), & \mathbf{c}_{14} = (2,1,1), & \mathbf{c}_{22} = (2,1,2),\\
\mathbf{c}_7 = (3,1,0), & \mathbf{c}_{15} = (3,1,1), & \mathbf{c}_{23} = (3,1,2).
\end{array}
\tag{3.22}
$$

The recursion of Algorithm 3.8 *can be cranked with the trivial collection* $\mathbf{c}_0$. *This gives*

$m=0$	$m=1$	$m=2$	$m=3$	$m=4$	$m=5$	$m=6$
$\mathbf{c}_0$	$\mathbf{c}_1$	$\mathbf{c}_2$	$\mathbf{c}_3$	$\mathbf{c}_7$	$\mathbf{c}_{15}$	$\mathbf{c}_{23}$
	$\mathbf{c}_4$	$\mathbf{c}_5$	$\mathbf{c}_6$	$\mathbf{c}_{11}$	$\mathbf{c}_{19}$	
	$\mathbf{c}_8$	$\mathbf{c}_9$	$\mathbf{c}_{10}$	$\mathbf{c}_{14}$	$\mathbf{c}_{22}$	
		$\mathbf{c}_{12}$	$\mathbf{c}_{13}$	$\mathbf{c}_{18}$		
		$\mathbf{c}_{16}$	$\mathbf{c}_{17}$	$\mathbf{c}_{21}$		
			$\mathbf{c}_{20}$			

(3.23)

It is easily checked that in each column of (3.23), *the collections have the same number sum* $\sum_{k=1}^{3} k_l$.

The addressing of Definition 3.9 gives a convenient representation of the collections when implementing Algorithm 3.5. In step 2 of this algorithm, the $(m+1)$-collections are recursively calculated using Algorithm 3.8. In step 2(a), the remaining m-collection for the current l is obtained by reducing the current k_l with 1 from the current $(m+1)$-collection. Using the addressing of Definition 3.9, this reduction can be made directly simply by subtracting $\prod_{j=1}^{l-1}(1+\kappa_j)$ from the address of the current $(m+1)$-collection. This results in the address to the corresponding m-collection. The subtracted products are found as the addresses for $m = 1$. In this way, the collections will not have to be explicitly represented by anything other than the addresses. In step 2(a), $k_l \neq 0$ can be checked for the addresses by successively dividing by the κ_l's in order to examine if these are contained as factors in the remainder of the current $(m+1)$-collection. This is shown in more detail in Appendix A. For example (from

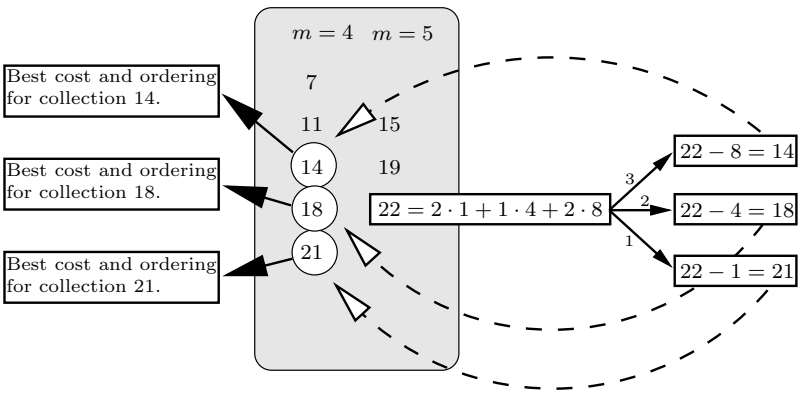


FIG. 3.3. *Addressing in Algorithm* 3.5.

Example 3.10 above), by subtracting 1, 4, and 8 respectively, the address 22 with number sum $m = 5$ will then directly point to the addresses 21, 18, and 14, all having number sums $m = 4$. This is illustrated in Figure 3.3.

Several implementation aspects were treated in this section, which concludes the derivation of a best basis selection scheme. The following section provides a way to calculate the expansion coefficients associated with each of the different basis functions. These coefficients are the essential input to Algorithm 3.5.

4. Decomposition. The expansion coefficients of a system can be obtained using system identification as described in section 2.2. This results in the coefficients for only one fixed basis. The selection scheme derived in section 3, however, needs the coefficients for all possible basis functions. In this section, an algorithm is given that uses the expansion coefficients in one single basis in order to produce all the coefficients needed for the best basis selection scheme. The algorithm uses local block two-by-two orthogonal matrix transformations and has the same complexity as the best basis selection scheme of section 3.

The local orthogonal transformations are derived in section 4.1 where a method is also given for generating coefficients in the case where all the all-pass filters are different from each other. The latter is provided for understanding of the general coefficient-generation scheme presented in section 4.2.

4.1. Calculating coefficients by orthonormal transformations. Suppose that the expansion coefficients are given in one basis. This can be achieved with the system-identification approach presented in section 2.2. Then, there is an orthonormal, and thus well-conditioned, matrix transformation of these coefficients to the coefficients of any other possible basis. Using the inherent recursive structure of the rational basis functions, this orthonormal transformation can be decomposed into local (block) two-by-two orthonormal transformations.

Given a family of different all-pass filters $\{H_k\}_{k=1}^n$, the basis functions for an ordering $\{l_k\}_{k=1}^n$ of the filters can then be written as in (3.2). Consider two basis functions of consecutive order

$$\Psi_k(z) = \Phi_{l_k}(z) \prod_{i=1}^{k-1} H_{l_i}(z), \tag{4.1}$$

$$\Psi_{k+1}(z) = \Phi_{l_{k+1}}(z) H_{l_k}(z) \prod_{i=1}^{k-1} H_{l_i}(z). \tag{4.2}$$

By interchanging the order of the output filter of $\Psi_k(z)$ and the output filter of $\Psi_{k+1}(z)$, two basis functions of a basis defined by another ordering are obtained. The functions can be written

$$\tilde{\Psi}_k(z) = \Phi_{l_{k+1}}(z) \prod_{i=1}^{k-1} H_{l_i}(z), \tag{4.3}$$

$$\tilde{\Psi}_{k+1}(z) = \Phi_{l_k}(z) H_{l_{k+1}}(z) \prod_{i=1}^{k-1} H_{l_i}(z). \tag{4.4}$$

The two pairs of functions relate via the linear transformation $\mathbf{T}(l_k, l_{k+1})$ as

$$\begin{bmatrix} \Psi_k(z) \\ \Psi_{k+1}(z) \end{bmatrix} = \mathbf{T}(l_k, l_{k+1}) \begin{bmatrix} \tilde{\Psi}_k(z) \\ \tilde{\Psi}_{k+1}(z) \end{bmatrix}. \tag{4.5}$$

Denote the outer product between two column vector valued functions $\Phi_j(z)$ and $\Phi_k(z)$ with possibly different dimension by

$$\langle \Phi_j, \Phi_k \rangle := \frac{1}{2\pi} \int_{-\pi}^{\pi} \Phi_j(e^{i\omega}) \left[\Phi_k(e^{i\omega})\right]^H d\omega, \tag{4.6}$$

where H denotes transposition and complex conjugation. Then, from the orthonormality of the basis functions, the transformation matrix $\mathbf{T}$ becomes

$$\mathbf{T}(l_k, l_{k+1}) = \begin{bmatrix} \langle \Psi_k, \tilde{\Psi}_k \rangle & \langle \Psi_k, \tilde{\Psi}_{k+1} \rangle \\ \langle \Psi_{k+1}, \tilde{\Psi}_k \rangle & \langle \Psi_{k+1}, \tilde{\Psi}_{k+1} \rangle \end{bmatrix}, \tag{4.7}$$

and since all the four basis functions have the common factor $\prod_{i=1}^{k-1} H_{l_i}$, T can be rewritten as

$$\mathbf{T}(l_k, l_{k+1}) = \begin{bmatrix} \langle \Phi_{l_k}, \Phi_{l_{k+1}} \rangle & \langle \Phi_{l_k}, \Phi_{l_k} H_{l_{k+1}} \rangle \\ \langle \Phi_{l_{k+1}} H_{l_k}, \Phi_{l_{k+1}} \rangle & \langle \Phi_{l_{k+1}} H_{l_k}, \Phi_{l_k} H_{l_{k+1}} \rangle \end{bmatrix}. \tag{4.8}$$

Rewriting (4.7) as (4.8) is crucial for the existence of a fast coefficient transformation scheme. Let $\mathbf{T}(j,k)$ be partitioned into blocks as

$$\mathbf{T}(j,k) = \begin{bmatrix} \mathbf{t}_{11}(j,k) & \mathbf{t}_{12}(j,k) \\ \mathbf{t}_{21}(j,k) & \mathbf{t}_{22}(j,k) \end{bmatrix}, \tag{4.9}$$

where the four blocks, given by

$$\mathbf{t}_{11}(j,k) = \langle \Phi_j, \Phi_k \rangle, \tag{4.10}$$

$$\mathbf{t}_{12}(j,k) = \langle \Phi_j, \Phi_j H_k \rangle, \tag{4.11}$$

$$\mathbf{t}_{21}(j,k) = \langle \Phi_k H_j, \Phi_k \rangle, \tag{4.12}$$

$$\mathbf{t}_{22}(j,k) = \langle \Phi_k H_j, \Phi_j H_k \rangle, \tag{4.13}$$

are orthonormal, i.e., the matrix satisfies $\mathbf{T}(j,k)\mathbf{T}^T(j,k) = \mathbf{I}$. (This is shown in Appendix B, where the orthonormality follows because the two pairs of orthonormal functions $\{\Phi_j, \Phi_k H_j\}$ and $\{\Phi_k, \Phi_j H_k\}$ have the same span.)

The transformation matrix $\mathbf{T}$ can be calculated as the solution of a Lyapunov-like equation. Define

$$\mathbf{X}_{j,k}(z) := \begin{bmatrix} \Phi_j(z) \\ \Phi_k(z) H_j(z) \end{bmatrix} \tag{4.14}$$

so that $\mathbf{T}(j,k) = \langle \mathbf{X}_{j,k}, \mathbf{X}_{k,j} \rangle$. Let $H_j(z)$ and $H_k(z)$ have the state-space realizations $(\mathbf{A}_j, \mathbf{B}_j, \mathbf{C}_j, D_j)$ and $(\mathbf{A}_k, \mathbf{B}_k, \mathbf{C}_k, D_k)$ respectively. From (2.15), $\mathbf{X}_{j,k}(z)$ has the (balanced) state-space realization

$$(\mathbf{A}_{j,k}, \mathbf{B}_{j,k}) = \left(\begin{bmatrix} \mathbf{A}_j & \mathbf{0} \\ \mathbf{B}_k \mathbf{C}_j & \mathbf{A}_k \end{bmatrix}, \begin{bmatrix} \mathbf{B}_j \\ \mathbf{B}_k D_j \end{bmatrix} \right). \tag{4.15}$$

In this way, $\mathbf{T}(j,k)$ can be written

$$\begin{aligned} \mathbf{T}(j,k) &= \langle \mathbf{X}_{j,k}, \mathbf{X}_{k,j} \rangle \\ &= \frac{1}{2\pi} \int_{-\pi}^{\pi} \left(\mathbf{I} e^{i\omega} - \mathbf{A}_{j,k} \right)^{-1} \mathbf{B}_{j,k} \left[\left(\mathbf{I} e^{-i\omega} - \mathbf{A}_{k,j} \right)^{-1} \mathbf{B}_{k,j} \right]^T d\omega \\ &= \sum_{n=0}^{\infty} \sum_{m=0}^{\infty} \mathbf{A}_{j,k}^n \mathbf{B}_{j,k} \left\{ \frac{1}{2\pi} \int_{-\pi}^{\pi} e^{-i\omega(n-m)} d\omega \right\} \mathbf{B}_{k,j}^T (\mathbf{A}_{k,j}^T)^m \\ &= \sum_{n=0}^{\infty} \mathbf{A}_{j,k}^n \mathbf{B}_{j,k} \mathbf{B}_{k,j}^T (\mathbf{A}_{k,j}^T)^n, \end{aligned} \tag{4.16}$$

which solves the equation

$$\mathbf{A}_{j,k} \mathbf{T}(j,k) \mathbf{A}_{k,j}^T + \mathbf{B}_{j,k} \mathbf{B}_{k,j}^T = \mathbf{T}(j,k). \tag{4.17}$$

As shown in Appendix C, this equation can easily be transformed into a Sylvester equation using bilinear transformations. This equation can then be solved with standard methods. Such equations are often used in feedback design of large-scale multivariable systems; see, e.g., [29].

In terms of the block components of $\mathbf{T}(j,k)$, this equation can be divided into the four coupled matrix equations

$$\mathbf{t}_{11} = \mathbf{A}_j \mathbf{t}_{11} \mathbf{A}_k^T + \mathbf{B}_j \mathbf{B}_k^T, \tag{4.18}$$

$$\mathbf{t}_{12} = \mathbf{A}_j \mathbf{t}_{11} \mathbf{C}_k^T \mathbf{B}_j^T + \mathbf{A}_j \mathbf{t}_{12} \mathbf{A}_j^T + \mathbf{B}_j D_k \mathbf{B}_j^T, \tag{4.19}$$

$$\mathbf{t}_{21} = \mathbf{B}_k \mathbf{C}_j \mathbf{t}_{11} \mathbf{A}_k^T + \mathbf{A}_k \mathbf{t}_{21} \mathbf{A}_k^T + \mathbf{B}_k D_j \mathbf{B}_k^T, \tag{4.20}$$

$$\begin{aligned} \mathbf{t}_{22} &= \mathbf{B}_k \mathbf{C}_j \mathbf{t}_{11} \mathbf{C}_k^T \mathbf{B}_j^T + \mathbf{B}_k \mathbf{C}_j \mathbf{t}_{12} \mathbf{A}_j^T \\ &\quad + \mathbf{A}_k \mathbf{t}_{21} \mathbf{C}_k^T \mathbf{B}_j^T + \mathbf{A}_k \mathbf{t}_{22} \mathbf{A}_j^T + \mathbf{B}_k D_j D_k \mathbf{B}_j^T, \end{aligned} \tag{4.21}$$

where the argument (j,k) of $\mathbf{t}_{11}$, $\mathbf{t}_{12}$, $\mathbf{t}_{21}$, and $\mathbf{t}_{22}$ was suppressed. The equations can either be solved "top-down" due to their triangular structure or as the total solution to (4.17). The following gives an example of a transformation matrix $\mathbf{T}(j,k)$.

EXAMPLE 4.1 (generalized Laguerre functions). *The generalized Laguerre functions with real poles*

$$\psi_k(z) = \frac{\sqrt{1-a_k^2}}{z - a_k} \prod_{j=1}^{k-1} \frac{1 - a_j z}{z - a_j}, \qquad k = 1, 2, \ldots, \tag{4.22}$$

are obtained from an array built of different first order all-pass filters. The transformation matrix (4.8) *becomes*

$$\mathbf{T}(j,k) = \frac{1}{1-a_j a_k}\begin{bmatrix} \sqrt{1-a_j^2}\sqrt{1-a_k^2} & a_j - a_k \\ a_k - a_j & \sqrt{1-a_j^2}\sqrt{1-a_k^2} \end{bmatrix}. \tag{4.23}$$

Suppose now that $\Psi_k(z)$ and $\Psi_{k+1}(z)$, defined by (4.1), have the respective expansion coefficients $\mathbf{g}_k$ and $\mathbf{g}_{k+1}$. Using (4.5), the expansion coefficients associated with the basis functions $\tilde{\Psi}_k(z)$ and $\tilde{\Psi}_{k+1}(z)$ are then given by

$$\begin{bmatrix} \tilde{\mathbf{g}}_k \\ \tilde{\mathbf{g}}_{k+1} \end{bmatrix} = \mathbf{T}^T(l_k, l_{k+1}) \begin{bmatrix} \mathbf{g}_k \\ \mathbf{g}_{k+1} \end{bmatrix}. \tag{4.24}$$

In this way, given the expansion coefficients of a pair of functions in one basis, the expansion coefficients for the functions obtained from alternating the two last all-pass filters can be obtained by a block two-by-two transformation. The question is whether the coefficients of all possible basis functions can be reached by repeating such transformations. Next, it will be shown that this is the case, but in order to facilitate the description, a more compact notation for the expansion coefficients will be introduced.

DEFINITION 4.2 (expansion-coefficient notation). *Given a family of different all-pass filters* $\{H_k(z)\}_{k=1}^n$, *an expansion coefficient associated with a basis function constructed from these filters can then be denoted by* $\mathbf{g}(p,k)$, *where* p *denotes the address to the subset of preceding all-pass filters and* k *is the number associated with the output filter. The addressing is performed as in Definition* 3.9, *which in the case where all the filters are different reduces to*

$$p = \sum_{l=1}^{n} d_l 2^{l-1}, \tag{4.25}$$

where d_l *is either* 1 *or* 0 *denoting whether or not the filter* $H_l(z)$ *is included in the set. The output filter* $H_k(z)$ *cannot belong to the subset* p.

Before introducing the scheme for calculating all the coefficients, the idea for coefficient generation will be illustrated with an example.

EXAMPLE 4.3 (coefficient generation). *Suppose that all coefficients belonging to functions generated from the filters* $H_1(z)$, $H_2(z)$, *and* $H_3(z)$ *are known. Suppose also that* $\mathbf{g}(\{1,2,3\},4)$ *is known. Here, the subset address is replaced by the subset itself in order to make the description easier to follow. To start with, all the coefficients generated from arrays of maximum length* 4 *become*

$$\begin{aligned}
\mathbf{g}(\{1,2,4\},3) &= \begin{bmatrix} \mathbf{t}_{12}^T & \mathbf{t}_{22}^T \end{bmatrix} \begin{bmatrix} \mathbf{g}(\{1,2\},3) \\ \mathbf{g}(\{1,2,3\},4) \end{bmatrix}, \\
\mathbf{g}(\{1,3,4\},2) &= \begin{bmatrix} \mathbf{t}_{12}^T & \mathbf{t}_{22}^T \end{bmatrix} \begin{bmatrix} \mathbf{g}(\{1,3\},2) \\ \mathbf{g}(\{1,2,3\},4) \end{bmatrix}, \\
\mathbf{g}(\{2,3,4\},1) &= \begin{bmatrix} \mathbf{t}_{12}^T & \mathbf{t}_{22}^T \end{bmatrix} \begin{bmatrix} \mathbf{g}(\{2,3\},1) \\ \mathbf{g}(\{1,2,3\},4) \end{bmatrix}, \\
\mathbf{g}(\{1,2\},4) &= \begin{bmatrix} \mathbf{t}_{11}^T & \mathbf{t}_{21}^T \end{bmatrix} \begin{bmatrix} \mathbf{g}(\{1,2\},3) \\ \mathbf{g}(\{1,2,3\},4) \end{bmatrix},
\end{aligned}$$

$$\mathbf{g}(\{1,3\},4) = \begin{bmatrix}\mathbf{t}_{11}^T & \mathbf{t}_{21}^T\end{bmatrix}\begin{bmatrix}\mathbf{g}(\{1,3\},2)\\ \mathbf{g}(\{1,2,3\},4)\end{bmatrix},$$

$$\mathbf{g}(\{2,3\},4) = \begin{bmatrix}\mathbf{t}_{11}^T & \mathbf{t}_{21}^T\end{bmatrix}\begin{bmatrix}\mathbf{g}(\{2,3\},1)\\ \mathbf{g}(\{1,2,3\},4)\end{bmatrix}.$$

*Next, coefficients coming from length-*3 *arrays are produced by*

$$\mathbf{g}(\{1,4\},2) = \begin{bmatrix}\mathbf{t}_{12}^T & \mathbf{t}_{22}^T\end{bmatrix}\begin{bmatrix}\mathbf{g}(\{1\},2)\\ \mathbf{g}(\{1,2\},4)\end{bmatrix},$$

$$\mathbf{g}(\{2,4\},1) = \begin{bmatrix}\mathbf{t}_{12}^T & \mathbf{t}_{22}^T\end{bmatrix}\begin{bmatrix}\mathbf{g}(\{2\},1)\\ \mathbf{g}(\{1,2\},4)\end{bmatrix},$$

$$\mathbf{g}(\{1,4\},3) = \begin{bmatrix}\mathbf{t}_{12}^T & \mathbf{t}_{22}^T\end{bmatrix}\begin{bmatrix}\mathbf{g}(\{1\},3)\\ \mathbf{g}(\{1,3\},4)\end{bmatrix},$$

$$\mathbf{g}(\{3,4\},1) = \begin{bmatrix}\mathbf{t}_{12}^T & \mathbf{t}_{22}^T\end{bmatrix}\begin{bmatrix}\mathbf{g}(\{3\},1)\\ \mathbf{g}(\{1,3\},4)\end{bmatrix},$$

$$\mathbf{g}(\{2,4\},3) = \begin{bmatrix}\mathbf{t}_{12}^T & \mathbf{t}_{22}^T\end{bmatrix}\begin{bmatrix}\mathbf{g}(\{2\},3)\\ \mathbf{g}(\{2,3\},4)\end{bmatrix},$$

$$\mathbf{g}(\{3,4\},2) = \begin{bmatrix}\mathbf{t}_{12}^T & \mathbf{t}_{22}^T\end{bmatrix}\begin{bmatrix}\mathbf{g}(\{3\},2)\\ \mathbf{g}(\{2,3\},4)\end{bmatrix},$$

$$\mathbf{g}(\{1\},4) = \begin{bmatrix}\mathbf{t}_{11}^T & \mathbf{t}_{21}^T\end{bmatrix}\begin{bmatrix}\mathbf{g}(\{1\},2)\\ \mathbf{g}(\{1,2\},4)\end{bmatrix},$$

$$\mathbf{g}(\{2\},4) = \begin{bmatrix}\mathbf{t}_{11}^T & \mathbf{t}_{21}^T\end{bmatrix}\begin{bmatrix}\mathbf{g}(\{2\},3)\\ \mathbf{g}(\{2,3\},4)\end{bmatrix},$$

$$\mathbf{g}(\{3\},4) = \begin{bmatrix}\mathbf{t}_{11}^T & \mathbf{t}_{21}^T\end{bmatrix}\begin{bmatrix}\mathbf{g}(\{3\},1)\\ \mathbf{g}(\{1,3\},4)\end{bmatrix}.$$

*Finally, the length-*2 *arrays give the coefficients*

$$\mathbf{g}(\{4\},1) = \begin{bmatrix}\mathbf{t}_{12}^T & \mathbf{t}_{22}^T\end{bmatrix}\begin{bmatrix}\mathbf{g}(\{\emptyset\},1)\\ \mathbf{g}(\{1\},4)\end{bmatrix},$$

$$\mathbf{g}(\{4\},2) = \begin{bmatrix}\mathbf{t}_{12}^T & \mathbf{t}_{22}^T\end{bmatrix}\begin{bmatrix}\mathbf{g}(\{\emptyset\},2)\\ \mathbf{g}(\{2\},4)\end{bmatrix},$$

$$\mathbf{g}(\{4\},3) = \begin{bmatrix}\mathbf{t}_{12}^T & \mathbf{t}_{22}^T\end{bmatrix}\begin{bmatrix}\mathbf{g}(\{\emptyset\},3)\\ \mathbf{g}(\{3\},4)\end{bmatrix},$$

$$\mathbf{g}(\{\emptyset\},4) = \begin{bmatrix}\mathbf{t}_{11}^T & \mathbf{t}_{21}^T\end{bmatrix}\begin{bmatrix}\mathbf{g}(\{\emptyset\},1)\\ \mathbf{g}(\{1\},4)\end{bmatrix}.$$

The reason the total $\mathbf{T}$ *is not used in the transforms is to avoid too many coefficients being generated. Note that above,* $\begin{bmatrix}\mathbf{t}_{11}^T & \mathbf{t}_{21}^T\end{bmatrix}$ *is used fewer times than* $\begin{bmatrix}\mathbf{t}_{12}^T & \mathbf{t}_{22}^T\end{bmatrix}$. *The number of coefficients known is* $3 \cdot 2^2 + 1 = 13$ *and the total number of coefficients associated with* 4 *filters is* $4 \cdot 2^3 = 32$. *In the example,* 19 *coefficients were generated, which is exactly the difference.*

It is now possible to formulate a coefficient generation scheme for the case where the all-pass filters are different. Assume as in the example above that all the coefficients associated with basis functions constructed from the filters $H_1(z), \ldots, H_{m-1}(z)$ are known. Furthermore, assume that $\mathbf{g}(2^{m-1}-1, m)$ also is known. Then, the rest of the coefficients associated with the basis functions constructed from the filters $H_1(z), \ldots, H_m(z)$ are generated by the following algorithm.

ALGORITHM 4.4 (recursive step for coefficient generation).

1. *Let* $l = m - 1$.
2. *For all possible* l*-collections* p_l *that can be constructed from the all-pass filters* $H_1(z), \ldots, H_{m-1}(z)$, *let*

$$\mathbf{g}(p_l - 2^{k-1} + 2^{m-1}, k) = \begin{bmatrix}\mathbf{t}_{12}^T(k,m) & \mathbf{t}_{22}^T(k,m)\end{bmatrix}\begin{bmatrix}\mathbf{g}(p_l - 2^{k-1}, k)\\ \mathbf{g}(p_l, m)\end{bmatrix}$$

for all k in the current l-collection.

3. *For all possible $(l-1)$-collections p_{l-1} that can be constructed from the all-pass filters $H_1(z), \ldots, H_{m-1}(z)$: Pick one k such that $H_k(z)$ does not belong to the current $(l-1)$-collection and let*

$$\mathbf{g}(p_{l-1}, m) = \begin{bmatrix} \mathbf{t}_{11}^T(k,m) & \mathbf{t}_{21}^T(k,m) \end{bmatrix} \begin{bmatrix} \mathbf{g}(p_{l-1}, k) \\ \mathbf{g}(p_{l-1} + 2^{k-1}, m) \end{bmatrix}.$$

4. *Terminate if $l = 1$, decrease l by 1 otherwise, and go to step 2.*

The following theorem guarantees that all coefficients can be generated given the coefficients for one basis.

THEOREM 4.5 (coefficient generation). *Given the coefficients $\mathbf{g}(2^{m-1}-1, m)$, $m = 1, \ldots, n$ of the canonical basis, all the other $n2^{n-1} - n$ coefficients are generated by performing Algorithm 4.4 recursively for $m = 1, \ldots, n$.*

Proof. Steps 2 and 3 of Algorithm 4.4 trivially generate coefficients not known before. It then remains to check that the number of coefficients generated equals $n2^{n-1} - n$. Step 2 is performed $\binom{m-1}{l} l$ times while step 4.4 is performed $\binom{m-1}{l-1}$ times. Thus for fixed m, Algorithm 4.4 generates

$$\sum_{l=1}^{m-1} \binom{m-1}{l} l + \binom{m-1}{l-1} = (m-1)2^{m-2} + 2^{m-1} - 1 \tag{4.26}$$

coefficients. Then, the algorithm is performed for $m = 1, \ldots, n$ so that the total number of coefficients produced becomes

$$\sum_{m=1}^{n} (m-1)2^{m-2} + 2^{m-1} - 1 = n2^{n-1} - n, \tag{4.27}$$

which is exactly the number of coefficients needed. ∎

The following section extends the coefficient generation procedure to cover the general rational basis case.

4.2. General coefficient calculation. Using the same two-by-two transformations as in the previous section, the coefficient generation scheme can now be derived for the general basis case where the all-pass filters can have multiplicities.

Given a family of all-pass filters $\{H_l\}_{l=1}^p$ with balanced realizations and multiplicities $\{\kappa_l\}_{l=1}^p$ denoting that each filter $H_l(z)$ appears κ_l times, the total number of filters is $n := \sum_{l=1}^p \kappa_l$. As in the previous section, denote an expansion coefficient $\mathbf{g}(\beta, l)$, meaning that the output filter in the corresponding basis function is $H_l(z)$ and that the filter is preceded by the collection of filters with address β. The addressing of the collections of filters is performed as suggested in Definition 3.9.

Suppose now that the expansion coefficients

$$\mathbf{g}(\beta_k, l_k), \qquad k = 1, \ldots, n, \tag{4.28}$$

for one basis are given. The basis is defined by the ordering $\{l_k\}_{k=1}^n$ of the filters in the array. In this way, β_k is a $(k-1)$-collection. The first collection will have the address $\beta_1 = 0$ and the other addresses are given by

$$\beta_{k+1} = \beta_k + \prod_{j=1}^{l_k - 1} (\kappa_j + 1), \qquad k = 1, \ldots, (n-1). \tag{4.29}$$

The following algorithm is now suggested for generating all the other coefficients that are obtained by rearranging the array of all-pass filters in all possible different ways.

ALGORITHM 4.6 (general coefficient generation). *Given the expansion coefficients for one basis, as described above, repeat the following steps for* $m = 2, \dots, n$*:*

1. *Let* $k \leftarrow m-1$ *and* $a_1 \leftarrow \beta_m$.
2. *For each* a_j*:*
   - *For every different* $H_l(z)$ *in the collection represented by* a_j *except for the filter* $H_{l_m}(z)$*:*

   (a) *Let*

$$\mathbf{g}\left(a_j - \prod_{i=1}^{l-1}(\kappa_i+1) + \prod_{i=1}^{l_m-1}(\kappa_i+1), l\right) = \begin{bmatrix}\mathbf{t}_{12}^T(l,l_m) & \mathbf{t}_{22}^T(l,l_m)\end{bmatrix}\begin{bmatrix}\mathbf{g}(a_j - \prod_{i=1}^{l-1}(\kappa_i+1), l) \\ \mathbf{g}(a_j, l_m)\end{bmatrix}. \tag{4.30}$$

   (b) *Let* $\tilde{a}_{j,l} \leftarrow a_j - \prod_{i=1}^{l-1}(\kappa_i+1)$.
3. *Delete all* a_j *and assign a new* $a_{j'} \leftarrow \tilde{a}_{j,l}$ *for each unique* $\tilde{a}_{j,l}$.
4. *For each* a_j*, pick one* $H_l(z)$ *from the collection represented by* $\beta_m - a_j$ *and let*

$$\mathbf{g}(a_j, l_m) = \begin{bmatrix}\mathbf{t}_{11}^T(l,l_m) & \mathbf{t}_{21}^T(l,l_m)\end{bmatrix}\begin{bmatrix}\mathbf{g}(a_j, l) \\ \mathbf{g}(a_j + \prod_{i=1}^{l-1}(\kappa_i+1), l_m)\end{bmatrix}. \tag{4.31}$$

5. *Terminate if* $k = 1$*, decrease* k *by* 1 *otherwise, delete all* $\tilde{a}_{j,l}$*, and go to step* 2.

Before the algorithm is proved, a couple of examples of how it is executed will be given. A different number of steps is performed for each m depending on the initially given ordering of the all-pass filters.

EXAMPLE 4.7 (general coefficient generation). *Let* $p = 3$ *and let* $\kappa_1 = 3$, $\kappa_2 = 1$, *and* $\kappa_3 = 2$. *The case is the same as in Example* 3.10. *Two different initial orderings will be examined in order to demonstrate how the algorithm executes for different cases. First, let the given basis be defined by the ordering*

$$\begin{pmatrix} l_1 & l_2 & l_3 & l_4 & l_5 & l_6 \end{pmatrix} = \begin{pmatrix} 1 & 1 & 1 & 2 & 3 & 3 \end{pmatrix}. \tag{4.32}$$

Then, Algorithm 4.6 *produces*

(0 0 0):1	(1 0 0):1	(2 0 0):1	(3 0 0):2	(3 1 0):3	(3 1 1):3
(0 0 0):2 (6)	(1 0 0):2 (4)	(2 0 0):2 (2)	(2 1 0):1 (1)	(2 1 1):1 (7)	(2 1 2):1 (24)
(0 0 0):3 (23)	(0 1 0):1 (5)	(1 1 0):1 (3)	(2 1 0):3 (9)	(3 0 1):2 (8)	(3 0 2):2 (25)
	(0 1 0):3 (19)	(1 1 0):3 (14)	(3 0 0):3 (10)	(2 1 1):3 (26)	
	(1 0 0):3 (20)	(2 0 0):3 (15)	(1 1 1):1 (11)	(3 0 1):3 (27)	
	(0 0 1):2 (21)	(0 1 1):1 (16)	(2 0 1):2 (12)	(1 1 2):1 (28)	
	(0 0 1):1 (22)	(1 0 1):2 (17)	(2 0 1):1 (13)	(2 0 2):2 (29)	
	(0 0 1):3 (40)	(1 0 1):1 (18)	(1 1 1):3 (31)	(2 0 2):1 (30)	
		(0 1 1):3 (36)	(2 0 1):3 (32)		
		(1 0 1):3 (37)	(0 1 2):1 (33)		
		(0 0 2):2 (38)	(1 0 2):2 (34)		
		(0 0 2):1 (39)	(1 0 2):1 (35)		

where the entries

$$(k_1\ k_2\ k_3) : l\ (\ j) \tag{4.33}$$

denote a coefficient for the basis function with l *as output filter, preceded by* k_i *filters* $H_i(z)$, $i = 1, 2, 3$. *The number* j *denotes that the entry is the* j*th coefficient generated by the algorithm. Grey-shaded entries are generated in step* 4 *of the algorithm while the nonshaded are produced in step* 2(a).

Let now the initially known coefficients be defined by the ordering

$$\begin{pmatrix} l_1 & l_2 & l_3 & l_4 & l_5 & l_6 \end{pmatrix} = \begin{pmatrix} 1 & 2 & 3 & 1 & 3 & 1 \end{pmatrix}. \tag{4.34}$$

The algorithm then executes as

(0 0 0):1	(1 0 0):2	(1 1 0):3	(1 1 1):1	(2 1 1):3	(2 1 2):1
(0 0 0):2 (2)	(0 1 0):1 (1)	(0 1 1):1 (3)	(2 0 1):2 (10)	(1 1 2):1 (17)	(3 0 2):2 (29)
(0 0 0):3 (9)	(0 1 0):3 (5)	(1 0 1):2 (4)	(2 1 0):3 (11)	(2 0 2):2 (18)	(3 1 1):3 (30)
	(1 0 0):3 (6)	(1 0 1):1 (12)	(1 1 1):3 (19)	(2 0 2):1 (31)	
	(0 0 1):2 (7)	(1 1 0):1 (13)	(2 0 1):3 (20)	(2 1 1):1 (32)	
	(0 0 1):1 (8)	(2 0 0):3 (14)	(0 1 2):1 (21)	(3 0 1):3 (33)	
	(1 0 0):1 (16)	(2 0 0):2 (15)	(1 0 2):2 (22)	(3 0 1):2 (34)	
	(0 0 1):3 (28)	(0 1 1):3 (24)	(1 0 2):1 (23)	(3 1 0):3 (35)	
		(1 0 1):3 (25)	(2 0 1):1 (36)		
		(0 0 2):2 (26)	(2 1 0):1 (37)		
		(0 0 2):1 (27)	(3 0 0):3 (38)		
		(2 0 0):1 (40)	(3 0 0):2 (39)		

which corresponds to a different execution pattern. In both cases, all 40 *coefficients are generated given the* 6 *coefficients in one basis. From Proposition* 3.2, *this is exactly the number of different basis functions in all the bases that can be constructed from the* 6 *all-pass filters.*

With the example above in mind, Algorithm 4.6 can now be proved.

Proof. Consider a fixed m in step 1 of the algorithm and suppose that all the coefficients in the previous steps have been calculated. Say now that the collection with address β_m has k_l filters $H_l(z)$, $l = 1, \ldots, p$. Then for fixed m, l_m is added to this collection. The new coefficients to be calculated will then be those based on exactly $k_{l_m} + 1$ filters $H_{l_m}(z)$, where the output filter also is counted. These can be divided into three parts:

(i) The new known coefficient $\mathbf{g}(\beta_m, l_m)$.
(ii) All the other $\mathbf{g}(a, l_m)$ where the collection a has k_{l_m} filters $H_{l_m}(z)$.
(iii) All coefficients $\mathbf{g}(b, l)$ where the collection b has $k_{l_m} + 1$ filters $H_{l_m}(z)$ and consequently $l \neq l_m$.

By construction, the coefficients of type (ii) are recursively generated in step 4 of the algorithm while the coefficients of type (iii) are generated recursively by step 2(a). That all coefficients are obtained is guaranteed by the fact that for each k in the algorithm the number of filters, except H_{l_m}, is reduced by 1 in all possible ways. □

REMARK 4.8. *The total number of coefficients of types* (i) *and* (ii) *becomes*

$$\prod_{\substack{j=1 \\ j \neq l_m}}^{p} (k_j + 1) \tag{4.35}$$

while the number of coefficients of type (iii) *is*

$$\sum_{\substack{l=1 \\ l \neq l_m}}^{p} k_l \prod_{\substack{j=1 \\ j \neq l \\ j \neq l_m}}^{p} (k_j + 1). \tag{4.36}$$

In this way the total number of new coefficients generated for a fixed m is the sum of (4.35) *and* (4.36)*, which does not depend on the number of the filter $H_{l_m}(z)$ in the array.*

This concludes the contributions of this paper. A complete scheme for decomposing a system in the coefficients of a large number of orthonormal bases and for selecting the best has been given. In the following section, the best basis scheme is examined for a number of examples.

5. Examples. The procedure for selecting the best orthonormal rational basis can now be investigated in some examples. The first example concerns an exact finite description in order to demonstrate the consequences of selecting different all-pass filter orderings. Section 5.2 then examines pole selection in a noise-free setting. Finally, in section 5.3, best basis selection is applied to delay estimation.

5.1. Finite description. In order to demonstrate best rational basis selection, this example assumes that the system is exactly described by a finite number of basis functions. This means that the poles of the system are exactly the poles of the basis functions. In this way, the problem is reduced to a pure problem of calculating the best basis with respect to some cost function, while the estimation procedure can be omitted. Let the system transfer function be

$$G(z) = \sum_{k=1}^{n} c_k \frac{\sqrt{1 - a_k^2}}{z - a_k}, \tag{5.1}$$

where all the a_k's are real and different with magnitude less than one, and where the square-root factor normalizes the norm of each of the terms to $|c_k|$. Choose as bases the different generalized Laguerre systems constructed with the very same poles. The all-pass filters will thus be

$$H_k(z) = \frac{1 - a_k z}{z - a_k}, \qquad k = 1, \ldots, n, \tag{5.2}$$

while the state transfer functions become

$$\phi_k(z) = \frac{\sqrt{1 - a_k^2}}{z - a_k}, \qquad k = 1, \ldots, n. \tag{5.3}$$

The basis functions for some given ordering $\{k_l\}_{l=1}^{n}$ are then given by

$$\psi_l(z) = \phi_{k_l}(z) \prod_{j=1}^{l-1} H_{k_j}(z), \qquad l = 1, \ldots, n. \tag{5.4}$$

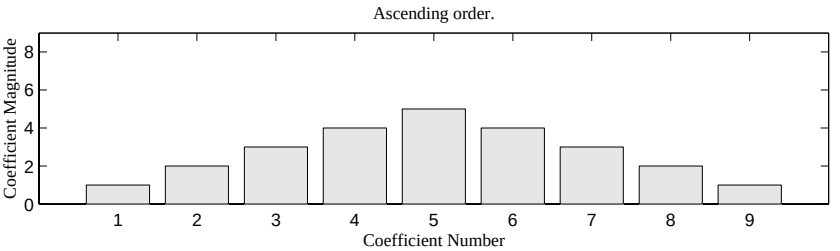


FIG. 5.1. *Coefficients for the ascending pole order.*

In this way, the system can now exactly be described by any of the $n!$ different bases that can be constructed. For fixed ordering, the expansion is written

$$G(z) = \sum_{l=1}^{n} g_l \psi_l(z), \tag{5.5}$$

where the coefficients $g_l = \langle G, \psi_l \rangle$ are explicitly calculated as

$$\begin{aligned} \langle G, \psi_l \rangle &= \frac{1}{2\pi} \int_{-\pi}^{\pi} G(e^{i\omega}) \psi_l(e^{-i\omega}) d\omega \\ &= \sqrt{1 - a_{k_l}^2} \sum_{m=1}^{n} c_m \frac{\sqrt{1 - a_m^2}}{1 - a_{k_l} a_m} \prod_{j=1}^{l-1} \frac{a_m - a_{k_j}}{1 - a_{k_j} a_m} \end{aligned} \tag{5.6}$$

for $l = 1, \ldots, n$.

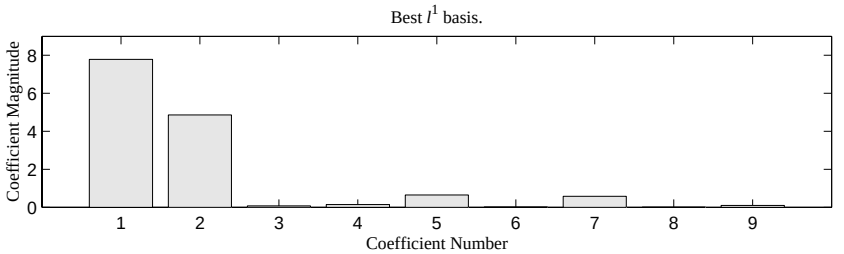
The best basis selection was performed with $n = 9$ and $a_k = k/10$, $k = 1, \ldots, 9$. In this way, $362,880$ bases are searched in $2,304$ steps. Two different cases were examined: In the first case, c_l was chosen so that $g_l = 5 - |l - 5|$ for the ascending pole ordering $n_l = l$, $l = 1, \ldots, 9$. Figure 5.1 shows the expansion coefficients for the ascending pole ordering. Two different cost functions were considered: the ℓ^1-cost

$$\sum_{l=1}^{n} |g_l| \tag{5.7}$$

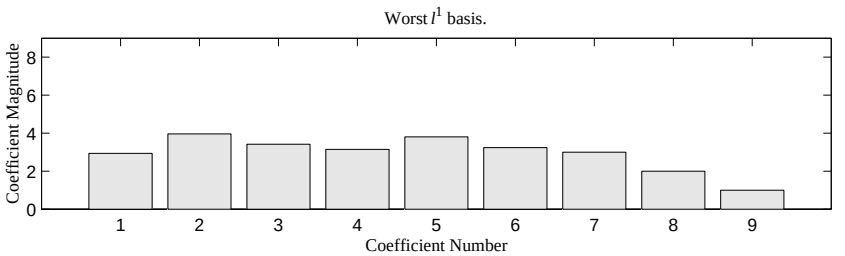
and the entropy

$$-\sum_{l=1}^{n} |g_l|^2 \log |g_l|. \tag{5.8}$$

The magnitude of the best and worst ℓ^1 expansion coefficients are shown in Figure 5.2. In the same way, this is shown for the best and worst entropy expansions in Figure 5.3. The second case considered is constructed so that $g_l = 1$ for the ascending order $k_l = l$, $l = 1, \ldots, 9$. The best and worst ℓ^1 expansion coefficients are displayed in Figure 5.4. For the entropy cost, exactly the same best and worst respective bases are selected as in the ℓ^1 case. The resulting costs and orderings for the two different cases are shown in Table 5.1. For case 1, the best expansions differ while the worst give the same result. As mentioned above, in case 2 both the best and the worst expansions match for the two different cost functions. Also the worst expansions for case 2 become the



(a) Best expansion.



(b) Worst expansion.

FIG. 5.2. *Best and worst ℓ^1 expansion coefficients for case* 1.

same. The latter should not be surprising since in this case, the solution is feasible and optimal for the two optimization problems

$$\min_{g_l} \sum_{l=1}^{n} |g_l| \quad \text{subject to} \quad \sum_{l=1}^{n} |g_l|^2 = \text{Constant} \tag{5.9}$$

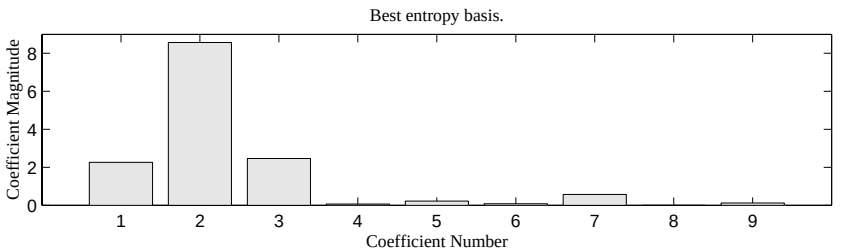
and

$$\min_{g_l} \sum_{l=1}^{n} |g_l|^2 \log |g_l|, \qquad \text{subject to} \quad \sum_{l=1}^{n} |g_l|^2 = \text{Constant}. \tag{5.10}$$

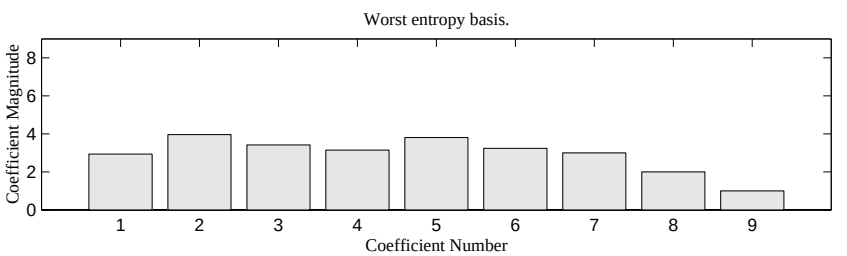
This example shows the difference in coefficient compression between different choices of basis and, in this way, the importance of selecting the order of the all-pass filters in the array properly.

5.2. Noise-free estimation. In this case, the system is estimated without measurement noise. As a priori information, three different alternatives are given for the complex poles of the system while the real pole is assumed to be known. One of the complex poles is identical to the true one and the other two are somewhat displaced. The system is given by

$$G(z) = C \frac{b_1(z) b_2(z)}{(z - K) a(z)^2}, \tag{5.11}$$



(a) Best expansion.



(b) Worst expansion.

FIG. 5.3. *Best and worst entropy expansion coefficients for case* 1.

TABLE 5.1
Costs and orderings for the two cases.

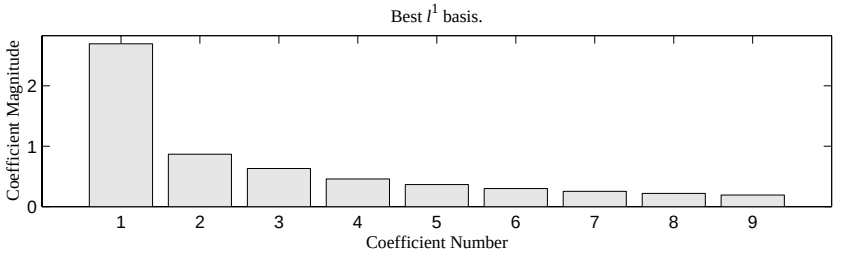
Case 1			
Type	ℓ^1-cost	Entropy cost	Ordering $k_1, \ldots, k_9$
Best ℓ^1	14.22	−161.21	9, 8, 2, 1, 7, 6, 3, 4, 5
Worst ℓ^1	26.53	−101.16	6, 4, 2, 1, 3, 5, 7, 8, 9
Best entropy	14.37	−167.00	5, 9, 7, 8, 1, 2, 4, 3, 6
Worst entropy	26.53	−101.16	6, 4, 2, 1, 3, 5, 7, 8, 9

Case 2			
Type	ℓ^1-cost	Entropy cost	Ordering $k_1, \ldots, k_9$
Best ℓ^1	5.99	−6.28	9, 8, 7, 6, 5, 4, 3, 2, 1
Worst ℓ^1	9.00	0.00	1, 2, 3, 4, 5, 6, 7, 8, 9
Best entropy	5.99	−6.28	9, 8, 7, 6, 5, 4, 3, 2, 1
Worst entropy	9.00	0.00	1, 2, 3, 4, 5, 6, 7, 8, 9

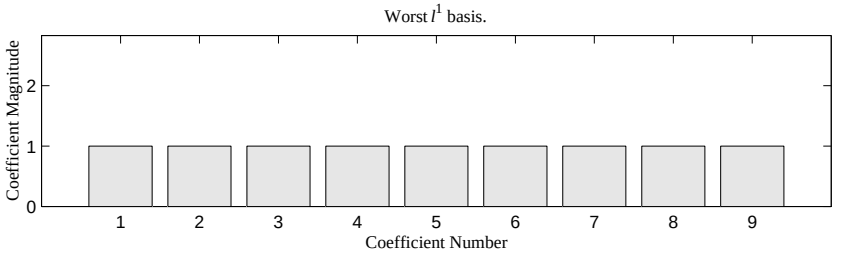
where

$$b_1(z) = z^2 - 2r\cos(\phi + \Delta\phi)z + r^2,$$
$$b_2(z) = z^2 - 2r\cos(\phi - \Delta\phi)z + r^2,$$
$$a(z) = z^2 - 2r\cos(\phi)z + r^2,$$

with $r = 0.80$, $\phi = 1.30\,\pi/4$, $\Delta\phi = 0.20\,\pi/4$, and $K = 0.50$. C is a normalizing constant achieving $G(1) = 1$. The system output y is generated as $y(t) = G(q)u(t)$, $t = 1, \ldots, N$, where q is the forward shift operator and the input sequence $\{u(t)\}_{t=1}^{N}$ is a unit pseudorandom binary signal (PRBS). The number of samples N is chosen



(a) Best expansion.



(b) Worst expansion.

FIG. 5.4. *Best and worst ℓ^1 expansion coefficients for case* 2. *These also match the best and worst entropy expansions for case* 2.

to be $1,024$.

The orthonormal basis used for estimation is constructed from the four different filters H_1, H_2, H_3, and H_4, where H_1 is a Laguerre filter with a pole in 0.50, H_2 is a Kautz filter with the same poles as the roots of $z^2 - 2r\cos(\phi_{d_1})z + r_d^2$, H_3 is a Kautz filter with the same poles as the roots of $z^2 - 2r\cos(\phi_{d_2})z + r_d^2$, and H_4 is a Kautz filter with the same poles as $a(z)$.

The displaced parameters ϕ_{d_1}, ϕ_{d_2}, and r_d are disturbed by 5% from the original so that $\phi_{d_1} = 1.05\,\phi$, $\phi_{d_2} = 0.95\,\phi$, and $r_d = 1.05\,r$. The corresponding multiplicities for the filters are

$$\begin{pmatrix} \kappa_1 & \kappa_2 & \kappa_3 & \kappa_4 \end{pmatrix} = \begin{pmatrix} 1 & 2 & 2 & 2 \end{pmatrix}. \tag{5.12}$$

With this setting, the best basis among 630 is chosen in 135 steps. The idea is to see if the best basis selection scheme chooses the correct complex pole over the displaced parameters so that the first three filters in the best array are H_1, H_4, and H_4, while the other basis functions have zero coefficients. The initial ordering of the filters was chosen as

$$\begin{pmatrix} l_1 & \cdots & l_7 \end{pmatrix} = \begin{pmatrix} 1 & 2 & 3 & 4 & 2 & 3 & 4 \end{pmatrix} \tag{5.13}$$

and estimation of the corresponding coefficients was performed as in (2.24) from section 2.2. The resulting coefficients are shown in Figure 5.5. Calculation of all the other possible coefficients and selection of the best basis was performed with the given conditions. For simplicity of the presentation, only the ℓ^1-cost was considered. The

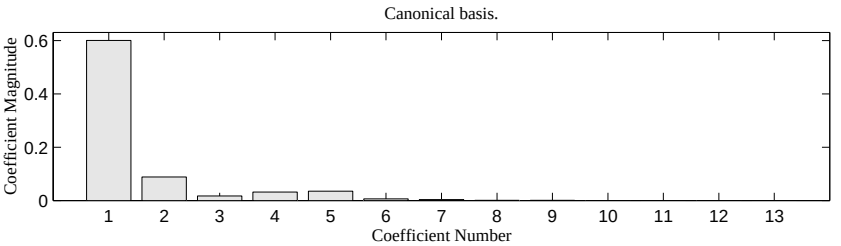


FIG. 5.5. *Estimated coefficients for the initial ordering of the all-pass filters.*

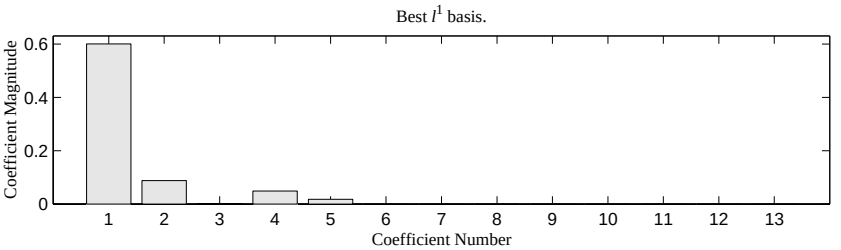


FIG. 5.6. *Coefficients for the best ℓ^1-ordering of the all-pass filters.*

TABLE 5.2
Costs and orderings for the initial, best, and worst bases.

Type	ℓ^1-cost	Ordering $l_1, \ldots, l_7$
Initial	0.786	$1, 2, 3, 4, 2, 3, 4$
Best ℓ^1	0.757	$1, 4, 4, 2, 3, 3, 2$
Worst ℓ^1	1.742	$2, 3, 2, 3, 4, 4, 1$

resulting coefficients are shown in Figure 5.6 while the ordering and cost are presented in Table 5.2. In the table, the worst cost and ordering are also shown.

The corresponding worst coefficients can be found in Figure 5.7. As intended, the best basis selection scheme chooses a basis constructed from an array starting with H_1, H_4, and H_4. This means that the system is exactly described by the first five coefficients. With the initial ordering, on the other hand, more coefficients are needed. One can also note that, not surprisingly, the worst basis has the first order all-pass filter last in the array.

This example shows how the coefficient calculation and selection scheme chooses a suitable ordering of the all-pass filters.

5.3. Multiple delays. In this example, estimation of a system with multiple delays is examined. The system is shown in Figure 5.8. The purpose is to estimate the delays from choosing the best basis. Thus, for clarity, the poles are assumed to be known so that

$$G_j(z) = \frac{1 - a_j}{z - a_j}, \qquad j = 1, 2, 3, \tag{5.14}$$

with $a_1 = 0.30$, $a_2 = 0.40$, and $a_3 = 0.50$. The delays are $d_1 = 5$, $d_2 = 10$, and $d_3 = 10$, and are unknown. The system can be viewed as echo measurements with two additional bounces.

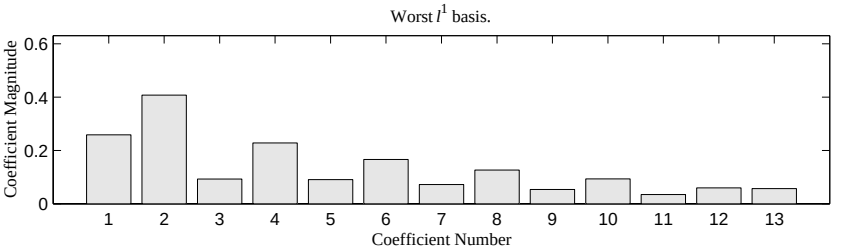


FIG. 5.7. *Coefficients for the worst ℓ^1-ordering of the all-pass filters.*

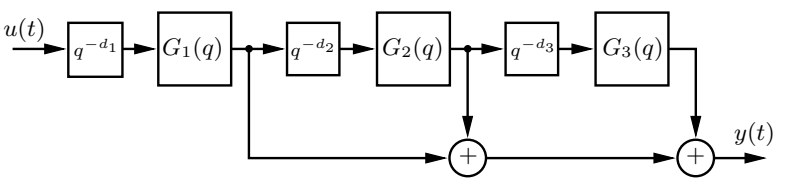


FIG. 5.8. *System with three delays.*

Simulations were performed with $\{u(t)\}_{t=1}^{1,024}$ chosen as PRBS; Gaussian measurement noise with standard deviation 1.00 was added to the output and the system was estimated in a basis consisting of the four different filters

$$H_j(z) = \frac{1 - a_j z}{z - a_j}, \qquad j = 1, 2, 3, \tag{5.15}$$

and

$$H_4(z) = z^{-1}. \tag{5.16}$$

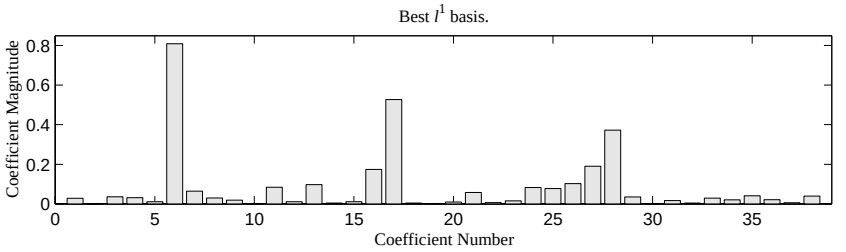
The multiplicities of these filters were

$$\begin{pmatrix} \kappa_1 & \kappa_2 & \kappa_3 & \kappa_4 \end{pmatrix} = \begin{pmatrix} 1 & 1 & 1 & 35 \end{pmatrix}. \tag{5.17}$$

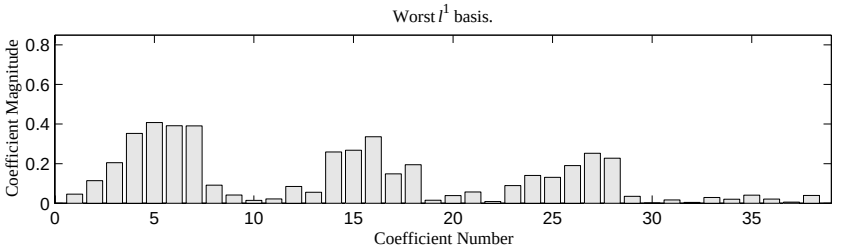
With a large number of delays in the bases, and with the remaining dynamics similar to that of the system, the best basis selection scheme should be able to identify the proper delays. The total number of possible bases is 50,616 while the number of different basis functions is only 712. Figure 5.9 shows the magnitude of the best and worst ℓ^1-coefficients.

Although the measurement noise is quite large, the best basis shows significantly larger coefficients in positions 6, 17, and 28. The corresponding filters are H_1, H_2, and H_3, respectively so that the rest of the all-pass filters are the unit delays H_4. The number of delays before coefficient 6 is 5 and corresponds to d_1. Between the coefficients 6 and 17 the number of delays is 10, corresponding to d_2. Finally, the number of delays between the coefficients 28 and 17 is 10, which is the same as d_3. The three different delays can thus be identified in this way. The ordering of the worst basis starts with H_3 and H_2 followed by 4 unit delays, which are followed by H_1. All the rest of the filters are delays.

The example shows that, for a system with this delay structure, the delays can be estimated by selecting the best orthonormal rational basis expansion even when substantial measurement noise is present.



(a) Best expansion.



(b) Worst expansion.

FIG. 5.9. *Best and worst ℓ^1-expansion coefficients for the system shown in Figure* 5.8.

6. Concluding remarks. This paper gives a method for calculating coefficients and selecting the best basis among a large number of different orthonormal rational bases.

The number of possible bases was first derived and the observation that a basis function can be a member of many different bases was made. This was exploited in order to derive a scheme for selecting the best orthonormal basis. The selection scheme is fast in the sense that the best basis is selected in a number of steps with a complexity no larger than the number of different basis functions. A similar scheme was then derived for calculating all possible different expansion coefficients, providing the coefficients in only one basis. The method is based on recursive local orthonormal transformations of the already calculated coefficients.

These two main results give a complete method for separately calculating all the possible coefficients and then selecting the best basis in an order of complexity which is the same as the number of different basis functions rather than the number of different bases. This property is very similar to that of the wavelet packet best basis methods of [7].

As indicated in the beginning of section 1, the different bases can be viewed as results from Gram–Schmidt orthogonalizations of simple filters having the given poles. Depending on the order in which the orthogonalization is performed, different bases for the finite-dimensional subspace are received. In this way, the best basis selection of Algorithm 3.5 is not only applicable to subspaces in $\mathcal{H}^2$ but can in fact be applied to any finite-dimensional space. However, the issue of multiplicities as well as the coefficient calculation of Algorithm 4.6 becomes specific to the construction of rational bases.

Appendix A. Address to collection conversion. Given an address

$$q = \sum_{l=1}^{p} k_l \prod_{j=1}^{l-1} (1 + \kappa_j)$$

to a collection $(k_1, \ldots, k_p)$ with respect to the multiplicities $\{\kappa_l\}_{l=1}^{p}$, the numbers $\{k_l\}_{l=1}^{p}$ are extracted from q by the following pseudocode:

for $l = 1, \ldots, p$
    $k_l \leftarrow \mathrm{rem}(q, \kappa_l + 1)$
    $q \leftarrow (q - k_l)/(\kappa_l + 1)$
end

Appendix B. Orthonormality of $\mathbf{T}(j,k)$. The matrix $\mathbf{T}(j,k)$ defined by (4.8) fulfills $\mathbf{T}(j,k)\mathbf{T}^T(j,k) = \mathbf{I}$ since

$$\mathrm{span}\{\Phi_j, \Phi_k H_j\} = \mathrm{span}\{\Phi_k, \Phi_j H_k\}. \tag{B.1}$$

These two pairs of functions have the same span since they both have the same poles and are strictly proper, and thus have the same partial fractional expansions.

Appendix C. Transforming $\mathbf{A_1 X A_2^T + B_1 B_2^T = X}$. Given the matrix equation

$$\mathbf{A}_1 \mathbf{X} \mathbf{A}_2^T + \mathbf{B}_1 \mathbf{B}_2^T = \mathbf{X}, \tag{C.1}$$

introduce

$$\tilde{\mathbf{A}}_k = (\mathbf{A}_k + \mathbf{I})^{-1}(\mathbf{A}_k - \mathbf{I}), \tag{C.2}$$

$$\tilde{\mathbf{B}}_k = \sqrt{2}(\mathbf{A}_k + \mathbf{I})^{-1}\mathbf{B}_k \tag{C.3}$$

for $k = 1, 2$. Then, $\mathbf{X}$ solves the Sylvester equation

$$\tilde{\mathbf{A}}_1 \mathbf{X} + \mathbf{X} \tilde{\mathbf{A}}_2^T + \tilde{\mathbf{B}}_1 \tilde{\mathbf{B}}_2^T = \mathbf{0}. \tag{C.4}$$

REFERENCES

[1] L. V. Ahlfors, *Complex Analysis*, McGraw–Hill, New York, 1953.

[2] P. Bodin, T. Oliveira e Silva, and B. Wahlberg, *On the construction of orthonormal basis functions for system identification*, in Proceedings of the 13th International Federation of Automatic Control World Congress, Vol. I, San Francisco, CA, 1996, pp. 369–374.

[3] P. Bodin and L. F. Villemoes, *Spectral subtraction in the time-frequency domain using wavelet packets*, in Proceedings of the IEEE Workshop on Speech Coding, Pocono Manor, PA, 1997, pp. 47–48.

[4] P. Bodin, L. F. Villemoes, and B. Wahlberg, *An algorithm for selection of best orthonormal rational basis*, in Proceedings of the 36th IEEE Conference on Decision and Control, San Diego, CA, 1997, pp. 1277–1282.

[5] P. Bodin and B. Wahlberg, *Thresholding in high order transfer function estimation*, in Proceedings of the 33rd IEEE Conference on Decision and Control, Vol. 4, Orlando, FL, 1994, pp. 3400–3405.

[6] P. W. Broome, *Discrete orthonormal sequences*, J. ACM, 12 (1965), pp. 151–168.

[7] R. R. Coifman and M. V. Wickerhauser, *Best-Adapted Wave Packet Bases*, Tech. report, Numerical Algorithms Group, Dept. of Mathematics, Yale University, New Haven, CT, 1990.

[8] R. R. Coifman and M. V. Wickerhauser, *Entropy-based algorithms for best basis selection*, IEEE Trans. Inform. Theory, 38 (1992), pp. 713–718.

[9] D. L. Donoho, *De-noising by soft-thresholding*, IEEE Trans. Inform. Theory, IT-41 (1995), pp. 613–649.

[10] D. L. Donoho and I. M. Johnstone, *Ideal denoising in an orthonormal basis chosen from a library of bases*, C. R. Acad. Sci. Paris Sér. I Math., 319 (1994), pp. 1317–1322.

[11] N. F. Dudley Ward and J. R. Partington, *Robust identification in the disc algebra using rational wavelets and orthonormal basis functions*, Internat. J. Control, 64 (1996), pp. 409–423.

[12] R. P. Grimaldi, *Discrete and Combinatorial Mathematics*, Addison–Wesley, Reading, MA, 1989.

[13] F. Gustafsson, *Identification of sparse linear regressions*, in Proceedings of the 13th International Federation of Automatic Control World Congress, Vol. J, San Francisco, CA, 1996, pp. 203–208.

[14] P. S. C. Heuberger, P. M. J. Van den Hof, and O. H. Bosgra, *A generalized orthonormal basis for linear dynamical systems*, IEEE Trans. Automat. Control, 40 (1995), pp. 451–465.

[15] P. S. C. Heuberger, P. M. J. Van den Hof, and O. H. Bosgra, *A generalized orthonormal basis for linear dynamical systems*, in Proceedings of the 32nd IEEE Conference on Decision and Control, San Antonio, TX, 1993, pp. 2850–2855.

[16] P. S. C. Heuberger, P. M. J. Van den Hof, and O. H. Bosgra, *Modeling linear dynamical systems through generalized orthonormal basis functions*, in Proceedings of the 12th International Federation of Automatic Control World Congress, Vol. 5, Sydney, Australia, 1993, pp. 283–286.

[17] T. Kitamori, *Applications of orthogonal functions to the determination of process dynamic characteristics and to the construction of self-optimizing control systems*, in Proceedings of the First International Congress of the International Federation of Automatic Control, Vol. II, Moscow, Soviet Union, 1960, pp. 613–617.

[18] M. R. Leadbetter, G. Lindgren, and H. Rootzén, *Extremes and Related Properties of Random Sequences and Processes*, Springer-Verlag, New York, 1983.

[19] L. Ljung, *System Identification: Theory for the User*, Prentice–Hall, Englewood Cliffs, NJ, 1987.

[20] C. T. Mullis and R. A. Roberts, *Roundoff noise in digital filters: Frequency transformations and invariants*, IEEE Trans. Acoustics, Speech and Signal Processing, 24 (1976), pp. 538–550.

[21] B. M. Ninness and F. Gustafsson, *A unifying construction of orthonormal bases for system identification*, IEEE Trans. Automat. Control, 42 (1997), pp. 515–521.

[22] B. M. Ninness, H. Hjalmarsson, and F. Gustafsson, *The fundamental role of orthonormal bases in system identification*, IEEE Trans. Automat. Control, 44 (1999), pp. 1384–1406.

[23] T. Oliveira e Silva, *A N-width result for the generalized orthonormal basis function model*, in Proceedings of the 13th International Federation of Automatic Control World Congress, Vol. 1, San Francisco, CA, 1996, pp. 375–380.

[24] Y. C. Pati and P. S. Krishnaprasad, *Rational wavelets in model reduction and system identification*, in Proceedings of the 33rd IEEE Conference on Decision and Control, Vol. 4, Orlando, FL, 1994, pp. 3394–3399.

[25] Y. C. Pati, R. Rezaiifar, P. S. Krishnaprasad, and W. P. Dayawansa, *A fast recursive algorithm for system identification and model reduction using rational wavelets*, in Proceedings of the 27th Annual Asilomar Conference on Signals Systems and Computers, Vol. 4, Pacific Grove, CA, 1993, pp. 35–39.

[26] K. Ramchandran and M. Vetterli, *Best wavelet packet bases in a rate-distortion sense*, IEEE Trans. Image Processing, 2 (1994), pp. 160–175.

[27] R. A. Roberts and C. T. Mullis, *Digital Signal Processing*, Addison–Wesley Series in Electrical Engineering: Digital Signal Processing, Addison–Wesley, Reading, MA, 1987.

[28] W. Rudin, *Real and Complex Analysis*, McGraw–Hill, New York, 1966.

[29] B. Shafai, *Design of state feedback for large-scale multivariable systems*, IEEE Trans. Automat. Control, 32 (1987), pp. 732–734.

[30] O. Szász, *On closed sets of rational functions*, Ann. Mat. Pura Appl. (4), 34 (1953), pp. 195–218.

[31] S. Takenaka, *On the orthogonal functions and a new formula of interpolation*, Japan. J. Math., 2 (1925), pp. 129–145.

[32] L. F. Villemoes, *Nonlinear approximation with Walsh atoms*, in Surface Fitting and Multiresolution Methods, A. Le Méhauté, C. Rabut, and L. L. Schumaker, eds., Vanderbilt University Press, Nashville, TN, 1997, pp. 329–336.

[33] B. Wahlberg, *System identification using Laguerre models*, IEEE Trans. Automat. Control, 36 (1991), pp. 551–562.

[34] B. Wahlberg, *System identification using Kautz models*, IEEE Trans. Automat. Control, 39 (1994), pp. 1276–1281.

[35] B. Wahlberg and E. J. Hannan, *Parametric signal modelling using Laguerre filters*, Ann. Appl. Probab., 3 (1993), pp. 467–496.

[36] B. Wahlberg and P. Mäkilä, *On approximation of stable linear dynamical systems using Laguerre and Kautz functions*, Automatica J. IFAC, 32 (1996), pp. 693–708.

[37] J. L. Walsh, *Interpolation and Approximation by Rational Functions in the Complex Domain*, 3rd ed., Amer. Math. Soc. Colloq. Publ. 20, American Mathematical Society, Providence, RI, 1960.

[38] M. V. Wickerhauser, *Adapted Wavelet Analysis, from Theory to Software*, A. K. Peters, Boston, 1994.

[39] T. Y. Young and W. H. Huggins, *Discrete orthonormal exponentials*, in Proceedings of the National Electronics Conference, Vol. 18, Chicago, IL, 1962, pp. 10–18.

SIAM J. CONTROL OPTIM.
Vol. 38, No. 4, pp. 1033–1049

LINEAR SYSTEMS WITH PRESCRIBED SIMILARITY STRUCTURAL INVARIANTS*

I. BARAGAÑA†, V. FERNÁNDEZ†, AND I. ZABALLA‡

Abstract. The problem of the existence of linear systems $\dot{x}(t) = Ax(t) + Bu(t)$ with prescribed structural invariants for system similarity is studied. Namely, we solve the problem of the existence of such a system with prescribed controllability indices, Hermite indices, and invariant factors when the invariant factors of A (which are also invariants under system similarity) are given.

Key words. system similarity, controllability indices, Hermite indices, Brunovsky indices, invariant factors, majorization

AMS subject classifications. 93B05, 93B10

PII. S0363012998337825

1. Introduction, notation, and preliminary results.

Let us assume that $A \in \mathbb{F}^{n\times n}$ and $B \in \mathbb{F}^{n\times m}$, $\mathbb{F}$ being the field of real or complex numbers. In the study of the structure of linear control systems

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

under systems similarity—(TAT^{-1}, TB), with T invertible—several systems of invariants can be found. In [12] Popov gave a complete system of independent invariants for system similarity but this is not the only complete system of invariants that one can obtain (see, for example, [9, p. 494], [5, p. 191], or [20]). An interesting feature of all these systems is that they are formed by two types of subsystems that, following [11, p. 48], we can call *invariants of structure* and *numerical invariants*. The former ones are nonnegative integers, and the latter ones are real or complex numbers depending on the underlying field where the elements of A and B are.

The most mentioned structural invariants are the controllability and the Hermite indices [9, Chap. 6], [5]. Both come up when searching for a basis of the controllability subspace, i.e., the one generated by the columns of the controllability matrix, $\mathcal{C}(A,B) = [B \;\; AB \;\; \dots \;\; A^{n-1}B]$, in the following table:

$$\begin{array}{ccccc} b_1 & b_2 & b_3 & \dots & b_m \\ Ab_1 & Ab_2 & Ab_3 & \dots & Ab_m \\ A^2b_1 & A_2b_2 & A^2b_3 & \dots & A^2b_m \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ A^{n-1}b_1 & A^{n-1}b_2 & A^{n-1}b_3 & \dots & A^{n-1}b_m, \end{array}$$

*Received by the editors April 22, 1998; accepted for publication (in revised form) April 19, 1999; published electronically April 4, 2000. This research was partially supported by CAICYT, Proyecto de Investigación PB94-1365-CO3-01.

http://www.siam.org/journals/sicon/38-4/33782.html

†Departamento de Ciencias de la Computación e IA, Facultad de Informática, Universidad del País Vasco, Apdo. Correos 649, 20080 Donostia-San Sebastián, Spain (ccpbagai@si.ehu.es, ccpfegov@si.ehu.es).

‡Departamento de Matemática Aplicada y EIO, Facultad de Ciencias, Universidad del País Vasco, Apdo. Correos 644, 48080 Bilbao, Spain (izaballa@picasso.lc.ehu.es).

where $b_i \in \mathbb{F}^{n\times 1}$ is the ith column of B. If $\operatorname{rank}\mathcal{C}(A,B) = r$ and we select by columns (from left to right) the first r linearly independent columns of the table and we write them as

$$b_1, \ldots, A^{h_1-1}b_1, \ldots, b_m, \ldots, A^{h_m-1},$$

then $h_1, h_2, \ldots, h_m$ are the Hermite indices of the system. Actually, as pointed out in [9, p. 476], these indices are the degrees of the polynomials appearing in the diagonal of the Hermite normal form of the right denominator of the transfer $(sI_n - A)^{-1}B$. This is why they were called Hermite indices in [20] (see also [14]). If we proceed similarly but searching by rows from top to bottom and we rearrange the indices in nonincreasing order, then we come up with the controllability indices [4] or input structural indices as they are called in [3, p. 156].

It is also worth noting that system similarity implies similarity of the corresponding state matrices. It is well known (see, for example, [6]) that two square matrices A_1 and A_2 are similar if and only if their corresponding characteristic matrices $sI - A_1$ and $sI - A_2$ are equivalent; i.e., they have the same invariant factors (the invariant factors of A being those of $sI - A$ as a polynomial matrix). Thus the invariant factors of the state matrix are invariants under system similarity. Furthermore, if we call invariant factors of (A, B) or of system (1.1) those of the polynomial matrix $[sI_n - A \quad B]$, then these polynomials are also invariants under system similarity. Notice that the invariant factors of system (1.1) are all equal to 1 if and only if the system is controllable [13].

The well-known Rosenbrock's theorem [13] on eigenstructure assignment under state feedback can be seen as a result of the relationship between the invariant factors of the state matrix A, i.e., those of $sI_n - A$, and the controllability indices of a controllable system (A, B). Similarly, the generalization of Rosenbrock's theorem to noncontrollable systems [17, 18] provides a characterization of the possible controllability indices and invariant factors of system (A, B) for a given matrix A. Following these ideas, the study of the relationship between the controllability and the Hermite indices of a given pair as well as the relationship between the invariant factors of A and the Hermite indices and invariant factors of (A, B) for all possible choices of B was carried out in [20]. In this paper, the four systems of invariants are considered together. Namely, we will deal with the following.

PROBLEM 1. *Let $A \in \mathbb{F}^{n\times n}$ and $\alpha_1 \mid \ldots \mid \alpha_n$ its invariant factors. Let $k_1 \geq \cdots \geq k_m > 0$ be positive integers, $h_1 \geq \cdots \geq h_m \geq 0$ nonnegative integers, and $\gamma_1 \mid \ldots \mid \gamma_n$ monic polynomials. Under what conditions does there exist a matrix $B \in \mathbb{F}^{n\times m}$ such that (A, B) has $k_1, \ldots, k_m$ as controllability indices, $h_1, \ldots, h_m$ as Hermite indices, and $\gamma_1, \ldots, \gamma_n$ as invariant factors?*

Our results will be of an algebraic nature and so we will not impose any restrictions on $\mathbb{F}$ that from now on will be considered arbitrary. We will use Greek letters to denote polynomials, $\alpha \mid \beta$ will mean that α divides β, and $d(\alpha)$ will be the degree of α.

As said before, the controllability indices are invariant under system similarity but in the case when (A, B) is controllable, they form a complete system of invariants for the feedback equivalence. Following Brunovsky [4], two matrix pairs $(A, B), (\hat{A}, \hat{B}) \in \mathbb{F}^{n\times n} \times \mathbb{F}^{n\times m}$ are said to be feedback equivalent if there are nonsingular matrices $P \in \mathbb{F}^{n\times n}$ and $Q \in \mathbb{F}^{m\times m}$ and a matrix $R \in \mathbb{F}^{m\times n}$ such that $(\hat{A}, \hat{B}) = (PAP^{-1} + PBR, PBQ)$. If (A, B) is not completely controllable, a complete system of invariants is given by the controllability indices and the invariant factors of (A, B).

The paper is organized as follows: in the next section we give a solution to the posed problem when the system is controllable, and in the following one when

this condition is not satisfied. Our proofs, although involved, are constructive and decidable.

Throughout this paper we will refer several times to the Brunovsky indices of a pair of matrices. We call Brunovsky indices of (A, B) to the components of the conjugate partition of that of the controllability indices; i.e., if $(k_1, \ldots, k_m)$ is the partition of the controllability indices of (A, B), then, by defining $r_j = \#\{i : k_i \geq j\}$, we have that $r_1, \ldots, r_n$ are its Brunovsky indices (or its r-numbers as they were called in [4]).

In general, if $a = (a_1, \ldots, a_n)$ is a partition of nonnegative integers, we will use $\bar{a} = \overline{(a_1, \ldots, a_n)}$ to mean its conjugate partition. Thus

$$(r_1, \ldots, r_n) = \overline{(k_1, \ldots, k_m)}.$$

We will consider that the components in each partition are arranged in nonincreasing order. Following [8], we say that a is majorized by b, and we write $a \prec b$ if

$$\sum_{j=1}^{k} a_j \leq \sum_{j=1}^{k} b_j, \ 1 \leq k < n,$$

and

$$\sum_{j=1}^{n} a_j = \sum_{j=1}^{n} b_j.$$

It can be proved that if $a \prec b$, then $\bar{b} \prec \bar{a}$ [10].

Before going into a solution of Problem 1, let us notice that this problem is equivalent, in the sense of Proposition 1.1 given below, to the following one.

PROBLEM 2. *Let $(A, B) \in \mathbb{F}^{n\times n} \times \mathbb{F}^{n\times m}$, and let $k_1 \geq \cdots \geq k_m > 0$ and $\gamma_1 \mid \ldots \mid \gamma_n$ be its controllability indices and invariant factors, respectively. Let $\alpha_1 \mid \ldots \mid \alpha_n$ and $h_1 \geq \cdots \geq h_m \geq 0$ be monic polynomials and nonnegative integers. Find necessary and sufficient conditions for the existence of a static state feedback matrix $F \in \mathbb{F}^{m\times n}$ and a nonsingular matrix $Q \in \mathbb{F}^{m\times m}$ such that $A + BF$ has $\alpha_1, \ldots, \alpha_n$ as invariant factors and $(A+BF, BQ)$ has $h_1, \ldots, h_m$ as Hermite indices.*

It should be remarked that a solution of this problem will provide us with a generalization of Rosenbrock's theorem. In fact we are characterizing not only the invariant factors that can be assigned by state feedback but also some additional invariants (the Hermite indices) of the similarity class where the closed loop system $(A + BF, BQ)$ may lie.

As said before the equivalence of Problems 1 and 2 is based on the following proposition.

PROPOSITION 1.1. *Let $A \in \mathbb{F}^{n\times n}$ and $\alpha_1 \mid \ldots \mid \alpha_n$ its invariant factors. Let $(\hat{A}, \hat{B}) \in \mathbb{F}^{n\times n} \times \mathbb{F}^{n\times m}$ with $k_1 \geq \cdots \geq k_m > 0$ and $\gamma_1 \mid \ldots \mid \gamma_n$ as controllability indices and invariant factors, respectively, and let $h_1 \geq \cdots \geq h_m \geq 0$ be nonnegative integers. Then there exists a matrix $B \in \mathbb{F}^{n\times m}$ such that (A, B) is feedback equivalent to $(\hat{A}, \hat{B})$ and has $h_1, \ldots, h_m$ as Hermite indices if and only if there are matrices $F \in \mathbb{F}^{m\times n}$ and $Q \in \mathbb{F}^{m\times m}$, nonsingular, such that $\hat{A} + \hat{B}F$ is similar to A and $(\hat{A}+\hat{B}F, \hat{B}Q)$ has $\gamma_1, \ldots, \gamma_n$ as invariant factors and $h_1, \ldots, h_m$ as Hermite indices.*

Proof. Assume that there is B so that (A, B) is feedback equivalent to $(\hat{A}, \hat{B})$ and has $h_1, \ldots, h_m$ as Hermite indices. Let us show that there are matrices F and

Q, nonsingular, such that $(\hat{A}+\hat{B}F, \hat{B}Q)$ is similar to (A, B). This is enough to prove that $(\hat{A}+\hat{B}F, \hat{B}Q)$ has $h_1, \ldots, h_m$ as Hermite indices and $\gamma_1, \ldots, \gamma_n$ as invariant factors because these two sequences are invariant under system similarity. In fact, since (A, B) and $(\hat{A}, \hat{B})$ are feedback equivalent, there are nonsingular matrices P and T and a matrix R such that

$$A = P(\hat{A}+\hat{B}RP)P^{-1} \quad \text{and} \quad B = P\hat{B}T.$$

So, $F = RP$ and $Q = T$ are the desired matrices.

Conversely, if A is similar to $\hat{A}+\hat{B}F$ for some matrix F, then there is an invertible matrix P such that $A = P(\hat{A}+\hat{B}F)P^{-1}$. Define $B = P\hat{B}Q$. Thus (A, B) and $(\hat{A}+\hat{B}F, \hat{B}Q)$ are system similar matrix pairs and so they have the same invariant factors $\gamma_1, \ldots, \gamma_n$ and the same Hermite indices $h_1, \ldots, h_m$. Furthermore, as $(\hat{A}, \hat{B})$ and $(\hat{A}+\hat{B}F, \hat{B}Q)$ are feedback equivalent, we conclude that (A, B) and $(\hat{A}, \hat{B})$ have the same controllability indices $k_1, \ldots, k_m$. ∎

From now on we will deal with Problem 1. A consequence of the next lemma is that we can substitute A with any matrix in its similarity class.

LEMMA 1.2. *Let $A \in \mathbb{F}^{n\times n}$. Let $k_1 \geq \cdots \geq k_m > 0$ and $h_1 \geq \cdots \geq h_m \geq 0$ be nonnegative integers and $\gamma_1, \ldots, \gamma_n$ monic polynomials. Suppose that $A \overset{s}{\sim} \hat{A}$. Then there exists $B \in \mathbb{F}^{n\times m}$ such that (A, B) has $k_1, \ldots, k_m$ as controllability indices, $h_1, \ldots, h_m$ as Hermite indices, and $\gamma_1, \ldots, \gamma_n$ as invariant factors if and only if there exists $\hat{B} \in \mathbb{F}^{n\times m}$ such that $(\hat{A}, \hat{B})$ has $k_1, \ldots, k_m$ as controllability indices, $h_1, \ldots, h_m$ as Hermite indices, and $\gamma_1, \ldots, \gamma_n$ as invariant factors.*

Proof. The proof is straightforward. ∎

Given a controllable matrix pair $(A, B) \in \mathbb{F}^{n\times n} \times \mathbb{F}^{n\times m}$, we are going to give a canonical form for the similarity of matrix pairs associated with the Hermite indices (see [9]).

LEMMA 1.3. *Let $(A, B) \in \mathbb{F}^{n\times n} \times \mathbb{F}^{n\times m}$ be a controllable pair and let $h_1 \geq \cdots \geq h_p > 0 = h_{p+1} = \cdots = h_m$ be its Hermite indices. Then there exists a nonsingular matrix $P \in \mathbb{F}^{n\times n}$ such that*

$$(PAP^{-1}, PB) = (A_c, B_c),$$

where

$$A_c = (A_{ij}) \begin{matrix} i = 1, \ldots, p, \\ j = 1, \ldots, p, \end{matrix} \qquad B_c = (B_{ij}) \begin{matrix} i = 1, \ldots, p, \\ j = 1, \ldots, m, \end{matrix}$$

are the blocks

$$A_{ii} = \begin{pmatrix} \begin{bmatrix} 0 & 0 & \ldots & 0 & x_{ii0} \\ 1 & 0 & \ldots & 0 & x_{ii1} \\ 0 & 1 & \ldots & 0 & x_{ii2} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \ldots & 1 & x_{iih_i-1} \end{bmatrix} \end{pmatrix} \in \mathbb{F}^{h_i\times h_i}, \ 1 \leq i \leq p,$$

$$A_{ij} = \begin{pmatrix} \begin{bmatrix} 0 & 0 & \ldots & 0 & x_{ji0} \\ 0 & 0 & \ldots & 0 & x_{ji1} \\ 0 & 0 & \ldots & 0 & x_{ji2} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \ldots & 0 & x_{jih_i-1} \end{bmatrix} \end{pmatrix} \in \mathbb{F}^{h_i\times h_j}, \ 1 \leq i < j \leq p,$$

$$A_{ij} = 0 \in \mathbb{F}^{h_i \times h_j},\ 1 \le j < i \le p,$$

$$B_{ii} = \begin{bmatrix} 1 & 0 & \dots & 0 \end{bmatrix}^T \in \mathbb{F}^{h_i \times 1},\ 1 \le i \le p,$$

$$B_{ij} = \begin{bmatrix} 0 & 0 & \dots & 0 \end{bmatrix}^T \in \mathbb{F}^{h_i \times 1},\ 1 \le i, j \le p,\ i \ne j,$$

$$B_{ij} = \begin{bmatrix} x_{ji0} & x_{ji1} & \dots & x_{jih_i-1} \end{bmatrix}^T \in \mathbb{F}^{h_i \times 1},\ 1 \le i \le p,\ p+1 \le j \le m.$$

2. The controllable case. In this section, we will deal with Problem 1 in the controllable case. The following two results provide necessary conditions for the problem to have a solution, i.e., for the existence of a matrix $B \in \mathbb{F}^{n\times m}$ such that (A, B) has prescribed controllability and Hermite indices for a given $A \in \mathbb{F}^{n\times n}$.

LEMMA 2.1 (see [20]). *Let $A \in \mathbb{F}^{n\times n}$, and let $\alpha_1 \mid \dots \mid \alpha_n$ be its invariant factors. Let $h_1 \ge \dots \ge h_m \ge 0$ be nonnegative integers . Then there exists a matrix $B \in \mathbb{F}^{n\times m}$, $m \le n$, such that (A, B) is controllable and has $h_1, \dots, h_m$ as Hermite indices if and only if there are m monic polynomials $\beta_1, \dots, \beta_m$ such that $d(\beta i) = h_i,\ 1 \le i \le m$, and*

$$\alpha_i = 1, \quad 1 \le i \le n - m, \tag{2.1}$$

$$\alpha_{n-m+1} \dots \alpha_{n-m+k} \mid g.c.d.\{\beta_{i_1} \dots \beta_{i_k} : 1 \le i_1 < \dots < i_k \le m\},$$

$$1 \le k \le m - 1, \tag{2.2}$$

$$\alpha_1 \dots \alpha_n = \beta_1 \dots \beta_m. \tag{2.3}$$

LEMMA 2.2 (see [20]). *Let $(A, B) \in \mathbb{F}^{n\times n} \times \mathbb{F}^{n\times m}$ be a controllable matrix pair. Let $k_1 \ge \dots \ge k_m \ge 0$ be its controllability indices, and let $h_1, \dots, h_m$ be its Hermite indices. Then*

$$(k_1, \dots, k_m) \prec (h_1, \dots, h_m). \tag{2.4}$$

The rest of this section is dedicated to show that conditions (2.1)–(2.4) are also sufficient for a matrix $B \in \mathbb{F}^{n\times m}$ to exist so that for a given matrix $A \in \mathbb{F}^{n\times n}$ the pair (A, B) is controllable and has prescribed controllability and Hermite indices. This will be shown as a consequence of several lemmas.

LEMMA 2.3 (see [2]). *Let $(A_1, B_1) \in \mathbb{F}^{n_1\times n_1} \times \mathbb{F}^{n_1\times m_1}$ and $(A_2, B_2) \in \mathbb{F}^{n_2\times n_2} \times \mathbb{F}^{n_2\times m_2}$ be two controllable matrix pairs, and let $r_1 \ge \dots \ge r_{n_1}$ and $m_2 = s_1 \ge \dots \ge s_{n_2}$ be their Brunovsky indices, respectively. Let $k_1 \ge \dots \ge k_{s_1}$ be the controllability indices of (A_2, B_2), and put $n = n_1 + n_2$. Assume that $t_1 = r_1 + m_2$ and that $t_1 \ge \dots \ge t_n$ are nonnegative integers. Then there exist matrices $X \in \mathbb{F}^{n_1\times n_2}$ and $Y \in \mathbb{F}^{n_1\times m_2}$ such that $t_1, \dots, t_n$ are the Brunovsky indices of*

$$\left(\begin{bmatrix} A_1 & X \\ 0 & A_2 \end{bmatrix}, \begin{bmatrix} B_1 & Y \\ 0 & B_2 \end{bmatrix} \right)$$

if and only if the following conditions hold:

$$\sum_{j=1}^{n_1} r_j + \sum_{j=1}^{n_2} s_j = \sum_{j=1}^{n} t_j, \tag{2.5}$$

$$t_j \le r_j + s_1,\ 1 \le j \le n, \tag{2.6}$$

$$t_j \ge s_j,\ 1 \le j \le n, \tag{2.7}$$

$$\sum_{j=1}^{d_p} (t_j - r_j - p) \ge \sum_{j=p+1}^{s_1} k_j,\ 1 \le p \le s_1, \tag{2.8}$$

$$\sum_{j=1}^{i} r_j \le (i-q)(t_q - s_q) + \sum_{j=1}^{q} (t_j - s_j), \quad 1 \le q \le k_1; \quad q+1 \le i \le n, \tag{2.9}$$

where $d_p = \max\{j : t_j - r_j \ge p\}$, $1 \le p \le s_1$.

Remark. Notice that when $s_1 = \cdots = s_{k_1}$, (2.5)–(2.8) imply $t_j = r_j + s_1$, $1 \le j \le k_1$, and (2.9) holds.

LEMMA 2.4 (see [1]). *Let* $(A, B) \in \mathbb{F}^{n\times n} \times \mathbb{F}^{n\times m}$ *be a controllable matrix pair, and let* $r_1 \ge \cdots \ge r_n$ *be its Brunovsky indices. Let* $t_1 = r_1 + s \ge \cdots \ge t_n$ *be nonnegative integers. Then there exists a matrix* $C \in \mathbb{F}^{n\times s}$ *such that* $t_1, \ldots, t_n$ *are the Brunovsky indices of* $(A, [B\ C])$ *if and only if the following conditions hold:*

$$\sum_{j=1}^{n} t_j = \sum_{j=1}^{n} r_j, \tag{2.10}$$

$$t_j \le r_j + s,\ 1 \le j \le n, \tag{2.11}$$

$$\sum_{j=1}^{d_p} (t_j - r_j - p) \ge 0,\ 1 \le p \le s, \tag{2.12}$$

where $d_p = \max\{j : t_j - r_j \ge p\}$, $1 \le p \le s$.

LEMMA 2.5. *Let* $k_1 \ge \cdots \ge k_m > 0$ *and* $h_1 \ge \cdots \ge h_m \ge 0$ *be nonnegative integers, and let* $(v_1, v_2, \ldots, v_n) = \overline{(h_1, \ldots, h_m)}$, $(t_1, \ldots, t_n) = \overline{(k_1, \ldots, k_m)}$ *and* $(v'_1, \ldots, v'_n) = \overline{(h_1, \ldots, h_{m-1})}$. *Assume that* (2.4) *holds and let*

$$r_i = m - 1,\ 1 \le i \le k_m,$$

and

$$r_i = \min\left\{ m - 1, t_i + \sum_{j=1}^{i-1} (t_j - r_j) - h_m \right\},\ k_m < i \le n.$$

Then the following conditions are satisfied:

$$r_1 \geq \cdots \geq r_n \geq 0, \tag{2.13}$$

$$(v_1', \ldots, v_n') \prec (r_1, \ldots, r_n), \tag{2.14}$$

$$t_j \leq r_j + 1,\ 1 \leq j \leq n, \tag{2.15}$$

$$\sum_{j=1}^{d_1} (t_j - r_j - 1) = 0, \tag{2.16}$$

where $d_1 = \max\{j : t_j - r_j \geq 1\}$.

Proof. We can assume without loss of generality that $k_m < n$.

Notice first that $t_i \leq m-1$, $k_m < i \leq n$, and $\sum_{j=1}^{k_m}(t_j - r_j) - h_m = k_m - h_m \geq 0$. Moreover, as $r_i \leq t_i + \sum_{j=1}^{i-1}(t_j - r_j) - h_m$, $k_m < i \leq n$, we have that $\sum_{j=1}^{i-1}(t_j - r_j) - h_m \geq 0$ for $k_m + 1 < i \leq n$. Hence,

$$r_i \geq t_i,\ k_m < i \leq n.$$

Therefore, $d_1 = k_m$ and (2.15)–(2.16) hold.

We will prove next that for some $i \in \{k_m + 1, \ldots, n\}$,

$$r_i = t_i + \sum_{j=1}^{i-1}(t_j - r_j) - h_m.$$

In fact, assume that

$$r_i \neq t_i + \sum_{j=1}^{i-1}(t_j - r_j) - h_m,\ k_m + 1 \leq i \leq n;$$

then

$$r_n = m - 1 < t_n + \sum_{j=1}^{n-1}(t_j - r_j) - h_m = \sum_{j=1}^{m-1} h_j - (n-1)(m-1).$$

So

$$n(m-1) < \sum_{j=1}^{m-1} h_j \leq n(m-1),$$

which is a contradiction.

Let $g = \min\{i \in \{k_m + 1 \ldots, n\} : r_i = t_i + \sum_{j=1}^{i-1}(t_j - r_j) - h_m\}$. Then

$$r_i = m - 1,\ 1 \leq i \leq g - 1,$$

$$r_g = \sum_{j=1}^{g} t_j - (g-1)(m-1) - h_m,$$

and

$$r_i = t_i,\ g < i \leq n.$$

Therefore,

$$r_{g-1} = m - 1 \geq r_g \geq t_g \geq t_{g+1} = r_{g+1},$$

and (2.13) follows.

On the other hand,

$$\begin{array}{l} v'_i = v_i - 1,\ 1 \leq i \leq h_m, \\ v'_i = v_i,\ h_m < i \leq n, \end{array}$$

and, bearing in mind that $h_m \leq k_m < g$, we have that

$$\sum_{j=1}^{i} r_j = (m-1)i \geq \sum_{j=1}^{i} v'_j,\ 1 \leq i < g,$$

$$\sum_{j=1}^{i} r_j = \sum_{j=1}^{i} t_j - h_m \geq \sum_{j=1}^{i} v_j - h_m = \sum_{j=1}^{i} v'_j,\ g \leq i \leq n,$$

and

$$\sum_{j=1}^{n} r_j = \sum_{j=1}^{n} t_j - h_m = \sum_{j=1}^{n} v_j - h_m = \sum_{j=1}^{n} v'_j,$$

from which we obtain (2.14). □

LEMMA 2.6. *Let $h_1 \geq \cdots \geq h_p > 0$ positive integers, and let*

$$(A_{ii}, B_{ii}) = \left(\begin{bmatrix} 0 & 0 & \dots & 0 & x_{ii0} \\ 1 & 0 & \dots & 0 & x_{ii1} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \dots & 1 & x_{iih_i-1} \end{bmatrix}, \begin{bmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix} \right) \in \mathbb{F}^{h_i \times hi} \times \mathbb{F}^{h_i \times 1},\ i = 1, \dots, p.$$

Let $k_1 \geq \cdots \geq k_m > 0$ be positive integers. If condition (2.4) *is satisfied, then there exist matrices $X_{ij} \in \mathbb{F}^{h_i \times h_j}$, $Y_{ij} \in \mathbb{F}^{h_i \times 1}$, $1 \leq i \leq p-1$, $i+1 \leq j \leq p$, and $Y_{ip+1} \in \mathbb{F}^{h_i \times (m-p)}$, $1 \leq i \leq p$, such that*

$$\left(\begin{bmatrix} A_{11} & X_{12} & \dots & X_{1p} \\ 0 & A_{22} & \dots & X_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & A_{pp} \end{bmatrix}, \begin{bmatrix} B_{11} & Y_{12} & \dots & Y_{1p} & Y_{1p+1} \\ 0 & B_{22} & \dots & Y_{2p} & Y_{2p+1} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \dots & B_{pp} & Y_{pp+1} \end{bmatrix} \right)$$

has $k_1, k_2, \dots, k_m$ as controllability indices.

Proof. If $m = 1$, then $k_1 = h_1$, and there is nothing to prove.

Assume now that condition (2.4) is sufficient for $m-1$, and let us show that it is sufficient for m.

From (2.4) we have that $m \geq p$ and $k_m \geq h_m$.

Let $(v_1, v_2, \ldots, v_n) = \overline{(h_1, \ldots, h_m)}$, $(t_1, \ldots, t_n) = \overline{(k_1, \ldots, k_m)}$, and $(v'_1, \ldots, v'_n) = \overline{(h_1, \ldots, h_{m-1})}$. From Lemma 2.5 there exist nonnegative integers $r_1, \ldots, r_n$ satisfying conditions (2.13)–(2.16) and $r_1 = m - 1$.

Let $(k'_1, \ldots, k'_{m-1}) = \overline{(r_1, \ldots, r_n)}$. Then $k'_{m-1} > 0$, and condition (2.14) is equivalent to

$$(k'_1, \ldots, k'_{m-1}) \prec (h_1, \ldots, h_{m-1}). \tag{2.17}$$

Moreover, from (2.4) and (2.14), we obtain

$$\sum_{j=1}^{n} r_j + h_m = \sum_{j=1}^{n} t_j, \tag{2.18}$$

and, since $k_1 \geq k_m \geq h_m$,

$$t_j \geq 1,\ 1 \leq j \leq h_m. \tag{2.19}$$

If $m = p$, from (2.17) and the induction hypothesis, there are matrices $X_{ij} \in \mathbb{F}^{h_i \times h_j}$, $Y_{ij} \in \mathbb{F}^{h_i \times 1}$, $1 \leq i \leq m-2$, $i+1 \leq j \leq m-1$, such that the matrix pair

$$(A', B') = \left(\begin{bmatrix} A_{11} & X_{12} & \ldots & X_{1m-1} \\ 0 & A_{22} & \ldots & X_{2m-1} \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & A_{m-1m-1} \end{bmatrix}, \begin{bmatrix} B_{11} & Y_{12} & \ldots & Y_{1m-1} \\ 0 & B_{22} & \ldots & Y_{2m-1} \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & B_{m-1m-1} \end{bmatrix} \right)$$

has $k'_1, \ldots, k'_{m-1}$ as controllability indices (and therefore $r_1, \ldots, r_n$ as Brunovsky indices). The Brunovsky indices of (A_{mm}, B_{mm}) are $1, \ldots, 1$ (h_m times).

From (2.15), (2.16), (2.18), and (2.19), by Lemma 2.3, there exist matrices $X \in \mathbb{F}^{(n-h_m) \times h_m}$, $Y \in \mathbb{F}^{(n-h_m) \times 1}$ such that the matrix pair

$$\left(\begin{bmatrix} A' & X \\ 0 & A_{mm} \end{bmatrix}, \begin{bmatrix} B' & Y \\ 0 & B_{mm} \end{bmatrix} \right)$$

has $k_1, \ldots, k_m$ as controllability indices.

If $m > p$, from (2.17) and the induction hypothesis, there are matrices $X_{ij} \in \mathbb{F}^{h_i \times h_j}$, $Y_{ij} \in \mathbb{F}^{h_i \times 1}$, $1 \leq i \leq p-1$, $i+1 \leq j \leq p$, and $Y_{ip+1} \in \mathbb{F}^{h_i \times (m-1-p)}$, $1 \leq i \leq p$, such that

$$(A', B') = \left(\begin{bmatrix} A_{11} & X_{12} & \ldots & X_{1p} \\ 0 & A_{22} & \ldots & X_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & A_{pp} \end{bmatrix}, \begin{bmatrix} B_{11} & Y_{12} & \ldots & Y_{1p} & Y_{1p+1} \\ 0 & B_{22} & \ldots & Y_{2p} & Y_{2p+1} \\ \vdots & \vdots & \ddots & \vdots & \\ 0 & 0 & \ldots & B_{pp} & Y_{pp+1} \end{bmatrix} \right)$$

has $r_1, \ldots, r_n$ as Brunovsky indices.

As $m > p$, it follows that $h_m = 0$. From (2.15), (2.16), and (2.18), by Lemma 2.4, there exists $y_{p+1} \in \mathbb{F}^{n \times 1}$ such that the matrix pair

$$\left(A', \begin{bmatrix} B' & y_{p+1} \end{bmatrix} \right)$$

has $k_1, \ldots, k_m$ as controllability indices. □

LEMMA 2.7. *Let $(A_i, B_i), (\hat{A}_i, \hat{B}_i) \in \mathbb{F}^{n_i \times n_i} \times \mathbb{F}^{n_i \times m_i}$, $i = 1, 2$, and assume that (A_i, B_i) is feedback equivalent to $(\hat{A}_i, \hat{B}_i)$, $i = 1, 2$. Let $m \geq m_1 + m_2$ and $k_1 \geq$*

$\cdots \geq k_m \geq 0$ *be nonnegative integers. Then there exist matrices* $X \in \mathbb{F}^{n_1 \times n_2}, Y_1 \in \mathbb{F}^{n_1 \times m_2}$, $Y_2 \in \mathbb{F}^{n_1 \times (m-m_1-m_2)}$, $Y_3 \in \mathbb{F}^{n_2 \times (m-m_1-m_2)}$ *such that*

$$(A, B) = \left(\begin{bmatrix} A_1 & X \\ 0 & A_2 \end{bmatrix}, \begin{bmatrix} B_1 & Y_1 & Y_2 \\ 0 & B_2 & Y_3 \end{bmatrix} \right)$$

has $k_1, \ldots, k_m$ *as controllability indices if and only if there exist matrices* $\hat{X} \in \mathbb{F}^{n_1 \times n_2}$, $\hat{Y}_1 \in \mathbb{F}^{n_1 \times m_2}$, $\hat{Y}_2 \in \mathbb{F}^{n_1 \times (m-m_1-m_2)}$, $\hat{Y}_3 \in \mathbb{F}^{n_2 \times (m-m_1-m_2)}$ *such that*

$$(\hat{A}, \hat{B}) = \left(\begin{bmatrix} \hat{A}_1 & \hat{X} \\ 0 & \hat{A}_2 \end{bmatrix}, \begin{bmatrix} \hat{B}_1 & \hat{Y}_1 & \hat{Y}_2 \\ 0 & \hat{B}_2 & \hat{Y}_3 \end{bmatrix} \right)$$

has $k_1, \ldots, k_m$ *as controllability indices.*

Proof. Assume that

$$(A, B) = \left(\begin{bmatrix} A_1 & X \\ 0 & A_2 \end{bmatrix}, \begin{bmatrix} B_1 & Y_1 & Y_2 \\ 0 & B_2 & Y_3 \end{bmatrix} \right)$$

has $k_1, \ldots, k_m$ as controllability indices.

There are nonsingular matrices $P_i \in \mathbb{F}^{n_i \times n_i}$ and $Q_i \in \mathbb{F}^{m_i \times m_i}$ and matrices $R_i \in \mathbb{F}^{m_i \times n_i}$, $i = 1, 2$, such that

$$(\hat{A}_i, \hat{B}_i) = (P_i A_i P_i^{-1} + P_i B_i R_i, P_i B_i Q_i), \quad i = 1, 2.$$

Put

$$P := \begin{bmatrix} P_1 & 0 \\ 0 & P_2 \end{bmatrix}, \quad Q := \begin{bmatrix} Q_1 & 0 & 0 \\ 0 & Q_2 & 0 \\ 0 & 0 & I_{m-m_1-m_2} \end{bmatrix}, \quad \text{and} \quad R := \begin{bmatrix} R_1 & 0 \\ 0 & R_2 \\ 0 & 0 \end{bmatrix}.$$

Then

$$(PAP^{-1} + PBR, PBQ) = \left(\begin{bmatrix} \hat{A}_1 & \hat{X} \\ 0 & \hat{A}_2 \end{bmatrix}, \begin{bmatrix} \hat{B}_1 & \hat{Y}_1 & \hat{Y}_2 \\ 0 & \hat{B}_2 & \hat{Y}_3 \end{bmatrix} \right),$$

where $\hat{X} = P_1 X P_2^{-1} + P_1 Y_1 R_2$, $\hat{Y}_1 = P_1 Y_1 Q_2$, $\hat{Y}_2 = P_1 Y_2$, and $\hat{Y}_3 = P_2 Y_3$. ∎

LEMMA 2.8. *Let* $(A_1, B_1) \in \mathbb{F}^{n_1 \times n_1} \times \mathbb{F}^{n_1 \times m_1}$ *be a controllable matrix pair,* $A_2 \in \mathbb{F}^{n_2 \times n_2}$ *and* $A_3 \in \mathbb{F}^{n_1 \times n_2}$. *Then there exist* $R \in \mathbb{F}^{m_1 \times n_2}$ *and* $L \in \mathbb{F}^{n_1 \times n_2}$ *such that*

$$A_1 L + B_1 R - L A_2 = A_3.$$

Proof. Let $k_1 \geq k_2 \geq \cdots \geq k_r$ be the controllability indices of the matrix pair (A_1, B_1), then [16] there exist nonsingular matrices $P \in \mathbb{F}^{n_1 \times n_1}$ and $Q \in \mathbb{F}^{m_1 \times m_1}$ and a matrix $F \in \mathbb{F}^{m_1 \times n_1}$ such that

$$P A_1 P^{-1} + P B_1 F = \hat{A}_1, \quad P B_1 Q = \hat{B}_1,$$

where

$$\hat{A}_1 = \operatorname{diag}(A_{11}, A_{12}, \ldots, A_{1r}), \quad \hat{B}_1 = \begin{bmatrix} B_{11} \\ B_{12} \\ \vdots \\ B_{1r} \end{bmatrix}$$

and

$$A_{1i} = \begin{bmatrix} 0 & I_{k_i-1} \\ 0 & 0 \end{bmatrix} \in \mathbb{F}^{k_i \times k_i},\ B_{1i} = \begin{bmatrix} 0 \\ e_i \end{bmatrix} \in \mathbb{F}^{k_i \times m_1},\ i = 1, \dots, r,$$

where e_i is the ith row of the identity matrix I_{m_1}.

Let us put

$$\hat{A}_3 = PA_3 = \begin{bmatrix} A_{31} \\ A_{32} \\ \vdots \\ A_{3r} \end{bmatrix},\quad A_{3i} \in \mathbb{F}^{k_i \times n_2},\ i = 1, \dots, r.$$

By Lemma 2.10 of [16], for $i = 1, \dots, r$ there are matrices $L_i \in \mathbb{F}^{k_i \times n_2}$, $d_i \in \mathbb{F}^{1 \times n_2}$ such that $-A_{1i}L_i + A_{3i} + L_iA_2 = \left[\begin{smallmatrix} 0 \\ d_i \end{smallmatrix}\right] \in \mathbb{F}^{k_i \times n_2}$.

Let

$$\hat{L} = \begin{bmatrix} L_1 \\ L_2 \\ \vdots \\ L_r \end{bmatrix}.$$

As all the elements of B_{1i} are zero except the one in position (k_i, i), it is easy to prove that there exists a matrix $\hat{R} \in \mathbb{F}^{m_1 \times n_2}$ such that $-\hat{A}_1\hat{L} + \hat{A}_3 + \hat{L}A_2 + \hat{B}_1\hat{R} = 0$.

Therefore,

$$-PA_1P^{-1}\hat{L} - PB_1F\hat{L} + PA_3 + \hat{L}A_2 + PB_1Q\hat{R} = 0,$$

and if we put $L = P^{-1}\hat{L}$ and $R = F\hat{L} - Q\hat{R}$, the lemma follows. □

LEMMA 2.9. *Let*

$$(A, B) = \left(\begin{bmatrix} A_{11} & A_{12} & \dots & A_{1p} \\ 0 & A_{22} & \dots & A_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & A_{pp} \end{bmatrix}, \begin{bmatrix} B_{11} & 0 & \dots & 0 \\ 0 & B_{22} & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & B_{pp} \end{bmatrix} \right),$$

with $(A_{ii}, B_{ii}) \in \mathbb{F}^{n_i \times n_i} \times \mathbb{F}^{n_i \times m_i}$ *controllable,* $1 \le i \le p-1$ *and* $(A_{pp}, B_{pp}) \in \mathbb{F}^{n_p \times n_p} \times \mathbb{F}^{n_p \times m_p}$. *Let* $m \ge \sum_{i=1}^{p} m_i$ *and* $k_1 \ge \cdots \ge k_m \ge 0$ *be nonnegative integers. Then there exist matrices* $Z_{ij} \in \mathbb{F}^{n_i \times m_j}$, $1 \le i \le p-1$, $i+1 \le j \le p$, *and* $Z_{ip+1} \in \mathbb{F}^{n_i \times (m - \sum_{i=1}^{p} m_i)}$, $1 \le i \le p$, *such that*

$$(A^{(p)}, B^{(p)}) = \left(\begin{bmatrix} A_{11} & A_{12} & \dots & A_{1p} \\ 0 & A_{22} & \dots & A_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & A_{pp} \end{bmatrix}, \begin{bmatrix} B_{11} & Z_{12} & \dots & Z_{1p} & Z_{1p+1} \\ 0 & B_{22} & \dots & Z_{2p} & Z_{2p+1} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \dots & B_{pp} & Z_{pp+1} \end{bmatrix} \right)$$

has $k_1, \dots, k_m$ *as controllability indices if and only if there exist matrices* $X_{ij} \in \mathbb{F}^{n_i \times n_j}, Y_{ij} \in \mathbb{F}^{n_i \times m_j}$, $1 \le i \le p-1$, $i+1 \le j \le p$, *and* $Y_{ip+1} \in \mathbb{F}^{n_i \times (m - \sum_{i=1}^{p} m_i)}$, $1 \le i \le p$, *such that*

$$(\hat{A}^{(p)}, \hat{B}^{(p)}) = \left(\begin{bmatrix} A_{11} & X_{12} & \dots & X_{1p} \\ 0 & A_{22} & \dots & X_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & A_{pp} \end{bmatrix}, \begin{bmatrix} B_{11} & Y_{12} & \dots & Y_{1p} & Y_{1p+1} \\ 0 & B_{22} & \dots & Y_{2p} & Y_{2p+1} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \dots & B_{pp} & Y_{pp+1} \end{bmatrix} \right)$$

has $k_1, \dots, k_m$ *as controllability indices.*

Proof. We will show by induction on p that if there exist matrices $X_{ij} \in \mathbb{F}^{n_i \times n_j}$, $Y_{ij} \in \mathbb{F}^{n_i \times m_j}$, $1 \leq i \leq p-1$, $i+1 \leq j \leq p$, and $Y_{ip+1} \in \mathbb{F}^{n_i \times (m - \sum_{i=1}^{p} m_i)}$, $1 \leq i \leq p$, such that $(\hat{A}^{(p)}, \hat{B}^{(p)})$ has $k_1, \ldots, k_m$ as controllability indices, then there exist matrices $Z_{ij} \in \mathbb{F}^{n_i \times m_j}$, $1 \leq i \leq p-1$, $i+1 \leq j \leq p$, and $Z_{ip+1} \in \mathbb{F}^{n_i \times (m - \sum_{i=1}^{p} m_i)}$, $1 \leq i \leq p$, such that $(A^{(p)}, B^{(p)})$ has $k_1, \ldots, k_m$ as controllability indices.

If $p = 1$, the proposition is trivial.

We will assume that it is true up to $p-1$, and let us prove that it also holds for p.

Let us suppose that there are matrices $X_{ij} \in \mathbb{F}^{n_i \times n_j}$, $Y_{ij} \in \mathbb{F}^{n_i \times m_j}$, $1 \leq i \leq p-1$, $i+1 \leq j \leq p$, and $Y_{ip+1} \in \mathbb{F}^{n_i \times (m - \sum_{i=1}^{p} m_i)}$, $1 \leq i \leq p$, such that $(\hat{A}^{(p)}, \hat{B}^{(p)})$ has $k_1, \ldots, k_m$ as controllability indices. By the induction hypothesis there exist matrices $Z_{ij} \in \mathbb{F}^{n_i \times m_j}$, $1 \leq i \leq p-1$, $i+1 \leq j \leq p-1$, such that

$$(A^{(p-1)}, B^{(p-1)}) = \left(\begin{bmatrix} A_{11} & A_{12} & \ldots & A_{1p-1} \\ 0 & A_{22} & \ldots & A_{2p-1} \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & A_{p-1p-1} \end{bmatrix}, \begin{bmatrix} B_{11} & Z_{12} & \ldots & Z_{1p-1} \\ 0 & B_{22} & \ldots & Z_{2p-1} \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & B_{p-1p-1} \end{bmatrix} \right)$$

has the same controllability indices as

$$(\hat{A}^{(p-1)}, \hat{B}^{(p-1)}) = \left(\begin{bmatrix} A_{11} & X_{12} & \ldots & X_{1p-1} \\ 0 & A_{22} & \ldots & X_{2p-1} \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & A_{p-1p-1} \end{bmatrix}, \begin{bmatrix} B_{11} & Y_{12} & \ldots & Y_{1p-1} \\ 0 & B_{22} & \ldots & Y_{2p-1} \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & B_{p-1p-1} \end{bmatrix} \right).$$

Furthermore, as (A_{ii}, B_{ii}) is controllable, $1 \leq i \leq p-1$; then the pairs $(A^{(p-1)}, B^{(p-1)})$ and $(\hat{A}^{(p-1)}, \hat{B}^{(p-1)})$ are controllable, and so they are feedback equivalent.

Let us assume that

$$X_p = \begin{bmatrix} X_{1p} \\ \vdots \\ X_{p-1p} \end{bmatrix} \in \mathbb{F}^{\sum_{i=1}^{p-1} n_i \times n_p}, \qquad Y_p = \begin{bmatrix} Y_{1p} \\ \vdots \\ Y_{p-1p} \end{bmatrix} \in \mathbb{F}^{\sum_{i=1}^{p-1} n_i \times m_p},$$

and

$$Y_{p+1} = \begin{bmatrix} Y_{1p+1} \\ \vdots \\ Y_{p-1p+1} \end{bmatrix} \in \mathbb{F}^{\sum_{i=1}^{p-1} n_i \times (m - \sum_{i=1}^{p} m_i)}$$

are matrices such that

$$(\hat{A}^{(p)}, \hat{B}^{(p)}) = \left(\begin{bmatrix} \hat{A}^{(p-1)} & X_p \\ 0 & A_{pp} \end{bmatrix}, \begin{bmatrix} \hat{B}^{(p-1)} & Y_p & Y_{p+1} \\ 0 & B_{pp} & Y_{pp+1} \end{bmatrix} \right)$$

has $k_1, \ldots, k_m$ as controllability indices.

By Lemma 2.7 there exist $\hat{X}_p, \hat{Y}_p, \hat{Y}_{p+1}, \hat{Y}_{pp+1}$ such that

$$\left(\begin{bmatrix} A^{(p-1)} & \hat{X}_p \\ 0 & A_{pp} \end{bmatrix}, \begin{bmatrix} B^{(p-1)} & \hat{Y}_p & \hat{Y}_{p+1} \\ 0 & B_{pp} & \hat{Y}_{pp+1} \end{bmatrix} \right)$$

is feedback equivalent to $(\hat{A}^{(p)}, \hat{B}^{(p)})$.

Because $(A^{(p-1)}, B^{(p-1)})$ is controllable, by Lemma 2.8 there exist matrices $L \in \mathbb{F}^{\sum_{i=1}^{p-1} n_i \times n_p}$ and $R \in \mathbb{F}^{\sum_{i=1}^{p-1} m_i \times n_p}$ such that

$$A^{(p-1)}L - LA_{pp} + B^{(p-1)}R = \hat{X}_p - A_p,$$

where

$$A_p = \begin{bmatrix} A_{1p} \\ \vdots \\ A_{p-1p} \end{bmatrix} \in \mathbb{F}^{\sum_{i=1}^{p-1} n_i \times n_p}.$$

Put

$$T = \begin{bmatrix} I & L \\ 0 & I \end{bmatrix}, \quad F = \begin{bmatrix} 0 & -R \\ 0 & 0 \\ 0 & 0 \end{bmatrix}.$$

Then

$$T \begin{bmatrix} A^{(p-1)} & \hat{X}_p \\ 0 & A_{pp} \end{bmatrix} T^{-1} + T \begin{bmatrix} B^{(p-1)} & \hat{Y}_p & \hat{Y}_{p+1} \\ 0 & B_{pp} & \hat{Y}_{pp+1} \end{bmatrix} F = \begin{bmatrix} A^{(p-1)} & A_p \\ 0 & A_{pp} \end{bmatrix} = A^{(p)}$$

and

$$T \begin{bmatrix} B^{(p-1)} & \hat{Y}_p & \hat{Y}_{p+1} \\ 0 & B_{pp} & \hat{Y}_{pp+1} \end{bmatrix} = \begin{bmatrix} B^{(p-1)} & \hat{Y}_p + LB_{pp} & \hat{Y}_{p+1} + L\hat{Y}_{pp+1} \\ 0 & B_{pp} & \hat{Y}_{pp+1} \end{bmatrix}.$$

Write $Z_{pp+1} = \hat{Y}_{pp+1}$ and

$$\hat{Y}_p + LB_{pp} = \begin{bmatrix} Z_{1p} \\ \vdots \\ Z_{p-1p} \end{bmatrix}, \quad \hat{Y}_{p+1} + L\hat{Y}_{pp+1} = \begin{bmatrix} Z_{1p+1} \\ \vdots \\ Z_{p-1p+1} \end{bmatrix},$$

where $Z_{ip} \in \mathbb{F}^{n_i \times m_p}, 1 \le i \le p-1$, and $Z_{ip+1} \in \mathbb{F}^{n_i \times (m - \sum_{i=1}^{p} m_i)}, 1 \le i \le p-1$. Put

$$B^{(p)} = \begin{bmatrix} B^{(p-1)} & \hat{Y}_p + LB_{pp} & \hat{Y}_{p+1} + L\hat{Y}_{pp+1} \\ 0 & B_{pp} & \hat{Y}_{pp+1} \end{bmatrix}.$$

Then $(A^{(p)}, B^{(p)})$ is feedback equivalent to $(\hat{A}^{(p)}, \hat{B}^{(p)})$, and therefore it has $k_1, \ldots, k_m$ as controllability indices. □

Finally, the following lemma, whose proof is straightforward, allows us to perform some transformations on matrix B without altering the Hermite indices of pair (A, B).

LEMMA 2.10. *Let $A \in \mathbb{F}^{n \times n}$ and $b_i \in \mathbb{F}^{n \times 1}, 1 \le i \le m$. Let $h_1, \ldots, h_m$ be the Hermite indices of the pair (A, B), where $B = [b_1, \ldots, b_m]$. Let t be a nonnegative integer and $l \in \{2, \ldots, m\}$. Put $b'_l = b_l + \sum_{j=0}^{t} \sum_{i=1}^{l-1} c_{ij} A^j b_i$, where $c_{ij} \in \mathbb{F}$ are arbitrary. Then, $h_1, \ldots, h_m$ are the Hermite indices of the matrix pair (A, B'), where $B' = [b_1, \ldots, b'_l, \ldots, b_m]$.*

As a consequence of this result, we have that the matrix pair

$$(A_c, B'_c) = \left(\begin{bmatrix} A_{11} & A_{12} & \ldots & A_{1p} \\ 0 & A_{22} & \ldots & A_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & A_{pp} \end{bmatrix}, \begin{bmatrix} B_{11} & Z_{12} & \ldots & Z_{1p} & Z_{1p+1} \\ 0 & B_{22} & \ldots & Z_{2p} & Z_{2p+1} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \ldots & B_{pp} & Z_{pp+1} \end{bmatrix} \right)$$

for arbitrary $Z_{ij} \in \mathbb{F}^{h_i \times 1}$, $1 \le i \le p-1$, $i+1 \le j \le p$, and $Z_{ip+1} \in \mathbb{F}^{h_i \times (m-p)}$, $1 \le i \le p$, has $h_1, \dots, h_m$ as Hermite indices.

Now we can prove our main result.

THEOREM 2.11. *Let $A \in \mathbb{F}^{n\times n}$, and let $\alpha_1 \mid \dots \mid \alpha_n$ be its invariant factors. Let $k_1 \ge \cdots \ge k_m > 0$ and $h_1 \ge \cdots \ge h_m \ge 0$ be nonnegative integers. Then there exists a matrix $B \in \mathbb{F}^{n\times m}$ such that (A, B) is controllable and has $k_1, \dots, k_m$ as controllability indices and $h_1, \dots, h_m$ as Hermite indices if and only if there are m monic polynomials $\beta_1, \dots, \beta_m$ such that $d(\beta_i) = h_i$, $1 \le i \le m$, and conditions* (2.1), (2.2), (2.3), *and* (2.4) *are satisfied.*

Proof. The necessity is a direct consequence of Lemmas 2.1 and 2.2.

Assume now that a matrix $A \in \mathbb{F}^{n\times n}$ is given with $\alpha_1 \mid \dots \mid \alpha_n$ as invariant factors. By Lemma 2.1 conditions (2.1)–(2.3) are sufficient for the existence of a matrix $B \in \mathbb{F}^{n\times m}$ such that (A, B) has $h_1, \dots, h_m$ as Hermite indices. By Lemma 1.3 (A, B) is similar to (A_c, B_c), where this pair has the form exhibited in that lemma. From Lemmas 2.6, 2.9, and 2.10 we have that if condition (2.4) is fulfilled, then there are matrices $Z_{ij} \in \mathbb{F}^{h_i \times 1}$, $1 \le i \le p-1$, $i+1 \le j \le p$, and $Z_{ip+1} \in \mathbb{F}^{h_i \times (m-p)}$, $1 \le i \le p$, such that

$$(A_c, B_c') = \left(\begin{bmatrix} A_{11} & A_{12} & \dots & A_{1p} \\ 0 & A_{22} & \dots & A_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & A_{pp} \end{bmatrix}, \begin{bmatrix} B_{11} & Z_{12} & \dots & Z_{1p} & Z_{1p+1} \\ 0 & B_{22} & \dots & Z_{2p} & Z_{2p+1} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \dots & B_{pp} & Z_{pp+1} \end{bmatrix} \right)$$

has $k_1, \dots, k_m$ as controllability indices and $h_1, \dots, h_m$ as Hermite indices. As A and A_c are similar matrices, there is a nonsingular matrix $P \in \mathbb{F}^{n\times n}$ such that $A = PA_cP^{-1}$. Put $B = PB_c'$. Then (A, B) has $k_1, \dots, k_m$ and $h_1, \dots, h_m$ as controllability and Hermite indices, respectively, and the theorem follows. ∎

It is worth noting that if $\mathbb{F}$ is algebraically closed in [20], it has been shown that conditions (2.1), (2.2), and (2.3) are equivalent to (2.1) and

$$(h_1, \dots, h_m) \prec (d(\alpha_n), \dots, d(\alpha_{n-m+1})). \tag{2.20}$$

Thus, in this case, we have the following consequence of Theorem 2.11.

COROLLARY 2.12. *Let $\mathbb{F}$ be an algebraically closed field. Let $A \in \mathbb{F}^{n\times n}$, and let $\alpha_1 \mid \dots \mid \alpha_n$ be its invariant factors. Let $k_1 \ge \cdots \ge k_m > 0$ and $h_1 \ge \cdots \ge h_m \ge 0$ be nonnegative integers. Then there exists a matrix $B \in \mathbb{F}^{n\times m}$ such that (A, B) is controllable and has $k_1, \dots, k_m$ as controllability indices and $h_1, \dots, h_m$ as Hermite indices if and only if conditions* (2.1), (2.20), *and* (2.4) *are satisfied.*

3. The noncontrollable case. Now we will generalize Theorem 2.11 to the noncontrollable case. It is well known (see, for example, [9, p. 361]) that if (A, B) is not completely controllable, then there is a nonsingular matrix P such that

$$PAP^{-1} = \begin{bmatrix} A_1 & A_2 \\ 0 & A_3 \end{bmatrix}, \quad PB = \begin{bmatrix} B_1 \\ 0 \end{bmatrix}, \tag{3.1}$$

where $(A_1, B_1) \in \mathbb{F}^{r\times r}$ is controllable and $r = \operatorname{rank} C(A, B)$. Furthermore, the invariant factors of (A, B) different from 1 and those of A_3 coincide [17].

As the Hermite indices and the controllability indices are invariant under similarity tranformations, it follows that the Hermite indices and the controllability indices of (A, B) are those of the matrix pair (A_1, B_1) in the decomposition of (A, B) as in (3.1).

The generalization of Lemma 2.1 to the noncontrollable case is given by the following lemma.

LEMMA 3.1 (see [20]). *Let $A \in \mathbb{F}^{n\times n}$, and let $\alpha_1 \mid \ldots \mid \alpha_n$ be its invariant factors. Let $h_1, \ldots, h_m$ and $\gamma_1, \ldots, \gamma_n$ be nonnegative integers and monic polynomials, respectively, such that $d(\gamma_1) + \cdots + d(\gamma_n) = q$. Let $r = n - q$. Then there exists a matrix $B \in \mathbb{F}^{n\times m}$, $m \le r$, such that (A, B) has $h_1, \ldots, h_m$ as Hermite indices and $\gamma_1, \ldots, \gamma_n$ as invariant factors if and only if there are m monic polynomials $\tau_1, \ldots, \tau_m$ such that $d(\tau_i) = h_i$, $1 \le i \le m$, and*

$$\alpha_{i-m} \mid \gamma_i \mid \alpha_i, \ 1 \le i \le n, \tag{3.2}$$

$$\sigma_1 \ldots \sigma_{r-m+j} \mid \text{g.c.d.}\{\tau_{i_1} \ldots \tau_{i_j} : 1 \le i_1 < \cdots < i_j \le m\}, \ 1 \le j \le m, \tag{3.3}$$

$$\sigma_1 \ldots \sigma_r = \tau_1 \ldots \tau_m, \tag{3.4}$$

where we agree that $\alpha_i := 1$ for $i < 1$ and $\sigma_j = \frac{\beta^j}{\beta^{j-1}}$, $1 \le j \le r$, and $\beta^j = \prod_{i=1}^{n+j} l.c.m.(\gamma_{i-j}, \alpha_{i-r})$, $0 \le j \le r$.

With the help of this result we can generalize Theorem 2.11 to give a complete solution to Problem 1.

THEOREM 3.2. *Let $A \in \mathbb{F}^{n\times n}$, and let $\alpha_1 \mid \ldots \mid \alpha_n$ be its invariant factors. Let $k_1 \ge \cdots \ge k_m > 0$ and $h_1 \ge \cdots \ge h_m \ge 0$ be nonnegative integers and $\gamma_1, \ldots, \gamma_n$ monic polynomials, such that $d(\gamma_1) + \cdots + d(\gamma_n) = q$. Let $r = n - q$. Then there exists a matrix $B \in \mathbb{F}^{n\times m}$, $m \le r$, such that (A, B) has $h_1, \ldots, h_m$ as Hermite indices, $k_1, \ldots, k_m$ as controllability indices, and $\gamma_1, \ldots, \gamma_n$ as invariant factors if and only if there are m monic polynomials $\tau_1, \ldots, \tau_m$ such that $d(\tau_i) = h_i$, $1 \le i \le m$, and the conditions* (3.2)–(3.4) *and* (2.4) *are satisfied.*

Proof of Theorem 3.2.

Necessity. If there exists B such that (A, B) has $k_1, \ldots, k_m$ as controllability indices, $h_1, \ldots, h_m$ as Hermite indices, and $\gamma_1, \ldots, \gamma_n$ as invariant factors, by Lemma 3.1 conditions (3.2)–(3.4) hold. Moreover, there exists a matrix P such that (3.1) holds and the controllability indices and the Hermite indices of (A, B) are those of the pair (A_1, B_1). Then, by Lemma 2.2 condition (2.4) holds.

Sufficiency. By Lemma 3.1, if conditions (3.2), (3.3), and (3.4) hold, there exists a matrix $B \in \mathbb{F}^{n\times m}$ such that (A, B) has $h_1, \ldots, h_m$ as Hermite indices and $\gamma_1, \ldots, \gamma_n$ as invariant factors.

If we assume that $h_1 \ge \cdots \ge h_p > 0 = h_{p+1} = \cdots = h_m$, then there exists a matrix P such that

$$PAP^{-1} = \begin{bmatrix} A_c & A_2 \\ 0 & A_3 \end{bmatrix} \qquad PB = \begin{bmatrix} B_c \\ 0 \end{bmatrix},$$

where $(A_c, B_c) \in \mathbb{F}^{r\times r} \times \mathbb{F}^{r\times m}$ has the structure of that of Lemma 1.3 with $h_1, \ldots, h_m$ as Hermite indices and $\gamma_{r+1}, \ldots, \gamma_n$ are the invariant factors of A_3.

Now, as in the proof of Theorem 2.11, if conditions (3.2)–(3.4) and (2.4) hold, then there exists B_c' such that

$$\left(\begin{bmatrix} A_c & A_2 \\ 0 & A_3 \end{bmatrix}, \begin{bmatrix} B_c' \\ 0 \end{bmatrix} \right)$$

has $k_1, \ldots, k_m$ as controllability indices, $h_1, \ldots, h_m$ as Hermite indices, and $\gamma_1, \ldots, \gamma_n$ as invariant factors.

As

$$A \overset{s}{\sim} \begin{bmatrix} A_c & A_2 \\ 0 & A_3 \end{bmatrix},$$

by applying Lemma 1.2, the theorem follows. □

COROLLARY 3.3. *Let $A \in \mathbb{F}^{n\times n}$, and let $\alpha_1 \mid \ldots \mid \alpha_n$ be its invariant factors, $\mathbb{F}$ algebraically closed. Let $k_1 \geq \cdots \geq k_m > 0$, $h_1 \geq \cdots \geq h_m \geq 0$ be nonnegative integers and $\gamma_1, \ldots, \gamma_n$ monic polynomials such that $d(\gamma_1) + \cdots + d(\gamma_n) = q$. Let $r = n - q$. Then there exists a matrix $B \in \mathbb{F}^{n\times m}$, $m \leq r$, such that (A, B) has $k_1, \ldots, k_m$ as controllability indices, $h_1, \ldots, h_m$ as Hermite indices, and $\gamma_1, \ldots, \gamma_n$ as invariant factors if and only if the conditions* (3.2), (2.4), *and*

$$(h_1, \ldots, h_m) \prec (d(\sigma_r), \ldots, d(\sigma_1)) \tag{3.5}$$

are satisfied.

A consequence of this corollary and Proposition 1.1 is the following theorem.

THEOREM 3.4. *Let $(A, B) \in \mathbb{F}^{n\times n} \times \mathbb{F}^{n\times m}$ be a matrix pair with $k_1 \geq \cdots \geq k_m > 0$ and $\gamma_1 \mid \ldots \mid \gamma_n$ as controllability indices and invariant factors, respectively. Let $\alpha_1 \mid \ldots \mid \alpha_n$ and $h_1 \geq \cdots \geq h_m \geq 0$ be monic polynomials and nonnegative integers. Then there exist a state feedback matrix $F \in \mathbb{F}^{m\times n}$ and a nonsingular matrix $Q \in \mathbb{F}^{m\times m}$ such that $A + BF$ has $\alpha_1, \ldots, \alpha_n$ as invariant factors and $(A + BF, BQ)$ has $h_1, \ldots, h_m$ as Hermite indices if and only if conditions* (3.2), (2.4), *and* (3.5) *hold.*

Acknowledgment. Thanks are given to the referees for their valuable suggestions for improving this paper. The authors are specially indebted to one of the referees for suggesting the equivalence of Problems 1 and 2.

REFERENCES

[1] I. BARAGAÑA AND I. ZABALLA, *Feedback invariants of pairs of matrices and restrictions*, Linear and Multilinear Algebra, 33 (1999), pp. 101–116.
[2] I. BARAGAÑA AND I. ZABALLA, *Feedback invariants of supplementary pairs of matrices*, Automatica J. IFAC, 33 (1997), pp. 2119–2130.
[3] G. BASILE AND G. MARRO, *Controlled and Conditioned Invariants in Linear System Theory*, Prentice Hall, Englewood Cliffs, NJ, 1992.
[4] P. BRUNOVSKY, *Classification of linear controllable systems*, Kybernetika, 3 (1970), pp. 173–188.
[5] C.-T. CHEN, *Linear System Theory and Design*, Oxford University Press, New York, 1984.
[6] R. GANTMACHER, *Théorie des Matrices:* I, Dunod, Paris, 1966.
[7] I. GOHBERG, P. LANCASTER, AND L. RODMAN, *Invariant Subspaces of Matrices with Applications*, John Wiley and Sons, New York, 1986.
[8] G. H. HARDY, J. E. LITTLEWOOD, AND G. PÓLYA, *Inequalities*, Cambridge University Press, Cambridge, UK, 1967.
[9] T. KAILATH, *Linear Systems*, Prentice Hall, Englewood Cliffs, NJ, 1980.
[10] I. G. MACDONALD, *Symetric Functions and Hall Polynomials*, Oxford University Press, Oxford, UK, 1979.
[11] C. C. MACDUFFEE, *The Theory of Matrices*, Chelsea Publishing Company, New York, 1933.
[12] V. M. POPOV, *Invariant description of linear, time-invariant controllable systems*, SIAM J. Control, 10 (1972), pp. 252–264.
[13] H. H. ROSENBROCK, *State-Space and Multivariable Theory*, Thomas Nelson and Sons, London, 1970.
[14] P. SIPARIS AND H. BLOMBERG, *Structured systems models part* 1*: Controllability and observability indices*, Int. J. Systems Sci., 2 (1991), pp. 1047–1069.

[15] W. A. Wolovich, *Linear Multivariable Systems*, Springer-Verlag, New York, 1974.
[16] I. Zaballa, *Matrices with prescribed rows and invariant factors*, Linear Algebra Appl., 87 (1987), pp. 113–146.
[17] I. Zaballa, *Interlacing inequalities and control theory*, Linear Algebra Appl., 101 (1988), pp. 9–31.
[18] I. Zaballa, *Interlacing and majorization in invariant factor assignment problems*, Linear Algebra Appl., 121 (1989), pp. 409–421.
[19] I. Zaballa, *Matrices with prescribed invariant factors*, Linear and Multilinear Algebra, 27 (1990), pp. 325–343.
[20] I. Zaballa, *Controllability and Hermite indices of matrix pairs*, Internat. J. Control, 68 (1997), pp. 61–86.

SIAM J. CONTROL OPTIM.
Vol. 38, No. 4, pp. 1050–1066

MINIMIZING EXPECTED LOSS OF HEDGING IN INCOMPLETE AND CONSTRAINED MARKETS*

JAKŠA CVITANIĆ†

Abstract. We study the problem of minimizing the expected discounted loss

$$E\left[e^{-\int_0^T r(u)du}(C - X^{x,\pi}(T))^+\right]$$

when hedging a liability C at time $t = T$, using an admissible portfolio strategy $\pi(\cdot)$ and starting with initial wealth x. The existence of an optimal solution is established in the context of continuous-time Ito process *incomplete* market models, by studying an appropriate dual problem. It is shown that the optimal strategy is of the form of a knock-out option with payoff C, where the "domain of the knock-out" depends on the value of the optimal dual variable. We also discuss a dynamic measure for the risk associated with the liability C, defined as the supremum over different scenarios of the minimal expected loss of hedging C.

Key words. expected loss, hedging, incomplete markets, portfolio constraints, dynamic measures of risk

AMS subject classifications. Primary, 90A09, 90A46; Secondary, 93E20, 60H30

PII. S036301299834185X

1. Introduction.

In a *complete* financial market which is free of arbitrage opportunities, any sufficiently integrable random payoff (*contingent claim*) C, whose value has to be delivered and is known at time $t = T$, can be hedged perfectly: starting with a large enough initial capital x, an agent can find a trading strategy π that will allow his wealth $X^{x,\pi}(\cdot)$ to hedge the liability C *without risk* at time $t = T$, that is,

$$X^{x,\pi}(T) \geq C \quad \text{almost surely (a.s.)} \quad \text{for some portfolio } \pi(\cdot), \tag{1.1}$$

while maintaining "solvency" throughout $[0, T]$. (For an overview of standard results in complete and some incomplete markets in continuous-time Ito process models, see, for example, Cvitanić (1997), or a recent book by Karatzas and Shreve (1998).) This is either no longer possible or too expensive to accomplish in a market which is incomplete due to various "market frictions," such as insufficient number of assets available for investment, transaction costs, portfolio constraints, problems with liquidity, presence of a "large investor," and so on. In this paper we concentrate on the case in which incompleteness arises due to some assets not being available for investment and the more general case of portfolio constraints. Popular approaches to the problem of hedging a claim C in such contexts have been to either maximize the expected utility of the difference $-D := X^{x,\pi}(T) - C$ or minimize the risk of D. In particular, one of the most studied approaches is to minimize $E[D^2]$, so-called quadratic hedging of Föllmer–Schweizer–Sondermann (for recent results and references see Pham,

*Received by the editors July 13, 1998; accepted for publication (in revised form) September 15, 1999; published electronically April 4, 2000. This research was supported in part by National Science Foundation grant DMS-97-32810.

http://www.siam.org/journals/sicon/38-4/34185.html

†Department of Mathematics, University of Southern California, 1042 W. 36th PL, DRB155, Los Angeles, CA 90089 (cvitanic@math.usc.edu). The author is on leave from the Department of Statistics, Columbia University, New York, NY 10027 (cj@stat.columbia.edu).

Rheinländer, and Schweizer (1998), for example). An obvious disadvantage of this approach is that one is penalized for high profits and not just high losses. On the other hand, Artzner, Delbaen, Eber, and Heath (1999) have shown in a static hedging setting that the only measure of risk that satisfies certain natural "coherence" properties is of the type $E[\bar{D}^+]$ (or a supremum of these over a set of probability measures), where $\bar{D}^+$ is the discounted value of the positive part of D. Motivated by this work, Cvitanić and Karatzas (1999) solve the problem of minimizing $E[\bar{D}^+]$ in a context of a *complete* continuous-time Ito process model for the financial market. We solve in this paper the same problem in a more difficult context of incomplete or constrained markets. Recently, Pham (1998) has solved the problem of minimizing $E[(D^+)^p]$ for $p > 1$ in discrete-time models and under cone constraints. Moreover, independently from Pham and the present paper, Föllmer and Leukert (1999b) analyze the problem of minimizing $E[l(D^+)]$ for a general loss function l and in general incomplete semimartingale models, emphasizing the Neyman–Pearson lemma approach, as opposed to the duality approach. The former approach was used by the same authors in Föllmer and Leukert (1999a) to solve the problem of maximizing the probability of perfect hedge $P[D \leq 0]$. Some early work on problems like these is presented in Dembo (1997), in a one-period setting. A very general study of the the duality approach and its use in the utility maximization context can be found in Kramkov and Schachermayer (1999).

As mentioned above, another approach would be to try to hedge away all the risk of the agent by superreplicating the claim he has to deliver at time T, namely to have $X^{x,\pi}(T) \geq C$, a.s. This has been done in the framework of constraints by Cvitanić and Karatzas (1993), Broadie, Cvitanić, and Soner (1998), and Cvitanić, Pham, and Touzi (1999). However, the cost $x = x_C$ of the least expensive strategy accomplishing the superreplication is typically very high, and hence the strategy is appropriate neither for pricing nor for hedging purposes. For example, if the agent sells a call option C on one share of stock S, and he cannot borrow money, then his cost of superreplicating the option is equal to the price of one share of S. Nobody would pay this much for the option if they can buy the stock itself. More interesting examples include newly deregulated energy markets, reinsurance markets, and emerging markets. The cost of superreplication in these markets is usually too high (even infinite), and one is forced to introduce preferences, typically in terms of a loss or a utility function. This is the approach taken also in this paper, with a linear loss function.

Suppose now that, in addition to the genuine risk that the liability C represents, the agent also faces some uncertainty regarding the model for the financial market itself. Following Cvitanić and Karatzas (1999), we capture such uncertainty by allowing a family $\mathcal{P}$ of possible "real-world probability measures," instead of just one measure. Thus, the "max-min" quantity

$$\underline{V}(x) := \sup_{P \in \mathcal{P}} \inf_{\pi} E^P[\bar{D}^+] \tag{1.2}$$

represents the maximal risk that the agent can encounter when faced with the "worst possible scenario" $P \in \mathcal{P}$. In the special case of incomplete markets and under the condition that all equivalent martingale measures are included in the set of possible real-world measures $\mathcal{P}$, we show that

$$\underline{V}(x) = \overline{V}(x) := \inf_{\pi} \sup_{P \in \mathcal{P}} E^P[\bar{D}^+]. \tag{1.3}$$

In other words, the corresponding fictitious "stochastic game" between the market and the agent has a value. The trading strategy attaining this value is shown to be

the one that corresponds to borrowing just enough money from the bank at time $t = 0$ as to be able to have at least the amount C at time $t = T$.

We describe the market model in section 2 and introduce the optimization problem in section 3. As is by now standard in financial mathematics, we define a dual problem, whose optimal solution determines the optimal terminal wealth $X^{x,\hat{\pi}}(T)$. It turns out that this terminal wealth is of the "knock-out" option type—namely, it is either equal to C or to 0 or to a certain (random) value $0 \leq B \leq C$, depending on whether the optimal dual variable is less than, larger than, or equal to one, respectively. What makes the dual problem more difficult than in the usual utility optimization problems (as in Cvitanić and Karatzas (1992)) is that the objective function fails to be everywhere differentiable, and the optimal dual variable (related to the Radon–Nikodym derivative of an "optimal change of measure") can be zero with positive probability. Nevertheless, we are able to solve the problem using nonsmooth optimization techniques for infinite dimensional problems, which can be found in Aubin and Ekeland (1984). We discuss in section 4 the stochastic game associated with (1.2) and (1.3).

2. The market model. We recall here the standard Ito process model for a financial market $\mathcal{M}$. It consists of one *bank account* and d *stocks*. Price processes $S_0(\cdot)$ and $S_1(\cdot), \ldots, S_d(\cdot)$ of these instruments are modeled by the equations

$$\begin{aligned} &dS_0(t) = S_0(t)r(t)dt\ , \quad S_0(0) = 1, \\ &dS_i(t) = S_i(t)\left[b_i(t)dt + \sum_{j=1}^{d} \sigma_{ij}(t)dW^j(t)\right], \quad S_i(0) = s_i > 0\ ;\ i = 1, \ldots, d. \end{aligned} \tag{2.1}$$

Here $W(\cdot) = (W^1(\cdot), \ldots, W^d(\cdot))'$ is a standard d-dimensional Brownian motion on a complete probability space $(\Omega, \mathcal{F}, P)$, endowed with a filtration $\mathbf{F} = \{\mathcal{F}(t)\}_{0 \leq t \leq T}$, the P-augmentation of $\mathcal{F}^W(t) := \sigma(W(s);\ 0 \leq s \leq t)\ ,\ \ 0 \leq t \leq T$, the filtration generated by the Brownian motion $W(\cdot)$. The *coefficients* $r(\cdot)$ (interest rate), $b(\cdot) = (b_1(\cdot), \ldots, b_d(\cdot))'$ (vector of stock return rates) and $\sigma(\cdot) = \{\sigma_{ij}(\cdot)\}_{1 \leq i,j \leq d}$ (matrix of stock-volatilities) of the model $\mathcal{M}$ are all assumed to be progressively measurable with respect to $\mathbf{F}$. Furthermore, the matrix $\sigma(\cdot)$ is assumed to be invertible, and all processes $r(\cdot)$, $b(\cdot)$, $\sigma(\cdot)$, $\sigma^{-1}(\cdot)$ are assumed to be bounded uniformly in $(t, \omega) \in [0, T] \times \Omega$.

The "risk premium" process

$$\theta_0(t) := \sigma^{-1}(t)[b(t) - r(t)\tilde{\mathbf{1}}]\ ,\ \ 0 \leq t \leq T, \tag{2.2}$$

where $\tilde{\mathbf{1}} = (1, \ldots, 1)' \in \mathbb{R}^d$, is then bounded and $\mathbf{F}$-progressively measurable. Therefore, the process

$$Z_0(t) := \exp\left[-\int_0^t \theta_0'(s)dW_0(s) - \frac{1}{2}\int_0^t ||\theta_0(s)||^2 ds\right],\ 0 \leq t \leq T, \tag{2.3}$$

is a P-martingale, and

$$P_0(\Lambda) := E[Z_0(T)1_\Lambda], \quad \Lambda \in \mathcal{F}(T) \tag{2.4}$$

is a probability measure equivalent to P on $\mathcal{F}(T)$. Under this *risk-neutral equivalent martingale measure* P_0, the discounted stock prices $\frac{S_1(\cdot)}{S_0(\cdot)}, \ldots, \frac{S_d(\cdot)}{S_0(\cdot)}$ become martingales, and the process

$$W_0(t)\ :=\ W(t) + \int_0^t \theta_0(s)ds\ ,\ \ 0 \leq t \leq T, \tag{2.5}$$

becomes Brownian motion, by the Girsanov theorem.

Consider now an agent who starts out with initial capital x and can decide, at each time $t \in [0,T]$, what proportion $\pi_i(t)$ of his (nonnegative) wealth to invest in each of the stocks $i = 1, \ldots, d$. However, the portfolio process $(\pi_1(\cdot), \ldots, \pi_d(\cdot))'$ has to take values in a given closed convex set $K \subset \mathbb{R}^d$ of constraints, for almost everywhere (a.e.) $t \in [0,T]$, a.s. We will also assume that K contains the origin. For example, if the agent can hold neither short nor long positions in the last $d-m$ stocks $S_{m+1}(\cdot), \ldots, S_d(\cdot)$, we get a typical example of an incomplete market, in the sense that not all square-integrable payoffs can be exactly replicated. (One of the best known examples of incomplete markets, the case of stochastic volatility, is included in this framework.) Another typical example is the case of an agent who has limits on how much he can borrow from the bank, or how much he can go short or long in a particular stock.

With $\pi(t) = (\pi_1(t), \ldots, \pi_d(t))' \in K$ chosen, the agent invests the amount $X(t)(1 - \sum_{i=1}^d \pi_i(t))$ in the bank account, at time t, where we have denoted $X(\cdot) \equiv X^{x,\pi,\kappa}(\cdot)$ his wealth process. Moreover, for reasons of mathematical convenience, we allow the agent to spend money outside of the market, and $\kappa(\cdot) \geq 0$ denotes the corresponding *cumulative consumption process*. The resulting wealth process satisfies the equation

$$\begin{aligned} dX(t) &= -d\kappa(t) + \left[X(t)(1 - \sum_{i=1}^d \pi_i(t))\right] r(t)dt \\ &\quad + \sum_{i=1}^d \pi_i(t)X(t)\left[b_i(t)dt + \sum_{j=1}^d \sigma_{ij}(t)dW^j(t)\right] \\ &= -d\kappa(t) + r(t)X(t)dt + \pi'(t)\sigma(t)X(t)dW_0(t); \quad X(0) = x. \end{aligned}$$

Denoting

$$\bar{X}(t) = e^{-\int_0^t r(u)du} X(t), \tag{2.6}$$

the discounted version of a process $X(\cdot)$, we get the equivalent equation

$$d\bar{X}(t) = -e^{-\int_0^t r(u)du} d\kappa(t) + \pi'(t)\sigma(t)\bar{X}(t)dW_0(t)\ ; \quad X(0) = x. \tag{2.7}$$

It follows that $\bar{X}(\cdot)$ is a nonnegative local P_0-supermartingale, hence also a P_0-supermartingale, by Fatou's lemma. Therefore, if τ_0 is defined to be the first time it hits zero, we have $X(t) = 0$ for $t \geq \tau_0$, so that the portfolio values $\pi(t)$ are irrelevant after that happens. Accordingly, we can and do set $\pi(t) = 0$ for $t \geq \tau_0$.

More formally, we have the following definition.

DEFINITION 2.1. (i) *A portfolio process* $\pi : [0,T] \times \Omega \to \mathbb{R}^d$ *is* $\mathbf{F}$*-progressively measurable and satisfies* $\int_0^T ||\pi(t)||^2 dt < \infty$*, a.s., as well as*

$$\pi(t) \in K \quad \text{for a.e. } t \in [0,T] \tag{2.8}$$

a.s. A consumption process $\kappa(\cdot)$ *is a nonnegative, nondecreasing, progressively measurable process with right-continuous with left limits (RCLL) paths, with* $\kappa(0) = 0$ *and* $\kappa(T) < \infty$.

(ii) *For a given portfolio and consumption processes* $\pi(\cdot)$, $\kappa(\cdot)$, *the process* $X(\cdot) \equiv X^{x,\pi,\kappa}(\cdot)$ *defined by* (2.7) *is called the wealth process corresponding to strategy* (π, κ) *and initial capital* x.

(iii) *A portfolio-consumption process pair* $(\pi(\cdot),\kappa(\cdot))$ *is called admissible for the initial capital* x*, and we write* $(\pi,\kappa)\in\mathcal{A}(x)$*, if*

$$X^{x,\pi,\kappa}(t)\geq 0,\qquad 0\leq t\leq T, \tag{2.9}$$

holds a.s.

We refer to the lower bound of (2.9) as a *margin requirement*. The no-arbitrage price of a contingent claim C in a complete market is unique and is obtained by multiplying ("discounting") the claim by $H_0(T):=Z_0(T)/S_0(T)$ and taking expectation. Since the market here is incomplete, there are more relevant *stochastic discount factors* other than $H_0(T)$. We identify them along the lines of Cvitanić and Karatzas (1993), hereafter [CK93], and Karatzas and Kou (1996), hereafter [KK96], as follows: Introduce the *support function*

$$\delta(\nu):=\sup_{\pi\in K}\{-\pi'\nu\} \tag{2.10}$$

of the set $-K$, as well as its *barrier cone*

$$\tilde{K}:=\{\nu\in\mathbb{R}^d\ /\ \delta(\nu)<\infty\}. \tag{2.11}$$

For the rest of the paper we assume the following mild conditions.

Assumption 2.2. The closed convex set $K\subset\mathbb{R}^d$ contains the origin; in other words, the agent is allowed not to invest in stocks at all. In particular, $\delta(\cdot)\geq 0$ on $\tilde{K}$. Moreover, the set K is such that $\delta(\cdot)$ is continuous on the barrier cone $\tilde{K}$ of (2.11).

Denote by $\mathcal{D}$ the set of all *bounded* progressively measurable process $\nu(\cdot)$ taking values in $\tilde{K}$ a.e. on $\Omega\times[0,T]$. In analogy with (2.2)–(2.5), introduce

$$\theta_\nu(t):=\sigma^{-1}(t)[\nu(t)+b(t)-r(t)\tilde{\mathbf{1}}]\ ,\quad 0\leq t\leq T, \tag{2.12}$$

$$Z_\nu(t):=\exp\left[-\int_0^t\theta_\nu'(s)dW(s)-\frac{1}{2}\int_0^t||\theta_\nu(s)||^2ds\right],\ 0\leq t\leq T, \tag{2.13}$$

$$P_\nu(\Lambda):=E[Z_\nu(T)1_\Lambda],\quad \Lambda\in\mathcal{F}(T), \tag{2.14}$$

$$W_\nu(t)\ :=\ W(t)+\int_0^t\theta_\nu(s)ds\ ,\quad 0\leq t\leq T, \tag{2.15}$$

a P_ν-Brownian motion. Also denote

$$H_\nu(t):=e^{-\int_0^t\delta(\nu(u))du}Z_\nu(t). \tag{2.16}$$

Note that

$$dZ_\nu(t)=-Z_\nu(t)\theta_\nu'(t)dW(t). \tag{2.17}$$

From this and (2.7) we get, by Ito's rule,

$$\begin{aligned}d(H_\nu(t)\bar{X}(t))=&-e^{-\int_0^t r(u)du}H_\nu(t)d\kappa(t)-[\delta(\nu(t))+\pi'(t)\nu(t)]H_\nu(t)\bar{X}(t)dt\\&+[\pi'(t)\sigma(t)-\theta_\nu(t)]H_\nu(t)\bar{X}(t)dW(t)\ ;\quad X(0)=x\end{aligned} \tag{2.18}$$

for all $\nu\in\mathcal{D}$. Therefore, $H_\nu(\cdot)\bar{X}(\cdot)$ is a P-local supermartingale (note that $\delta(\nu)+\pi'\nu\geq 0$ for $\pi\in K$ and $\nu\in\tilde{K}$), and from (2.9) thus also a P-supermartingale, by Fatou's lemma. Consequently,

$$E[H_\nu(T)\bar{X}^{x,\pi,\kappa}(T)]\leq x\quad\forall\ (\pi,\kappa,\nu)\ \in\ \mathcal{A}(x)\times\mathcal{D}\ . \tag{2.19}$$

3. The minimization problem and its dual. Suppose now that, at time $t = T$, the agent has to deliver a payoff given by a contingent claim C, a random variable in $\mathbf{L}^2(\Omega, \mathcal{F}(T), P)$, with

$$P[C \geq 0] = 1 \quad \text{and} \quad P[C > 0] > 0. \tag{3.1}$$

Introduce a (possibly infinite) process

$$\bar{C}(t) := ess \sup_{\nu \in \mathcal{D}} E\left[H_\nu(T)\bar{C} \;\middle|\; \mathcal{F}(t) \right], \qquad 0 \leq t \leq T, \tag{3.2}$$

a.s., the discounted version of the process

$$C(\cdot) := S_0(\cdot)\bar{C}(\cdot). \tag{3.3}$$

We have denoted

$$\bar{C} := \frac{C}{S_0(T)} \tag{3.4}$$

the discounted value of the $\mathcal{F}(T)$-measurable random variable C. We impose the following assumption throughout the rest of the paper (see Remark 3.14 for a discussion on the relevance of this assumption).

Assumption 3.1. We assume

$$C(0) = \sup_{\nu \in \mathcal{D}} E[H_\nu(T)\bar{C}] < \infty. \tag{3.5}$$

The following theorem is taken from the literature on constrained financial markets (see, for example, [CK93], [KK96], or Cvitanić (1997)).

THEOREM 3.2. (Cvitanić and Karatzas (1993)). *Let $C \geq 0$ be a given contingent claim. Under Assumption* 3.1*, the process $C(\cdot)$ of* (3.3) *is finite, and it is equal to the minimal admissible wealth process hedging the claim C. More precisely, there exists a pair $(\pi_C, \kappa_C) \in \mathcal{A}(C(0))$ such that*

$$C(\cdot) \equiv X^{C(0),\pi_C,\kappa_C}(\cdot), \tag{3.6}$$

and, if for some $x \geq 0$ and some pair $(\pi, \kappa) \in \mathcal{A}(x)$ we have

$$X^{x,\pi,\kappa}(T) \geq C, \;\; P - a.s., \tag{3.7}$$

then

$$X^{x,\pi,\kappa}(t) \geq C(t), \quad 0 \leq t \leq T, \;\; P - a.s.$$

Consequently, if $x \geq C(0)$, there exists then an admissible pair $(\pi, \kappa) \in \mathcal{A}(x)$ such that $X^{x,\pi,\kappa}(T) \geq C$. Achieving a "hedge without risk" is not possible for $x < C(0)$. Motivated by results of Artzner et al. (1999) (and similarly as in a complete market setting of Cvitanić and Karatzas (1999)) we choose the following risk function to be minimized:

$$V(x) \equiv V(x; C) := \inf_{(\pi,\kappa) \in \mathcal{A}(x)} E\left[\bar{C} - \bar{X}^{x,\pi,\kappa}(T)\right]^+. \tag{3.8}$$

In other words, we are minimizing the expected discounted net loss, over all admissible trading strategies.

If $x \geq C(0)$, we have $V(x) = 0$, because, as mentioned above, we can find a wealth process that hedges C. Moreover, the margin requirement (2.9) implies that $x \geq 0$, so we assume from now on that

$$0 < x < C(0). \tag{3.9}$$

Note that we can (and do) assume $X^{x,\pi,\kappa}(T) \leq C$, $P-$a.s., in our optimization problem (3.8), since the agent can always consume down to the value of C, in case he has more than C at time T. In particular, if $C(\omega) = 0$, we can (and do) assume $X^{x,\pi,\kappa}(T,\omega) = 0$, too. This means that the set $\{C = 0\} \in \mathcal{F}(T)$ is not relevant for the problem (3.8), which motivates us to define a new probability measure

$$P^C(\Lambda) = \frac{1}{E[\bar{C}]} E[\bar{C}\mathbf{1}_\Lambda], \quad \Lambda \in \mathcal{F}(T) \tag{3.10}$$

(see also Remark 3.14 (ii)). Denote by E^C the associated expectation operator.

The problem (3.8) has then an equivalent formulation

$$V(x) = E[\bar{C}] \inf_{(\pi,\kappa)\in\mathcal{A}(x)} E^C \left[1 - \frac{X^{x,\pi,\kappa}(T)}{C}\right]^+. \tag{3.11}$$

We approach the problem (3.11) by recalling familiar tools of convex duality: starting with the convex loss function $R(y) = (1-y)^+$, consider its Legendre–Fenchel transform

$$\tilde{R}(z) := \min_{0\leq y\leq 1} [R(y) + yz] = z \wedge 1, \qquad z \geq 0 \tag{3.12}$$

(where $z \wedge 1 = \min\{z, 1\}$). The minimum in (3.12) is attained by any number $I(z;b)$ of the form

$$I(z;b) := \left\{ \begin{array}{lll} 0 & ; & z > 1 \\ 1 & ; & 0 \leq z < 1 \\ b & ; & z = 1 \end{array} \right\}, \tag{3.13}$$

where $0 \leq b \leq 1$.

Consequently, denoting

$$Y^{x,\pi,\kappa} := \frac{X^{x,\pi,\kappa}(T)}{C},\ P^C - a.s., \tag{3.14}$$

we conclude from (3.12) that for any initial capital $x \in (0, C(0))$ and any $(\pi,\kappa) \in \mathcal{A}(x)$, $\nu \in \mathcal{D}$, $z \geq 0$ we have

$$(1 - Y^{x,\pi,\kappa})^+ \geq \tilde{R}(zH_\nu(T)) - zH_\nu(T)Y^{x,\pi,\kappa},\ P^C - a.s. \tag{3.15}$$

Thus, multiplying by $E[\bar{C}]$, taking expectations, and in conjunction with (2.19), we obtain

$$\begin{aligned} E[\bar{C}]E^C\left[1 - Y^{x,\pi,\kappa}\right]^+ &\geq E[\bar{C}]E^C\left[\tilde{R}(zH_\nu(T))\right] - zE[\bar{C}]E^C\left[H_\nu(T)Y^{x,\pi,\kappa}\right] \\ &\geq E[\bar{C}]E^C\left[\tilde{R}(zH_\nu(T))\right] - xz. \end{aligned} \tag{3.16}$$

This is a type of a duality relationship that has proved to be very useful in constrained portfolio optimization studied in Cvitanić and Karatzas (1992). The difference here is that we have to extend it to the random variables in the set

$$\mathcal{H} := \{H \in \mathbf{L}^1(\Omega, \mathcal{F}(T), P^C) \ / \ H \geq 0 \ P^C - a.s., \\ E[\bar{C}]E^C[HY^{x,\pi,\kappa}] \leq x, \ \forall \ (\pi,\kappa) \in \mathcal{A}(x)\}. \tag{3.17}$$

Remark 3.3. As the referee points out, it should be noted that the set $\mathcal{H}$ does not actually depend on x. This is because $X^{x,\pi,\kappa}(\cdot) = xX^{1,\pi,\kappa/x}(\cdot)$, so that the inequality in the definition (3.17) of $\mathcal{H}$ is equivalent to

$$E[\bar{C}]E^C[HY^{1,\pi,\kappa}] \leq 1 \quad \forall \ (\pi,\kappa) \in \mathcal{A}(1). \tag{3.18}$$

It is clear that $\mathcal{H}$ is a convex set. It is also closed in $\mathbf{L}^1(\Omega, \mathcal{F}(T), P^C)$. Indeed, if $H_n \to H$ in $\mathbf{L}^1(\Omega, \mathcal{F}(T), P^C)$, then there exists a (relabeled) subsequence $\{H_n\}_{n\in\mathbb{N}}$ converging to H, P^C-a.s.; therefore $H \geq 0$, $P^C - a.s.$, and, by Fatou's lemma, $E[\bar{C}]E^C[HY^{x,\pi,\kappa}] \leq x \quad \forall (\pi,\kappa) \in \mathcal{A}(x)$.

By Theorem 3.2 we have $C(\cdot) = X^{C(0),\pi_C,\kappa_C}(\cdot)$ for some $(\pi_C, \kappa_C) \in \mathcal{A}(x)$. Consequently, we have $Y^{C(0),\pi_C,\kappa_C} = 1$, $P^C - a.s.$, and therefore

$$E[\bar{C}]E^C[H] = E[\bar{C}H] \leq C(0) \quad \forall \ H \in \mathcal{H}, \tag{3.19}$$

where we extend a random variable H to the probability space $(\Omega, \mathcal{F}(T), P)$ by setting $H = 0$ on $\{C = 0\}$. Similarly, since $0 \in K$, taking $\tilde{X}^{x,0,0}(T) = x$ in the definition (3.17) of $\mathcal{H}$, we see that

$$E[\bar{C}]E^C[H/C] = E[H] \leq 1 \quad \forall \ H \in \mathcal{H}. \tag{3.20}$$

Moreover, since $E[\bar{C}]E^C[H_\nu(T)] \leq C(0) < \infty \quad \forall \nu \in \mathcal{D}$, and by (2.19), we get

$$\mathcal{H}_{\mathcal{D}} := \{H_\nu(T) \ / \ \nu \in \mathcal{D}\} \ \subset \ \mathcal{H}. \tag{3.21}$$

Remark 3.4. The idea of introducing the set $\mathcal{H}$ is similar to and inspired by the approach of Kramkov and Schachermayer (1999), who work with the set of all nonnegative processes $G(\cdot)$ such that $G(\cdot)\tilde{X}(\cdot)$ is a P-supermartingale for all admissible wealth processes $X(\cdot)$.

Next, arguing as above (when deducing (3.16)), we obtain

$$E[\bar{C}]E^C\left[1 - Y^{x,\pi,\kappa}\right]^+ \geq E[\bar{C}]E^C\left[\tilde{R}(zH)\right] - xz =: \tilde{J}(H;z) - xz \quad \forall \ H \in \mathcal{H}, \ z \geq 0, \tag{3.22}$$

where we have denoted

$$\tilde{J}(H;z) := E[\bar{C}]E^C[(zH) \wedge 1]. \tag{3.23}$$

It is easily seen that $-\tilde{J}(\cdot;z) : \mathbf{L}^1(\Omega, \mathcal{F}(T), P^C) \to \mathbb{R}$ is a convex, lower-semicontinuous and proper functional, in the terminology of convex analysis; see, for example, Aubin and Ekeland (1984), henceforth [AE84].

Remark 3.5. It is straightforward to see that the inequality of (3.22) holds as equality for some $(\hat{\pi}, \hat{\kappa}) \in \mathcal{A}(x)$ and $\hat{z} \geq 0$, $\hat{H} \in \mathcal{H}$, *if and only if* we have

$$E[\bar{C}]E^C\left[\hat{H}Y^{x,\hat{\pi},\hat{\kappa}}\right] = x \tag{3.24}$$

and

$$Y^{x,\hat{\pi},\hat{\kappa}} = I(\hat{z}\hat{H};\hat{B}) = 1_{\{\hat{z}\hat{H}<1\}} + \hat{B}1_{\{\hat{z}\hat{H}=1\}},\ P^C - a.s. \tag{3.25}$$

for some $\mathcal{F}(T)$-measurable random variable $\hat{B}$ that satisfies $0 \le \hat{B} \le 1$, P^C- a.s. We also set

$$Y^{x,\hat{\pi},\hat{\kappa}} = 0 \ \text{ on } \ \{C = 0\}.$$

If (3.24) and (3.25) are satisfied, then $(\hat{\pi},\hat{\kappa})$ is optimal for the problem (3.11), under the "change of variables" (3.14), since the lower bound of (3.22) is attained. Moreover, $\hat{H} \in \mathcal{H}$ is optimal for the auxiliary dual problem

$$\tilde{V}(z) = \sup_{H\in\mathcal{H}} \tilde{J}(H;z) \tag{3.26}$$

with $z = \hat{z}$. If we let

$$X^{x,\hat{\pi},\hat{\kappa}}(T) = CY^{x,\hat{\pi},\hat{\kappa}}, \quad P - a.s., \tag{3.27}$$

the conditions (3.24) and (3.25) become

$$E\left[\hat{H}\bar{X}^{x,\hat{\pi},\hat{\kappa}}(T)\right] = x \tag{3.28}$$

and

$$X^{x,\hat{\pi},\hat{\kappa}}(T) = C\left(1_{\{\hat{z}\hat{H}<1\}} + \hat{B}1_{\{\hat{z}\hat{H}=1\}}\right),\ P - a.s. \tag{3.29}$$

for some $\mathcal{F}(T)$-measurable random variable $\hat{B}$ that satisfies $0 \le \hat{B} \le 1$, a.s., and $X^{x,\hat{\pi},\hat{\kappa}}(T)$ is the terminal wealth of the strategy $(\hat{\pi},\hat{\kappa})$ which is optimal for the problem (3.8). □

In light of the preceding remark, our approach will be the following: we will try to find a number $\hat{z} > 0$, a solution $\hat{H}$ to the auxiliary dual problem (3.26) with $z = \hat{z}$, a number $\hat{z} > 0$, a random variable $\hat{B}$ as above, and a pair $(\hat{\pi},\hat{\kappa}) \in \mathcal{A}(x)$ such that (3.24) and (3.25) (or, equivalently, (3.28) and (3.29)) are satisfied.

THEOREM 3.6. *For any given $z > 0$, there exists an optimal solution $\hat{H} = \hat{H}_z \in \mathcal{H}$ for the auxiliary dual problem* (3.26).

Proof. Let $H_n \in \mathcal{H}$ be a sequence that attains the supremum in (3.26), so that

$$\lim_n \tilde{J}(H_n;z) = \tilde{V}(z).$$

Note that, by (3.19), $\mathcal{H}$ is a bounded set in $\mathbf{L}^1(\Omega,\mathcal{F}(T),P^C)$, so that by Komlós theorem (see Schwartz (1986), for example) there exists a random variable $\hat{H} \in \mathbf{L}^1(\Omega,\mathcal{F}(T),P^C)$ and a (relabeled) subsequence $\{H_i\}_{i\in\mathbb{N}}$ such that

$$G_n := \frac{1}{n}\sum_{i=1}^n H_i \to \hat{H}, \quad P^C - a.s.$$

Fatou's lemma then implies $\hat{H} \in \mathcal{H}$. Since $0 \le (zG_n) \wedge 1 \le 1$, by the dominated convergence theorem and concavity of $\tilde{J}(\cdot;z)$ we get

$$\tilde{J}(\hat{H};z) = \lim_n \tilde{J}\left(\frac{1}{n}\sum_{i=1}^n H_i;z\right) \ge \lim_n \left[\frac{1}{n}\sum_{i=1}^n \tilde{J}(H_i;z)\right] = \tilde{V}(z).$$

Thus, $\hat{H} \in \mathcal{H}$ is optimal. □

LEMMA 3.7. *The function $\tilde{V}(z)$ is continuous on $[0,\infty)$.*

Proof. Let $H \in \mathcal{H}$ and assume first $z_1, z_2 > 0$. We have

$$\begin{aligned}\tilde{J}(H;z_1) &= E[\bar{C}]E^C[(z_1H)\wedge 1] = E[\bar{C}]E^C\big[(z_2H)\wedge 1 + (z_1H)\wedge 1 - (z_2H)\wedge 1\big]\\ &= \tilde{J}(H;z_2) + E[\bar{C}]E^C\big[H(z_1-z_2)\mathbf{1}_{\{z_1H<1,z_2H<1\}} + (z_1H-1)\mathbf{1}_{\{z_1H<1,z_2H\geq 1\}}\\ &\qquad + (1-z_2H)\mathbf{1}_{\{z_1H\geq 1,z_2H<1\}}\big]\\ &\leq \tilde{V}(z_2) + 2E[\bar{C}](1-z_2/z_1)^+.\end{aligned}$$

Taking the supremum over $H \in \mathcal{H}$ we get $\tilde{V}(z_1)-\tilde{V}(z_2) \leq 2E[\bar{C}](1-z_2/z_1)^+$. Since we can do the same while interchanging the roles of z_1 and z_2, we have shown continuity on $(0,\infty)$. To prove continuity at $z_2 = 0$, note that, by duality and (3.20), we have

$$\tilde{J}(H;z_1) = E[\bar{C}]E^C[(z_1H)\wedge 1] \leq E[(\bar{C}-y)^+] + yz_1E[H] \leq E[(\bar{C}-y)^+] + yz_1$$

for all $z_1 > 0, y > 0$. Choosing first y large enough and then z_1 small enough, we can make the two terms on the right-hand side arbitrarily close to zero, uniformly in $H \in \mathcal{H}$. □

PROPOSITION 3.8. *For every $0 < x < C(0)$ there exists $\hat{z} = \hat{z}_x \in (0,\infty)$ that attains the supremum $\sup_{z\geq 0}[\tilde{V}(z) - xz]$.*

Proof. Denote

$$\alpha(z) := \tilde{V}(z) - xz.$$

Note that $\alpha(0) = 0$. It is clear that

$$\limsup_{z\to\infty} \alpha(z) < 0, \tag{3.30}$$

so that the supremum of $\alpha(z)$ over $[0,\infty)$ cannot be attained at $z = \infty$. Consequently, being continuous by Lemma 3.7, function $\alpha(z)$ either attains its supremum at some $\hat{z} > 0$, or else $\alpha(z) \leq \alpha(0) = 0$ for all $z > 0$. Suppose that the latter is true. We have then

$$x \geq \frac{\tilde{V}(z)}{z} \geq E[\bar{C}]E^C\left[H \wedge \frac{1}{z}\right] \tag{3.31}$$

for all $z > 0$ and $H \in \mathcal{H}$. In particular, we can use the dominated convergence theorem while letting $z \to 0$ to get

$$x \geq E[H\bar{C}]$$

for all $H \in \mathcal{H}_{\mathcal{D}}$. Taking the supremum over $H \in \mathcal{H}_{\mathcal{D}}$ we obtain $x \geq C(0)$, a contradiction. □

Denote $\hat{H} = \hat{H}_{\hat{z}}$ the optimal dual variable for problem (3.26), corresponding to $z = \hat{z}$ of Proposition 3.8. We want to show that there exists an $\mathcal{F}(T)$-measurable random variable $0 \leq \hat{B} \leq 1$ such that the optimal wealth for the primal problem is given by $CI(\hat{z}\hat{H}, \hat{B})$, where $I(z;b)$ is given in (3.13). In order to do that, we recall some notions and results from convex analysis, as presented, for example, in [AE84].

First, introduce the space

$$\mathbf{L} := \mathbf{L}^1(\Omega, \mathcal{F}(T), P^C) \times \mathbb{R} \tag{3.32}$$

with the norm

$$\|(Z,z)\| := E[\bar{C}]E^C|Z| + |z|$$

and its subset

$$\mathcal{G} := \{(zH,z) \in \mathbf{L} \ / \ z \geq 0, H \in \mathcal{H}\}. \tag{3.33}$$

It is easily seen that $\mathcal{G}$ is convex, by the convexity of $\mathcal{H}$. It is also closed in $\mathbf{L}$. Indeed, if we are given subsequences $z_n \geq 0$ and $H_n \in \mathcal{H}$ such that $(z_nH_n, z_n) \to (Z,z)$ in $\mathbf{L}$, then $z_n \to z$; we also have, from (3.19),

$$E^C|z_nH_n - zH_n| \leq |z_n - z|E^C[H_n] \leq \frac{C(0)}{E[\bar{C}]}|z_n - z|,$$

so that $zH_n \to Z$ in $\mathbf{L}^1(\Omega, \mathcal{F}(T), P^C)$. If $z = 0$ we get $Z = 0$, and we are done. If $z > 0$, we get $H_n \to Z/z$ in $\mathbf{L}^1(\Omega, \mathcal{F}(T), P^C)$, therefore $Z/z \in \mathcal{H}$ because $\mathcal{H}$ is closed in $\mathbf{L}^1(\Omega, \mathcal{F}(T), P^C)$, and we are done again. The closedness of $\mathcal{G}$ has been confirmed.

We now define a functional $\tilde{U} : \mathbf{L} \to \mathbb{R}$ by

$$\tilde{U}(Z,z) := -E[\bar{C}]E^C[Z \wedge 1] + xz = -\tilde{J}(Z;1) + xz. \tag{3.34}$$

It is easy to check that $\tilde{U}$ is convex, lower-semicontinuous, and proper on $\mathbf{L}$. Moreover, since we have

$$\begin{aligned}\tilde{J}(\hat{H},\hat{z}) - x\hat{z} &= \tilde{V}(\hat{z}) - x\hat{z} \geq \tilde{V}(z) - xz \\ &\geq \tilde{J}(H,z) - xz \ \ \forall\ (H,z) \in \mathcal{H} \times [0,\infty)\end{aligned}$$

from Proposition 3.8 and in the notation of Theorem 3.6, it follows that the pair $\hat{G} := (\hat{z}\hat{H}, \hat{z}) \in \mathcal{G}$ is optimal for the *dual problem*

$$\inf_{(Z,z)\in\mathcal{G}} \tilde{U}(Z,z). \tag{3.35}$$

Let $\mathbf{L}^* := \mathbf{L}^\infty(\Omega, \mathcal{F}(T), P^C) \times \mathbb{R}$ be the dual space to $\mathbf{L}$ and let $N(\hat{z}\hat{H}, \hat{z})$ be the *normal cone* to the set $\mathcal{G}$ at the point $(\hat{z}\hat{H}, \hat{z})$, given by

$$N(\hat{z}\hat{H},\hat{z}) = \{(Y,y) \in L^* \ / \ E[\bar{C}]E^C[\hat{z}\hat{H}Y] + \hat{z}y = \max_{(zH,z)\in\mathcal{G}}(E[\bar{C}]E^C[zHY] + zy)\} \tag{3.36}$$

by Proposition 4.1.4 in [AE84]. Let $\partial\tilde{U}(\hat{z}\hat{H},\hat{z})$ denote the subdifferential of $\tilde{U}$ at $(\hat{z}\hat{H},\hat{z})$, which, by Proposition 4.3.3 in [AE84], is given by

$$\begin{aligned}\partial\tilde{U}(\hat{z}\hat{H},\hat{z}) = \{(Y,y) \in \mathbf{L}^* \ / \ &\tilde{U}(\hat{z}\hat{H},\hat{z}) - \tilde{U}(Z,z) \\ &\leq E[\bar{C}]E^C[Y(\hat{z}\hat{H} - Z)] + y(\hat{z} - z) \quad \forall\ (Z,z) \in \mathbf{L}\}.\end{aligned} \tag{3.37}$$

Then, by Corollary 4.6.3 in [AE84], since $(\hat{z}\hat{H},\hat{z})$ is optimal for the problem (3.35), we obtain the following proposition.

PROPOSITION 3.9. *The pair $(\hat{z}\hat{H},\hat{z}) \in \mathcal{G}$ is a solution to*

$$0 \in \partial\tilde{U}(\hat{z}\hat{H},\hat{z}) + N(\hat{z}\hat{H},\hat{z}). \tag{3.38}$$

In other words, there exists a pair $(\hat{Y}, \hat{y}) \in \mathbf{L}^*$ *which belongs to the normal cone* $N(\hat{z}\hat{H}, \hat{z})$ *and such that* $-(\hat{Y}, \hat{y})$ *belongs to the subdifferential* $\partial\tilde{U}(\hat{z}\hat{H}, \hat{z})$.

From (3.36) and (3.37), this is equivalent to

$$E[\bar{C}]E^C[\hat{z}\hat{H}\hat{Y}] + \hat{z}\hat{y} \geq E[\bar{C}]E^C[zH\hat{Y}] + z\hat{y} \quad \forall z \geq 0, H \in \mathcal{H}, \tag{3.39}$$

and

$$\begin{aligned} &E[\bar{C}]E^C[\hat{Y}(\hat{z}\hat{H} - Z)] + \hat{y}(\hat{z} - z) \\ &\quad \leq E[\bar{C}]E^C[(\hat{z}\hat{H}) \wedge 1] - E[\bar{C}]E^C[Z \wedge 1] + x(z - \hat{z}) \quad \forall\ (Z, z) \in \mathbf{L}. \end{aligned} \tag{3.40}$$

It is clear from (3.40) (by letting $z \to \pm\infty$ while keeping Z fixed) that, necessarily,

$$\hat{y} = -x.$$

On the other hand, if we let $\hat{z} = z$ in (3.39), we get

$$E^C[\hat{Y}\hat{H}] \geq E^C[\hat{Y}H] \quad \forall H \in \mathcal{H}. \tag{3.41}$$

Moreover, letting $z = (\hat{z} + \varepsilon)$ for some $\varepsilon > 0$, $H = \hat{H}$ in (3.39), and recalling $\hat{y} = -x$, we obtain

$$x \geq E[\bar{C}]E^C[\hat{Y}\hat{H}].$$

Similarly, we get the reverse inequality by letting $\hat{z} = z - \varepsilon$ and $H = \hat{H}$ in (3.39) (recall that $\hat{z} > 0$ by Proposition 3.8), to obtain, finally,

$$E[\bar{C}]E^C[\hat{Y}\hat{H}] = x. \tag{3.42}$$

This last equality will correspond to (3.24) with $\hat{Y} = Y^{x,\hat{\pi},\hat{\kappa}}$, if we can show the following result and recall (3.14).

PROPOSITION 3.10. *There exists an admissible pair* $(\hat{\pi}, \hat{\kappa}) \in \mathcal{A}(x)$ *such that*

$$X^{x,\hat{\pi},\hat{\kappa}}(T) = C\hat{Y},\ P - a.s.$$

and such that (3.28) *is satisfied.*

(Here we set $\hat{Y} = 0$ on $\{C = 0\}$.)

Proof. This follows immediately from (3.41) and (3.42), which can be written as

$$x = E[\bar{C}\hat{Y}\hat{H}] = \sup_{H \in \mathcal{H}} E[\bar{C}\hat{Y}H]$$

(with $H = 0$ on $\{C = 0\}$). Indeed, Theorem 3.2 tells us that the right-hand side is no smaller than the minimal amount of initial capital needed to hedge $C\hat{Y}$; thus, there exists a pair $(\hat{\pi}, \hat{\kappa}) \in \mathcal{A}(x)$ that does the hedge. □

In order to "close the loop," it only remains to show (3.25).

PROPOSITION 3.11. *Let* $-(Y, y) \in \partial\tilde{U}(\hat{z}\hat{H}, \hat{z})$. *Then* $y = -x$ *and* Y *is of the form*

$$Y = 1_{\{\hat{z}\hat{H} < 1\}} + B1_{\{\hat{z}\hat{H} = 1\}}\ ,\ P^C - a.s. \tag{3.43}$$

for some $\mathcal{F}(T)$*-measurable random variable* B *that satisfies* $0 \leq B \leq 1$, P^C *a.s.*

Proof. We have already seen that $y = -x$. Define a random variable A by

$$Y = \mathbf{1}_{\{\hat{z}\hat{H}<1\}} + A. \tag{3.44}$$

From (3.40) with $\hat{y} = -x$ and $\hat{Y} = Y$, we get

$$\begin{aligned} &E^C[A(\hat{z}\hat{H} - Z)] - E^C[Z\mathbf{1}_{\{\hat{z}\hat{H}<1\}}] \\ &\quad \le E^C[\mathbf{1}_{\{\hat{z}\hat{H}\ge 1\}}] - E^C[Z \wedge 1] \quad \forall\ Z \in \mathbf{L}^1(\Omega, \mathcal{F}(T), P^C). \end{aligned} \tag{3.45}$$

Let $Z \in \mathbf{L}^1(\Omega, \mathcal{F}(T), P^C)$ be such that

$$\{\hat{z}\hat{H} < 1\} = \{Z < 1\}.$$

Then,

$$E^C[A(\hat{z}\hat{H} - Z)] = E^C[A(\hat{z}\hat{H} - Z)\mathbf{1}_{\{\hat{z}\hat{H}<1\}}] + E^C[A(\hat{z}\hat{H} - Z)\mathbf{1}_{\{\hat{z}\hat{H}\ge 1\}}] \le 0, \tag{3.46}$$

by (3.45). This implies

$$A \le 0 \ \text{ on } \ \{\hat{z}\hat{H} < 1\}, \quad A \ge 0 \ \text{ on } \ \{\hat{z}\hat{H} \ge 1\}, \quad P^C - a.s., \tag{3.47}$$

for otherwise we could make Z arbitrarily small (respectively, large) on $\{\hat{z}\hat{H} < 1\} \cap \{A > 0\}$ (respectively, on $\{\hat{z}\hat{H} \ge 1\} \cap \{A < 0\}$) to get a contradiction in (3.46).

Suppose now that $P^C[A < 0, \hat{z}\hat{H} < 1] > 0$. There exists then $\delta > 0$ such that $E^C[A(\hat{z}\hat{H} - 1)\mathbf{1}_{\{\hat{z}\hat{H}<1\}}] > \delta$, because of (3.47). For a given $\varepsilon > 0$, let $Z = 1 - \varepsilon$ on $\{\hat{z}\hat{H} < 1\}$ and $Z = 1$ on $\{\hat{z}\hat{H} \ge 1\}$, in (3.46). This gives

$$E^C[A(\hat{z}\hat{H} - 1 + \varepsilon)\mathbf{1}_{\{\hat{z}\hat{H}<1\}}] + E^C[A(\hat{z}\hat{H} - 1)\mathbf{1}_{\{\hat{z}\hat{H}\ge 1\}}] \le 0. \tag{3.48}$$

The left-hand side is greater than $\delta + \varepsilon E^C[A\mathbf{1}_{\{\hat{z}\hat{H}<1\}}]$ for all $\varepsilon > 0$ (recall (3.47) again), a contradiction to (3.48). Thus, we have shown

$$A = 0 \ \text{ on } \ \{\hat{z}\hat{H} < 1\}, \quad P^C - a.s. \tag{3.49}$$

Going back to (3.46), this implies

$$E^C[A(\hat{z}\hat{H} - Z)\mathbf{1}_{\{\hat{z}\hat{H}\ge 1\}}] \le 0 \tag{3.50}$$

for all $Z \in \mathbf{L}^1(\Omega, \mathcal{F}(T), P^C)$ such that $\{Z < 1\} = \{\hat{z}\hat{H} < 1\}$. If we set now $Z = 1$ on $\{\hat{z}\hat{H} \ge 1\}$, we get from (3.50) and (3.47)

$$A = 0 \ \text{ on } \ \{\hat{z}\hat{H} > 1\}, \quad P^C - a.s. \tag{3.51}$$

Using (3.49) and (3.51) in (3.45), we obtain

$$\begin{aligned} &E^C[A(1 - Z)\mathbf{1}_{\{\hat{z}\hat{H}=1\}}] - E^C[Z(\mathbf{1}_{\{\hat{z}\hat{H}<1\}} - \mathbf{1}_{\{Z<1\}})] \\ &\quad \le E^C[\mathbf{1}_{\{\hat{z}\hat{H}\ge 1\}} - \mathbf{1}_{\{Z\ge 1\}}] \quad \forall\ Z \in \mathbf{L}^1(\Omega, \mathcal{F}(T), P^C). \end{aligned} \tag{3.52}$$

Suppose now that $P^C[A > 1, \hat{z}\hat{H} = 1] > 0$. There exists then $\delta > 0$ such that $E^C[A\mathbf{1}_{\{\hat{z}\hat{H}=1, A>1\}}] > \delta + P^C[\hat{z}\hat{H} = 1, A > 1]$. Setting $Z = 0$ on $\{\hat{z}\hat{H} = 1, A > 1\}$, $Z = 1$ on $\{\hat{z}\hat{H} = 1, A \le 1\}$, and $Z = 1 - \varepsilon$ otherwise (for a given $\varepsilon > 0$), (3.52) implies

$$\begin{aligned} &E^C[A\mathbf{1}_{\{\hat{z}\hat{H}=1, A>1\}}] - (1 - \varepsilon)E^C[\mathbf{1}_{\{\hat{z}\hat{H}<1\}} - \mathbf{1}_{\{\hat{z}\hat{H}\ne 1\}}] \\ &\quad \le P^C[\hat{z}\hat{H} \ge 1] - P^C[\hat{z}\hat{H} = 1, A \le 1]. \end{aligned} \tag{3.53}$$

The left-hand side is greater than $\delta + P^C[\hat{z}\hat{H} = 1, A > 1] + (1-\varepsilon)(P^C[\hat{z}\hat{H} \neq 1] - P^C[\hat{z}\hat{H} < 1])$, so that from (3.53) we conclude $\delta - \varepsilon(P^C[\hat{z}\hat{H} \neq 1] - P^C[\hat{z}\hat{H} < 1]) \leq 0$ for all $\varepsilon > 0$, a contradiction. Therefore,

$$A \leq 1 \ , \quad \text{on} \ \ \{\hat{z}\hat{H} = 1\}, \quad P^C - a.s. \tag{3.54}$$

Together with (3.44), (3.47), (3.49), and (3.51), this completes the proof. □

We now state the main result of the paper.

THEOREM 3.12. *For any initial wealth x with $0 < x < C(0) < \infty$, there exists an optimal pair $(\hat{\pi}, \hat{\kappa}) \in \mathcal{A}(x)$ for the problem* (3.8) *of minimizing the expected loss of hedging the claim C. It can be taken as that strategy for which the terminal wealth $X^{x,\hat{\pi},\hat{\kappa}}(T)$ is given by* (3.29)*, i.e.,*

$$X^{x,\hat{\pi},\hat{\kappa}}(T) = C\left(1_{\{\hat{z}\hat{H}<1\}} + \hat{B}1_{\{\hat{z}\hat{H}=1\}}\right) \ , \ P - a.s. \tag{3.55}$$

Here $(\hat{z}, \hat{H})$ is an optimal solution for the dual problem (3.35)*, and $\hat{B}$ can be taken as the random variable B in Proposition* 3.11*, with (Y, y) replaced by some $(\hat{Y}, \hat{y}) \in \{-\partial\tilde{U}(\hat{z}\hat{H}, \hat{z}) \cap N(\hat{z}\hat{H}, \hat{z})\}$, which exists by Proposition* 3.9.

Proof. It follows from Remark 3.5. Indeed, it was observed in that remark that a pair $(\hat{\pi}, \hat{\kappa}) \in \mathcal{A}(x)$ is optimal for the problem (3.8) if it satisfies (3.28) and (3.55) for some $\mathcal{F}(T)$-measurable random variable $\hat{B}$, $0 \leq \hat{B} \leq 1$, and some $\hat{z} \geq 0$, $\hat{H} \in \mathcal{H}$. The existence of such a pair $(\hat{\pi}, \hat{\kappa}) \in \mathcal{A}(x)$ was established in Proposition 3.10 in conjunction with Proposition 3.11, with $\hat{B}$, $\hat{z}$, and $\hat{H}$ as in the statement of the theorem. □

The following simple example is mathematically interesting from several points of view. It shows that the optimal dual variable $\hat{H}$ can be equal to zero with positive probability, unlike the case of classical utility maximization under constraints (as in Cvitanić and Karatzas (1992)). Moreover, $\hat{z}\hat{H}$ can be equal to one with positive probability, so that the use of nonsmooth optimization techniques and subdifferentials for the dual problem is really necessary. It also shows why it can be mathematically convenient to allow nonzero consumption. Finally, it confirms that condition (3.5) is not always necessary for the dual approach to work.

Example 3.13. Suppose $r(\cdot) \equiv 0$ for simplicity, and let $C \geq 0$ be any contingent claim such that $P[C \geq x] > 0$. We consider the trivial primal problem for which $K = \{0\}$, so that there is only one possible admissible portfolio strategy $\hat{\pi}(\cdot) \equiv 0$ (in other words, the agent can invest only in the riskless asset). We do not assume condition (3.5), which, for these constraints, is equivalent to C being bounded. It is clear that the value $V(x)$ of the primal problem is $E[C - x]^+$, and duality implies

$$E[C - x]^+ \geq E[C((zH) \wedge 1)] - xz \tag{3.56}$$

for all $z \geq 0$, $H \in \mathcal{H}$ (see (3.22)). Here we can take $\mathcal{H}$ to be the set of all nonnegative random variables such that $E[H] \leq 1$. Let $\hat{z} := P[C \geq x] > 0$ and $\hat{z}\hat{H} := \mathbf{1}_{\{C \geq x\}}$. It is then easily checked that $\hat{H} \in \mathcal{H}$ and that the pair $(\hat{H}, \hat{z})$ attains equality in (3.56), so that the pair $(\hat{z}\hat{H}, \hat{z}) \in \mathcal{G}$ is optimal for the dual problem (3.35). One possible choice for the optimal terminal wealth is

$$X^{x,\hat{\pi},\hat{\kappa}}(T) = x\mathbf{1}_{\{C \geq x\}} + C\mathbf{1}_{\{C<x\}}.$$

According to (3.55), this corresponds to $\hat{B} = x/C$ on $\{C \geq x\}$, and $\hat{\kappa}(t) = 0$ for $t < T$, while $\hat{\kappa}(T) = (x - C)\mathbf{1}_{\{C<x\}}$.

Remark 3.14. (i) Assumption 3.1 is satisfied, for example, if C is bounded. We need it in order to get existence for the dual problem (3.35), due to our use of the Komlós theorem. Example 3.13 shows that this assumption is not always necessary: in this example the dual problem has a solution and there is no gap between the primal and the dual problem, even when (3.5) is not satisfied.

(ii) If we, in fact, assumed that C is bounded, the switch to the equivalent formulation (3.11) from (3.8) would not be necessary. (The reason for this is that the dual spaces of $\mathbf{L}^1(\Omega, \mathcal{F}(T), P)$ and $\mathbf{L}^1(\Omega, \mathcal{F}(T), P^C)$ are then the same, up to the equivalence class determined by the set $\{C = 0\}$.)

Remark 3.15. Since there are almost no examples with explicit solutions to this problem, it is of interest to study possible numerical algorithms. In Markovian continuous-time models this would involve solving Hamilton–Jacobi–Bellman PDEs, while in discrete models one could apply standard linear or convex programming techniques; see Blumenstein (1999) for some of these issues.

4. Dynamic measures of risk. Suppose now that we are not quite sure whether our subjective probability measure P is equal to the real-world measure. We would like to measure the risk of hedging the claim C under constraints given by set K, and under uncertainty about the real-world measure. According to Artzner et al. (1999), and Cvitanić and Karatzas (1999), it makes sense to consider the following quantities as the lower and upper bounds for the measure of such a risk, where we denote by $\mathcal{P}$ a set of possible real-world measures:

$$\underline{V}(x) := \sup_{Q \in \mathcal{P}} \inf_{(\pi,\kappa) \in \mathcal{A}(x)} E^Q \left[\bar{C} - \bar{X}^{x,\pi,\kappa}(T) \right]^+, \tag{4.1}$$

the *maximal risk that can be incurred, over all possible real-world measures*, dominated by its "min-max" counterpart

$$\overline{V}(x) := \inf_{(\pi,\kappa) \in \mathcal{A}(x)} \sup_{Q \in \mathcal{P}} E^Q \left[\bar{C} - \bar{X}^{x,\pi,\kappa}(T) \right]^+, \tag{4.2}$$

the upper-value of a fictitious *stochastic game* between an agent (who tries to choose $(\pi, \kappa) \in \mathcal{A}(x)$ so as to minimize his risk) and "the market" (whose "goal" is to choose the real-world measure that is least favorable for the agent). Here, E^Q is expectation under measure Q. A question is whether the "upper-value" (4.2) and the "lower-value" (4.1) of this game coincide and, if they do, to compute this common value. We shall answer this question only in a very specific setting as follows. Let P be the "reference" probability measure, as in the previous sections. We first change the margin requirement (2.9) to a more flexible requirement

$$\bar{X}^{x,\pi,\kappa}(t) \geq -k, \qquad 0 \leq t \leq T, \quad P - a.s., \tag{4.3}$$

where k is a constant such that $\infty > k \geq C(0) - x > 0$. We still assume $0 < x < C(0)$, and we look at the special case of the constraints given by

$$K = \{\pi \in \mathbb{R}^d \ / \ \pi_{m+1} = \cdots = \pi_d = 0\} \tag{4.4}$$

for some $m < d$. In other words, we only consider the case of a market which is incomplete due to the insufficient number of assets available for investment. In this case,

$$\tilde{K} = \{\nu \in \mathbb{R}^d \ / \ \nu_1 = \cdots = \nu_m = 0\}$$

and

$$\mathcal{D} = \{\text{bounded progress. meas. processes } \nu(\cdot) \ / \ \nu_1(\cdot) \equiv \cdots \equiv \nu_m(\cdot) \equiv 0\}.$$

We define the set $\mathcal{P}$ of possible real-world probability measures as follows. Let $\mathcal{E}$ be a set of progressively measurable and bounded processes $\nu(\cdot)$ and such that

$$\mathcal{D} \subset \mathcal{E}. \tag{4.5}$$

We set

$$\mathcal{P} := \{P_\nu \ / \ \nu \in \mathcal{E}\}, \tag{4.6}$$

in the notation of (2.14) (note that the reference measure P is not necessarily in $\mathcal{P}$). In other words, our set of all possible real-world probability measures includes all the "equivalent martingale measures" for our market, corresponding to bounded "kernels" $\nu(\cdot)$. This way, under a possible real-world probability measure $P_\nu \in \mathcal{P}$, the model $\mathcal{M}$ of (2.1) becomes

$$\begin{aligned} dS_0(t) &= S_0(t)r(t)dt, \quad S_0(0) = 1, \\ dS_i(t) &= S_i(t)\left[(r(t) - \nu_i(t))dt + \sum_{j=1}^{d} \sigma_{ij}(t)dW_\nu^j(t)\right], \\ S_i(0) &= s_i \in (0,\infty) \ ; \quad i = 1, \ldots, d, \end{aligned} \tag{4.7}$$

in the notation of (2.15). The resulting modified model $\mathcal{M}_\nu$ is similar to that of (2.1); now $W_\nu(\cdot)$ plays the role of the driving Brownian motion (under P_ν), but the stock return rates are different for different "model measures" P_ν.

The following theorem shows that, if the uncertainty about the real-world probability measure is large enough (in the sense that all equivalent martingale measures corresponding to bounded kernels are possible candidates for the real-world measure), then the optimal thing to do in order to minimize the expected risk of hedging a claim C in the market is the following: *borrow exactly as much money from the bank as is needed to hedge C.*

THEOREM 4.1. *Under the above assumptions we have*

$$\overline{V}(x) = \underline{V}(x) = C(0) - x. \tag{4.8}$$

In other words, the stochastic game defined by (4.1) *and* (4.2) *has a value that is equal to the expected loss of the strategy which borrows $C(0) - x$ from the bank and then invests according to the least expensive strategy for hedging the claim C.*

Proof. Let (π^*, κ^*) be the strategy from the statement of the theorem, namely the one for which we have

$$\bar{X}^*(t) := \bar{X}^{x,\pi^*,\kappa^*}(t) = \bar{C}(t) - (C(0) - x), \ P - a.s.,$$

in the notation of (3.2). Such a strategy exists by Theorem 3.2. It is clear that (4.3) is then satisfied, so that $(\pi^*, \kappa^*) \in \mathcal{A}(x)$. Since for this strategy $E^Q[\bar{C} - \bar{X}^{x,\pi^*,\kappa^*}(T)]^+ = C(0) - x$ for all $Q \in \mathcal{P}$, it also follows that

$$\overline{V}(x) \leq C(0) - x. \tag{4.9}$$

On the other hand, we have here $\delta(\nu) = 0$ for $\nu \in \tilde{K}$, so that $H_\nu(\cdot) = Z_\nu(\cdot)$ for $\nu \in \mathcal{D}$, and Ito's rule gives, in analogy to (2.7) and in the notation of (2.15),

$$d\bar{X}^*(t) = -e^{-\int_0^t r(u)du} d\kappa^*(t) + (\pi^*(t))'\sigma(t)\bar{X}^*(t)dW_\nu(t) \ ; \quad X^*(0) = x \tag{4.10}$$

for all $\nu \in \mathcal{D}$, since $\nu'(\cdot)\pi^*(\cdot) \equiv 0$. Therefore, $\bar{X}^*(\cdot)$ is a P_ν-local supermartingale bounded from below, thus also a P_ν-supermartingale, by Fatou's lemma. Consequently,

$$E_\nu[\bar{X}^*(T)] \leq x \quad \forall\, \nu \ \in \ \mathcal{D}, \tag{4.11}$$

where E_ν is the expectation under P_ν measure. Since $P_\nu \in \mathcal{P}$ for all $\nu \in \mathcal{D}$, (4.11) and Jensen's inequality imply

$$\underline{V}(x) \geq \sup_{\nu\in\mathcal{D}} \inf_{(\pi,\kappa)\in\mathcal{A}(x)} (E_\nu[\bar{C}] - x)^+ = \sup_{\nu\in\mathcal{D}} (E_\nu[\bar{C}] - x)^+ = C(0) - x. \tag{4.12}$$

Since $\underline{V}(x) \leq \overline{V}(x)$, (4.8) is a consequence of (4.9) and (4.12). □

Acknowledgments. I wish to thank Ioannis Karatzas for suggesting the use of Komlós theorem and for providing me with the reference Schwartz (1986), as well for thorough readings and helpful comments on the paper. Thanks are also due to the two anonymous referees for useful suggestions and comments.

REFERENCES

Ph. Artzner, F. Delbaen, J. M. Eber, and D. Heath (1999), *Coherent measures of risk*, Math. Finance, 9, pp. 203–228.

J. Aubin and I. Ekeland (1984), *Applied Nonlinear Analysis*, John Wiley and Sons, New York.

J. Blumenstein (1999), *Minimizing Expected Loss of Hedging in Discrete Time*, Doctoral thesis, Department of Mathematics, Columbia University, New York.

M. Broadie, J. Cvitanić, And H. M. Soner (1998), *Optimal replication of contingent claims under portfolio constraints*, The Review of Financial Studies, 11, pp. 59–79.

J. Cvitanić (1997), *Optimal trading under constraints*, in Financial Mathematics, Lecture Notes in Math. 1656, W. J. Runggaldier, ed., Springer, Berlin, pp. 123–190.

J. Cvitanić And I. Karatzas (1992), *Convex duality in constrained portfolio optimization*, Ann. Appl. Probab., 2, pp. 767–818.

J. Cvitanić and I. Karatzas (1993), *Hedging contingent claims with constrained portfolios*, Ann. Appl. Probab., 3, pp. 652–681.

J. Cvitanić and I. Karatzas (1999), *On dynamic measures of risk*, Finance and Stochastics, 4, pp. 451–482.

J. Cvitanić, H. Pham, and N. Touzi (1999), *Super-replication in stochastic volatility models under portfolio constraints*, J. Appl. Probab., 36, pp. 523–545.

R. Dembo (1997), *Optimal Portfolio Replication*, preprint, Algorithmics Inc., Toronto, Canada.

H. Föllmer and P. Leukert (1999a), *Quantile hedging*, Finance and Stochastics, 3, pp. 251–274.

H. Föllmer and P. Leukert (1999b), *Partial hedging with minimal shortfall risk*, Finance and Stochastics, to appear.

I. Karatzas and S. Kou (1996), *On the pricing of contingent claims under constraints*, Ann. Appl. Probab., 6, pp. 321–369.

I. Karatzas and S. Shreve (1998), *Methods of Mathematical Finance*, Appl. Math. 39, Springer-Verlag, New York.

D. Kramkov and W. Schachermayer (1999), *The asymptotic elasticity of utility functions and optimal investment in incomplete markets*, Ann. Appl. Probab., 9, pp. 904–950.

H. Pham (1998), *Dynamic L^p-Hedging in Discrete Time under Cone Constraints*, preprint, Universitè Marne-la-Vallée, France.

H. Pham, T. Rheinländer, and M. Schweizer (1998), *Mean-variance hedging for continuous processes: new proofs and examples*, Finance and Stochastics, 2, pp. 173–198.

M. Schwartz (1986), *New proofs of a theorem of Komlós*, Acta Math. Hungar., 47, pp. 181–185.

SIAM J. CONTROL OPTIM.
Vol. 38, No. 4, pp. 1067–1085

DISCONTINUOUS SOLUTIONS OF THE HAMILTON–JACOBI EQUATION FOR EXIT TIME PROBLEMS*

J. J. YE†

Abstract. In general, the value function associated with an exit time problem is a discontinuous function. We prove that the lower (upper) semicontinuous envelope of the value function is a supersolution (subsolution) of the Hamilton–Jacobi equation involving the proximal subdifferentials (superdifferentials) with subdifferential-type (superdifferential-type) mixed boundary condition. We also show that if the value function is upper semicontinuous, then it is the maximum subsolution of the Hamilton–Jacobi equation involving the proximal superdifferentials with the natural boundary condition, and if the value function is lower semicontinuous, then it is the minimum solution of the Hamilton–Jacobi equation involving the proximal subdifferentials with a natural boundary condition. Futhermore, if a compatibility condition is satisfied, then the value function is the unique lower semicontinuous solution of the Hamilton–Jacobi equation with a natural boundary condition and a subdifferential type boundary condition. Some conditions ensuring lower semicontinuity of the value functions are also given.

Key words. Hamilton–Jacobi equation, dynamic programming principle, exit time problems, proximal subdifferentials

AMS subject classifications. 49L05, 49L20

PII. S0363012997326234

1. Introduction.

In this paper we study the exit time problem (also called the control problem with a boundary condition as in [10]). In its simplest form, the exit time problem involves a given open set E in R^n, and asks for choices for the time $t^* \geq 0$ and the measurable function u on $[0, t^*)$ which will

$$\begin{aligned} \text{minimize} \quad & J(x,u) := \int_0^{t^*} e^{-\lambda s} f(y(s), u(s)) ds + e^{-\lambda t^*} h(y(t^*)) \\ \text{subject to (s.t.)} \quad & \dot{y}(t) = g(y(t), u(t)) \quad \text{a.e. } t \in [0, t^*], \\ & u(t) \in U \quad \text{a.e. } t \in [0, t^*), \\ & y(0) = x, y(t) \in E, 0 \leq t < t^*, y(t^*) \notin E. \end{aligned}$$

By the classical Hamilton–Jacobi (H–J) theory (or the so-called dynamic programming theory), if the value function V is continuously differentiable, then it is the unique solution of the following H–J equation:

$$\lambda V(x) + H(x, -\nabla V(x)) = 0 \quad \forall x \in E, \tag{1}$$

where the Hamiltonian $H(x,p) := \max\{p \cdot g(x,u) - f(x,u) : u \in U\}$, with the natural boundary condition

$$V(x) = h(x) \quad \forall x \in \partial E.$$

Due to the complicated behavior of the trajectories at the boundary of the state space, the value function for the exit time problem is in general discontinuous, even

*Received by the editors August 22, 1997; accepted for publication (in revised form) July 8, 1999; published electronically April 4, 2000. This work was supported by NSERC and a University of Victoria internal research grant.

http://www.siam.org/journals/sicon/38-4/32623.html

†Department of Mathematics and Statistics, University of Victoria, P.O. Box 3045, MS 7437, Victoria, BC, Canada V8W 3P4 (janeye@math.uvic.ca).

if all the problem data are Lipschitz continuous, unless some nontangency condition is imposed on the boundary (see, e.g., [12, 23, 10] for the Lipschitz continuity of the value function). Solving the H–J equation (1) with appropriate boundary conditions in some nonclassical sense has become an active research area. Gonzalez and Rofman [13] proved that the value function is an upper bound of a suitable set of subsolutions of the H–J equation. Dempster and Ye [10] characterized the Lipschitz value function as a solution of the H–J equation involving the Clarke generalized gradient. Bardi and Soravia [2], Barles and Perthame [4, 5], Blanc [6], Ishii [14], and Soravia [17, 18] have studied the solution of the H–J equation (1) with various boundary conditions in the framework of the viscosity solutions first introduced by Crandall and Lions [9] for continuous functions and later defined for discontinuous functions by Ishii [14, 15] and modified by Barron and Jenson [3] for the case of convex Hamiltonians. The reader is also referred to the recent monograph of Bardi and Capuzzo-Dolcetta [1] for the history and the recent development of the H–J equation using the viscosity approach.

Under assumptions that reduce the exit time problem to a generalized optimal stopping time problem, Ye and Zhu [24] showed that the value function of the exit time problem with relaxed controls is the unique lower semicontinuous solution of the H–J equation with the usual gradient replaced by the proximal subdifferential $\partial_p V(x)$ (see Definition 2.1) with the natural boundary condition

$$V(x) = h(x) \ \forall x \in E^c,$$

where E^c denotes the complement of the state space E and the subdifferential type boundary condition, i.e.,

$$\lambda V(x) + H(x, -\partial_p V(x)) \leq 0 \quad \forall x \in \partial E.$$

The purpose of this paper is to extend the H–J theory using the equivalence between the invariance and the H–J equation to treat exit time problems under assumptions that are much more general than those in [24]. In particular, we allow the discount rate λ to be zero and the exit cost h to be unbounded. In Theorem 2.2 we show that the lower (upper) semicontinuous envelope of the value function is a supersolution (subsolution) of the H–J equation involving the proximal subdifferentials (superdifferentials) with subdifferential-type (superdifferential-type) mixed boundary condition. In Theorems 2.3 and 2.4 we show that if the value function is upper semicontinuous, then it is the maximum subsolution of the H–J equation involving the proximal superdifferentials with the natural boundary condition, and if the value function is lower semicontinuous, then it is the minimum solution of the H–J equation involving the proximal subdifferentials with a natural boundary condition. Some conditions ensuring lower semicontinuity of the value functions are given in Proposition 2.5.

The technique of treating semicontinuous solutions to the H–J equation by using equivalence between the invariance property and the H–J equation was first introduced by Subbotin [19] for differential games (see also Subbotin [20]) and has been used in [8, 11] for finite horizon problems and in [22] for minimal time problems. The equivalence of the various concepts of the solution to the H–J equation in an open set was also given in [8].

We arrange the paper as follows: In the next section we state the problem formulation for the exit time problem and our main results. In section 3 we establish the equivalence among the optimality principle, the invariance property, and the H–J equations. The proofs of the main results are contained in section 4.

2. The exit time problems and the H–J equation. Let U be a compact subset of R^m and Prob(U) the set of all Borel probability measures on U. Consider Prob(U) as a subset of the dual of $C(U)$ endowed with the weak star topology, where $C(U)$ is the Banach space of continuous functions on U with the supremum norm. For any $\phi \in C(U)$ and $u \in$ Prob(U), we denote the pairing of ϕ and u by $\phi(u) := \int_U \phi(r)u(dr)$. Let $\mathcal{U}$ be the set of all Lebesgue measurable mappings from R to Prob(U). For finite real numbers $a < b$, define $\mathcal{U}_{[a,b]} := \{u|_{[a,b]} : u \in \mathcal{U}\}$. Then $\mathcal{U}_{[a,b]}$ is a weak star compact subset of $L^1([a,b]; C(U))^*$. We endow $\mathcal{U}$ with the following topology: u^n converges to u in $\mathcal{U}$ provided that $u^n|_{[a,b]}$ converges to $u|_{[a,b]}$ in $\mathcal{U}_{[a,b]}$ for any finite real numbers $a < b$. The set $\mathcal{U}_{[a,b]}$ is the collection of relaxed control functions defined in Warga [21]. It is the compactification of the set of usual control functions in the weak star topology of $L^1([a,b]; C(U))^*$. Elements of $\mathcal{U}_{[a,b]}$ are called relaxed controls. Using the set of relaxed controls ensures the existence of the optimal solution and also ensures the convexity of the velocity set so that the invariance theorems can be used. Any relaxed control can be approximated by usual controls. We refer to [21] for more details.

Let the state space E be an open subset of R^d, $\bar{E}$ be the closure of E, and O be an open set containing $\bar{E}$. Assume that $g : O \times U \to R^d$ satisfies the following condition.

(H1) $g(x,u)$ is continuous, bounded, and Lipschitz in x uniformly in $u \in U$.

Under such a condition, for each $x \in O$ and $u \in \mathcal{U}$, the differential equation

$$\dot{y}(s) = g(y(s), u(s)) \quad \text{a.e.}$$

has a unique solution defined on R that satisfies the side condition $y(0) = x$. We denote this solution by $y[x,u](\cdot)$ to indicate its dependence on x and u.

For each initial state $x \in E$ and control function u, define the exit time $t^*[x,u]$ to be the first time the trajectory starting from $x \in E$ corresponding to the control u exits from the state space E, or infinity if it never exits the state space; i.e.,

$$t^*[x,u] := \inf\{t > 0 : y[x,u](t) \notin E\},$$

where $\inf \emptyset = \infty$ by convention. For any $x \in E^c$, we define $t^*[x,u] := 0$. Where there is no confusion, we will simply use t^* instead of $t^*[x,u]$.

Let $\lambda \geq 0$ be the discount rate. Consider the following exit time problem:

$$P_x \quad \text{minimize} \quad J(x,u) := \int_0^{t^*} e^{-\lambda s} f(y[x,u](s), u(s))ds + e^{-\lambda t^*} h(y[x,u](t^*))$$
$$\text{s.t.} \quad u \in \mathcal{U}.$$

We state some further basic assumptions:

(H2) The running cost $f(x,u) : O \times U \to R^d$ is continuous, bounded, and Lipschitz in x uniformly in $u \in U$. The exit cost $h(x) : O \to R$ is lower semicontinuous. Furthermore, when $t^*[x,u] = \infty$ the integral $\int_0^{t^*} e^{-\lambda s} f(y[x,u](s), u(s))ds$ converges and the limit $e^{-\lambda\infty} h(y[x,u](\infty)) := \lim_{r\to\infty} e^{-\lambda r} h(y[x,u](r))$ exists and is finite.

REMARK 1. *The exit time problem we consider in this paper is more general than that usually considered in the literature (see, e.g.,* [1, 6]*) in that we allow the discount rate λ to be zero and the exit cost h to be unbounded. Notice that under the assumption that f is bounded, the integral $\int_0^\infty e^{-\lambda s} f(y[x,u](s), u(s))ds$ converges*

automatically for the case $\lambda > 0$ *so the assumption* (H2) *is mainly for the case when* $\lambda = 0$.

Under our assumptions, it is known that there exists an optimal control for the exit time problem for each $x \in E$. Define the value function of the family of problems P_x as

$$V(x) := \min \left\{ \int_0^{t^*} e^{-\lambda s} f(y[x,u](s), u(s)) ds + e^{-\lambda t^*} h(y[x,u](t^*)) : u \in \mathcal{U} \right\}.$$

Unlike a standard free end point optimal control problem whose value function is continuous if the terminal cost is continuous, the value function for the exit time problem is in general discontinuous even in the case where the terminal cost h is smooth. To see this we examine two simple examples.

EXAMPLE 1. *Let* $E = (0,1)$ *be the state space and the control set* $U = \{-1\}$. *Consider the following exit time problem where* $f(x,u) \equiv 0, g(x,u) = u, h(x) = x$, $\lambda = 0$:

$$\begin{aligned} \min \quad & y(t^*) \\ s.t. \quad & \dot{y} = u, u(t) = -1, \\ & y(0) = x. \end{aligned}$$

It is easy to see that the value function

$$V(x) = \begin{cases} 0 & \text{if } x \in [0,1), \\ x & \text{if } x \notin [0,1) \end{cases}$$

is upper semicontinuous with discontinuity at $x = 1$.

EXAMPLE 2. *In Example* 1, *change the control set to* $U = \{1\}$. *Then the value function becomes*

$$V(x) = \begin{cases} 1 & \text{if } x \in (0,1], \\ x & \text{if } x \notin (0,1], \end{cases}$$

which is lower semicontinuous with discontinuity at $x = 0$.

In order to see the connections between the value function and the H–J equations we define the lower and upper semicontinuous envelopes of a function $W : O \to R$ as

$$W_*(x) := \liminf_{y \to x} W(y)$$

and

$$W^*(x) := \limsup_{y \to x} W(y),$$

respectively. Then it is easy to see that W_* is lower semicontinuous and W^* is upper semicontinuous.

We will use the concept of proximal subdifferentials (superdifferentials) for any lower (upper) semicontinuous functions defined as follows.

DEFINITION 2.1 (see, e.g., Clarke [7] and Loewen [16]). *Let* $\phi : R^d \to (-\infty, \infty]$ *be an extended-valued lower semicontinuous function. The proximal subdifferential of* ϕ *at* $x \in R^d$ *where* $\phi(x) \neq \infty$ *is a set-valued map defined by*

$$\begin{aligned} \partial_p \phi(x) := \{ \xi \in R^d : & \exists \sigma > 0, \delta > 0 \\ & s.t.\ \phi(y) \geq \phi(x) - \sigma \|y - x\|^2 + \langle \xi, y - x \rangle \quad \forall y \in x + \delta B \}, \end{aligned}$$

where $\langle a, b\rangle$ denotes the inner product of the vectors a and b and B denotes the open unit ball. Let $\phi : R^d \to [-\infty, \infty)$ be an extended-valued upper semicontinuous function. The proximal superdifferential of ϕ at x where $\phi(x) \neq \infty$ is defined by

$$\partial^p \phi(x) := -\partial_p(-\phi)(x),$$

i.e.,

$$\partial^p \phi(x) := \{\xi \in R^d : \exists \sigma > 0, \delta > 0 \quad \text{s.t. } \phi(y) \leq \phi(x) + \sigma\|y - x\|^2 + \langle \xi, y - x\rangle \quad \forall x + \delta B\}.$$

REMARK 2. *Since the function $y \to \phi(x) - \sigma\|y - x\|^2 + \langle \xi, y - x\rangle$ in the right-hand side of the inequality in the definition of the proximal subdifferential is a quadratic, it is easy to see that $\xi \in \partial_p \phi(x)$ if and only if there is a parabola fitting under the epigraph of ϕ at $(x, \phi(x))$ with ξ as the slope of ϕ at x. Hence, in the case where there does not exist a parabola fitting under the epigraph of ϕ at $(x, \phi(x))$, the proximal subdifferential of ϕ at x may be empty (e.g., $\phi(x) = -|x|$ has an empty proximal subdifferential at 0). Similarly the proximal superdifferential may be empty. However, we shall see later that the emptiness of the proximal subdifferential (superdifferential) is actually an advantage.*

We now state our main results. The first result gives the connection between the semicontinuous envelopes of the value function and the H–J inequalities.

THEOREM 2.2. *Under assumptions* (H1)–(H2) *the lower semicontinuous envelope of the value function $V_*(x)$ is a supersolution of the H–J equation involving the proximal subdifferentials (in E), i.e.,*

$$\lambda V_*(x) + H(x, -\partial_p V_*(x)) \geq 0 \quad \forall x \in E \tag{2}$$

with the subdifferential-type mixed boundary condition

$$\max\{V_*(x) - h(x), \lambda V_*(x) + H(x, -\partial_p V_*(x))\} \geq 0 \quad \forall x \in \partial E, \tag{3}$$

and the upper semicontinuous envelope of the value function $V^*(x)$ is a subsolution of the H–J equation involving the proximal superdifferentials (in E), i.e.,

$$\lambda V^*(x) + H(x, -\partial^p V^*(x)) \leq 0 \quad \forall x \in E, \tag{4}$$

with the superdifferential-type mixed boundary condition

$$\min\{V^*(x) - h^*(x), \lambda V^*(x) + H(x, -\partial^p V^*(x))\} \leq 0 \quad \forall x \in \partial E, \tag{5}$$

where ∂E denotes the boundary of E.

REMARK 3. *Equation* (2) *should be understood in the following sense: At any point $x \in E$ where $\partial_p V_*(x) \neq \emptyset$,*

$$\lambda V_*(x) + H(x, -\xi) \geq 0 \quad \forall \xi \in \partial_p V_*(x).$$

Hence the points x where $\partial_p V_*(x) = \emptyset$ can be neglected. Equation (4) *is understood in a similar way. Equation* (3) *means that if $x \in \partial E$ is a point where $V_*(x) < h(x)$ and $\partial_p V_*(x) \neq \emptyset$, then*

$$\lambda V_*(x) + H(x, -\xi) \geq 0 \quad \forall \xi \in \partial_p V_*(x).$$

Similarly, (5) *means that if* $x \in \partial E$ *is a point where* $V^*(x) > h^*(x)$ *and* $\partial^p V^*(x) \neq \emptyset$, *then*

$$\lambda V^*(x) + H(x, -\xi) \leq 0 \quad \forall \xi \in \partial^p V^*(x).$$

Note that a similar result was given in Theorem 2.9 *of Blanc* [6] *in the viscosity solution sense for the case* $\lambda > 0$ *and bounded exit cost* h. *In general, as in Remark* 2.7 *of Blanc* [6], *we do not expect to have a unique function that satisfies* (2)–(5).

When the value function has a semicontinuity property, the following two theorems give connections between the value function (instead of its semicontinuous envelopes) and the H–J equation with natural boundary condition (instead of the mixed boundary condition).

THEOREM 2.3. *In additions to assumptions* (H1)–(H2), *assume that the value function is upper semicontinuous. Then it is the maximum upper semicontinuous function that is a subsolution of the H–J equation involving the proximal superdifferentials* (*in* E), *i.e.,*

$$\lambda W(x) + H(x, -\partial^p W(x)) \leq 0 \quad \forall x \in E,$$

with the natural boundary condition

$$W(x) = h(x) \quad \forall x \in \partial E.$$

THEOREM 2.4. *In additions to assumptions* (H1)–(H2), *assume that the value function is lower semicontinuous. Then it is the minimum lower semicontinuous solution of the H–J equation involving the proximal subdifferentials* (*in* E), *i.e.,*

$$\lambda W(x) + H(x, -\partial_p W(x)) = 0 \quad \forall x \in E,$$

with the natural boundary condition

$$W(x) = h(x) \quad \forall x \in \partial E.$$

We now give some conditions which ensure lower semicontinuity of the value function. First we state the required assumptions.

(A) $\forall\ x \in \partial E$ and $u \in \mathcal{U}$ such that $y[x,u](t) \in \bar{E} \quad \forall t \in [0,\infty)$, the limit $\lim_{r\to\infty} e^{-\lambda r} h(y[x,u](r))$ exists. Also, $\forall x \in \partial E$, all controls $u \in \mathcal{U}$ and $r \geq 0$ such that $y[x,u](t) \in \bar{E} \forall t \in [0,r], y[x,u](r) \in \partial E$, or $r = \infty$,

$$h(x) \leq \int_0^r e^{-\lambda s} f(y[x,u](s), u(s)) ds + e^{-\lambda r} h(y[x,u](r)).$$

PROPOSITION 2.5. *In addition to assumptions* (H1)–(H2), *if* (A) *is satisfied, then the value function* $V(x)$ *is lower semicontinuous on* $\bar{E}$.

Proof. Let $x \in \bar{E}$ and $x_n \in \bar{E}, x_n \to x$. By definition of the value function for each n, there exists $u^n \in \mathcal{U}$ and $t^*[x_n, u_n] := t_n^*$ such that

$$V(x_n) = \int_0^{t_n^*} e^{-\lambda s} f(y[x_n,u_n](s), u_n(s)) ds + e^{-\lambda t_n^*} h(y[x_n,u_n](t_n^*)). \tag{6}$$

We now consider two cases.

Case 1. The sequence $\{t_n^*\}$ is bounded. Without loss of generality we may assume that t_n^* converges to r, $t_n^* \in [0, r+1]$ and $u_n|_{[0,r+1]}$ converges to $u \in \mathcal{U}_{[0,r+1]}$ in the

topology of $\mathcal{U}_{[0,r+1]}$. Then $y[x_n,u_n](s)$ uniformly converges to $y[x,u](s)$ in $[0,r+1]$. Since $\bar{E}$ and ∂E are closed, $y[x,u](t)\in\bar{E}\ \forall t\in[0,r]$ and $y[x,u](r)\in\partial E$. Taking lim inf in (6) when $n\to\infty$ yields by virtue of lower semicontinuity of h

$$\begin{aligned}\liminf_{n\to\infty} V(x_n) &\geq \int_0^r e^{-\lambda s} f(y[x,u](s),u(s))ds + e^{-\lambda r} h(y[x,u](r))\\ &= \int_0^{t^*} e^{-\lambda s} f(y[x,u](s),u(s))ds\\ &\quad + \int_{t^*}^r e^{-\lambda s} f(y[x,u](s),u(s))ds + e^{-\lambda r} h(y[x,u](r))\\ &\geq \int_0^{t^*} e^{-\lambda s} f(y[x,u](s),u(s))ds + e^{-\lambda t^*} h(y[x,u](t^*))\\ &\qquad\qquad \text{(by assumption (A))}\\ &\geq V(x).\end{aligned}$$

Case 2. The sequence $\{t_n^*\}$ is unbounded. Without loss of generality we may assume that $t_n^*\to\infty$. For each integer $m>0$, consider the restriction of u_n to $[0,m]$. Since $\mathcal{U}_{[0,m]}$ is compact we can extract a convergent subsequence from $\{u_n|_{[0,m]}\}$. Using the diagonal method we can choose a subsequence $\{u_{n_i}\}$ of $\{u_n\}$ and an element $u\in\mathcal{U}$ such that $u_{n_i}|_{[0,m]}$ converges to $u|_{[0,m]}$ in $\mathcal{U}|_{[0,m]}$ for any m. We may assume that

$$\lim_{i\to\infty} V(x_{n_i}) = \liminf_{n\to\infty} V(x_n).$$

Taking lim inf in (6) when $n_i\to\infty$ yields in the case $t^*[x,u]=\infty$ that

$$\begin{aligned}\liminf_{i\to\infty} V(x_{n_i}) &\geq \int_0^{\infty} e^{-\lambda s} f(y[x,u](s),u(s))ds + e^{-\lambda\infty} h(y[x,u](\infty))\\ &\geq V(x)\end{aligned}$$

and in the case $t^*[x,u]<\infty$ that

$$\begin{aligned}\liminf_{i\to\infty} V(x_{n_i}) &\geq \int_0^{\infty} e^{-\lambda s} f(y[x,u](s),u(s))ds + e^{-\lambda\infty} h(y[x,u](\infty))\\ &= \int_0^{t^*} e^{-\lambda s} f(y[x,u](s),u(s))ds\\ &\quad + \int_{t^*}^{\infty} e^{-\lambda s} f(y[x,u](s),u(s))ds + e^{\lambda\infty} h(y[x,u](\infty))\\ &\geq \int_0^{t^*} e^{-\lambda s} f(y[x,u](s),u(s))ds + e^{-\lambda t^*} h(y[x,u](t^*))\\ &\qquad\qquad \text{(by assumption (A))}\\ &\geq V(x).\end{aligned}$$

The proof of the proposition is complete. □

REMARK 4. *In its essentials the above result says that if we allow the trajectory to continue after the first exit time but assume that it is cheaper to stop at the first exit time than to continue until the trajectory reaches the boundary again, then the value function is lower semicontinuous. The assumption* (A) *is not satisfied at the boundary*

point $x = 1$ for Example 1 since the only r in this case is $r = 1$ and the trajectory starting at $x = 1$ using the only control $u = -1$ reaches the boundary state $x = 0$ at time $r = 1$, but $1 = h(1) \not\leq h(y[1, u](1)) = 0$. On the other hand, the assumption (A) *is satisfied by Example 2 and hence the value function is lower semicontinuous.*

Combining Theorem 2.4 and Proposition 2.5, we have the following corollary.

COROLLARY 2.6. *Under assumptions* (H1), (H2), *and* (A), *the value function is the minimum lower semicontinuous solution of the H–J equation involving the proximal subdifferentials (in E), i.e.,*

$$\lambda W(x) + H(x, -\partial_p W(x)) = 0 \quad \forall x \in E, \tag{7}$$

with the natural boundary condition

$$W(x) = h(x) \quad \forall x \in \partial E. \tag{8}$$

One may wonder whether the natural boundary condition (8) is enough for the uniqueness of the solution to (7). The following example gives a negative answer.

EXAMPLE 3. *Let $E = R^2 \backslash \{0\}$ be the state space and the control set $U = [-1, 1]$. Consider the exit time problem where $f(x, u) \equiv 1, g(x, u) = (u, 0), h(x) = 0, \lambda = 1$. It is easy to see that*

$$V(x) = \begin{cases} \int_0^{|x_1|} e^{-s} ds = 1 - e^{-|x_1|} & \text{if } x_2 = 0, \\ \int_0^{\infty} e^{-s} ds = 1 & \text{if } x_2 \neq 0, \end{cases}$$

$$H(x, p) = \max\{p_1 u : u \in [-1, 1]\} - 1 = |p_1| - 1,$$

and

$$\partial_p V(x_1, 0) = \begin{cases} (e^{-x_1}, 0) & \text{if } x_1 > 0, \\ (-e^{x_1}, 0) & \text{if } x_1 < 0, \\ [-1, 1] \times \{0\} & \text{if } x_1 = 0. \end{cases}$$

Hence the value function is a lower semicontinuous solution of the H–J equation (7) *with the natural boundary condition* (8). *However, the function $W(x) = 1$ if $x \neq 0$ and $W(0) = 0$ is also a lower semicontinuous solution of* (7), (8). *Indeed, by Corollary 2.6, the value function is the minimum solution of the H–J equation* (7) *with the natural boundary condition $V(0) = 0$.*

The above example shows that the natural boundary condition (8) may not be enough to ensure the uniqueness of the solution to the H–J equation involving the proximal subdifferentials (7). However, V satisfies the subdifferential-type boundary condition

$$\lambda V(0) + H(x, -\partial_p V(0)) \leq 0$$

while $W(x)$ does not satisfy the above boundary condition. (Note $\partial_p W(0) = R^2$.) We now give a compatibility condition stronger than assumption (A) under which the value function is not only lower semicontinuous but also a unique lower semicontinuous solution to the H–J equation involving the proximal subdifferentials with the natural boundary condition and the subdifferential-type boundary condition.

(H3) $\forall\, x \in O \setminus E$, all controls $u \in \mathcal{U}$

$$h(x) \leq \int_0^r e^{-\lambda s} f(y[x, u](s), u(s)) ds + e^{-\lambda r} h(y[x, u](r)) \quad \forall 0 \leq r < \tau^*, \tag{9}$$

where

$$\tau^* := \inf\{t > 0 : y[x,u](t) \notin O \setminus E\}$$

and when

$$\tau^* = \infty, \int_0^{\tau^*} e^{-\lambda s} f(y[x,u](s), u(s))ds$$

converges and the limit $\lim_{r\to\infty} e^{-\lambda r} h(y[x,u])(r))$ exists and is finite.

REMARK 5. *When $E = R^d$ and $h = 0$ the exit time problem becomes an infinite horizon problem and assumption* (H3) *is satisfied vacuously. If $f \geq 0$ and $h = 0$, then assumption* (H3) *is also satisfied. This includes the time optimal control problem since solving a time optimal control problem is equivalent to solving an exit time problem with $f = 1$ and $h = 0$. Note that* (H3) *is a local version of the assumption* (5.8) *in Bardi and Capuzzo-Dolcetta* [1] *and* (H3′) *in Ye and Zhu* [24].

The statement of the following theorem is known from Corollary 4.5 of Ye and Zhu [24] for the case where $\lambda > 0$ and h is bounded under the assumption that (9) is satisfied globally $\forall x \in R^d$. However, the proof we give here is independent and different.

THEOREM 2.7. *Under assumptions* (H1)–(H3), *the value function $V(x)$ is a unique lower semicontinuous solution of the H–J equation involving the proximal subdifferentials (in E), i.e.,*

$$\lambda V(x) + H(x, -\partial_p V(x)) = 0 \quad \forall x \in E$$

with the natural boundary condition

$$V(x) = h(x) \qquad \forall x \in O \setminus E$$

and the subdifferential-type boundary condition, i.e.,

$$\lambda V(x) + H(x, -\partial_p V(x)) \leq 0 \forall x \in \partial E.$$

REMARK 6. *Note that a similar result was proved in Theorem* 5.5 *of Bardi and Capuzzo-Dolcetta* [1] *for the lower semicontinuous envelope of the value function in the viscosity solution sense in the case where $\lambda > 0$ and h is bounded under the assumption that* (9) *is satisfied globally $\forall x \in R^d$.*

3. Optimality principle, invariance, and the H–J equation.

DEFINITION 3.1. *Let $W(x) : G :\to R$ where $G \subseteq R^d$ is an open set. We say that $W(x)$ satisfies*

(a) *the* superoptimality principle *in G if and only if $\forall x \in G$ there exists a control $u \in \mathcal{U}$ such that*

$$W(x) \geq \int_0^{\tau} e^{-\lambda s} f(y[x,u](s), u(s))ds + e^{-\lambda\tau} W(y[x,u](\tau)) \quad \forall 0 \leq \tau < \tau^*;$$

(b) *the* suboptimality principle *in G if and only if $\forall x \in G$ and $\forall u \in \mathcal{U}$,*

$$W(x) \leq \int_0^{\tau} e^{-\lambda s} f(y[x,u](s), u(s))ds + e^{-\lambda\tau} W(y[x,u](\tau)) \qquad \forall 0 \leq \tau < \tau^*.$$

Here $\tau^* := \inf\{t > 0 : y[x,u](t) \notin G\}$ *is the exit time from set* G.

The following facts are well known from the Bellman optimality principle.

PROPOSITION 3.2. *The value function* $V(x)$ *satisfies the superoptimality principle in* E *and suboptimality principle in* E.

Furthermore, the following results indicate that not only the value function but also its lower (upper) semicontinuous envelope satisfies the superoptimality (suboptimality) principle in E.

PROPOSITION 3.3. *The lower semicontinuous envelope of the value function* $V_*(x)$ *satisfies the superoptimality principle in* E *and the upper semicontinuous envelope of the value function* $V^*(x)$ *satisfies the suboptimality principle in* E.

Proof. Fix $x \in E$, and suppose $V_*(x) = \lim_{n\to\infty} V(x_n)$ where $\lim_{n\to\infty} x_n = x, x_n \in E$. Then by Proposition 3.2, there exists $u_n \in \mathcal{U}$ and $t^*[x_n, u_n] = t_n^*$ such that

$$V(x_n) \geq \int_0^\tau e^{-\lambda s} f(y[x_n,u_n](s), u_n(s))ds + e^{-\lambda\tau} V(y[x_n,u_n](\tau)) \quad \forall 0 \leq \tau < t_n^*.$$

Without loss of generality assume that $t_n^* \to t^*$ where t^* may be finite or infinity. Let $0 \leq \tau < t^*$. Hence, for n large enough $\tau \leq t_n^*$. Taking limits and using the compactness of relaxed controls, we find a control $u \in \mathcal{U}$ such that

$$V_*(x) \geq \int_0^\tau e^{-\lambda s} f(y[x,u](s), u(s))ds + e^{-\lambda\tau} V_*(y[x,u](\tau)) \quad \forall 0 \leq \tau < t^*.$$

Similarly we can prove that $V^*(x)$ satisfies the suboptimality principle. □

In the following proposition, we show that either semicontinuity on $\bar{E}$ and the optimality principle in E or the suboptimality principle in an open set containing $\bar{E}$ gives the comparison results.

PROPOSITION 3.4.

(a) *Suppose that* W *satisfies the superoptimality principle in* E. *If* W *is lower semicontinuous on* $\bar{E}$ *and*

$$W(x) \geq h(x) \quad \forall x \in \partial E,$$

then $W(x) \geq V(x) \quad \forall x \in \bar{E}$.

(b) *Suppose that* W *satisfies the suboptimality principle in* E. *If* W *is upper semicontinuous on* $\bar{E}$ *and*

$$W(x) \leq h(x) \quad \forall x \in \partial E,$$

then $W(x) \leq V(x) \quad \forall x \in \bar{E}$.

(b′) *Suppose that* W *satisfies the suboptimality principle in the open set* O *containing* $\bar{E}$ *and*

$$W(x) \leq h(x) \quad \forall x \in \partial E.$$

Then $W(x) \leq V(x) \quad \forall x \in \bar{E}$.

Proof. (a) Suppose that W satisfies the superoptimality in E and $W(x) \geq h(x)\ \forall x \in \partial E$. Then $\forall x \in E$ there exists $u \in \mathcal{U}$ such that $\forall \tau_n \in [0, t^*)$.

$$W(x) \geq \int_0^{\tau_n} e^{-\lambda s} f(y[x,u](s), u(s))ds + e^{-\lambda\tau_n} W(y[x,u](\tau)).$$

Without loss of generality, assume that $\tau_n \to t^*$. Taking limits in the above inequality, we have by the compactness of relaxed controls and the lower semicontinuity of the function W that

$$\begin{aligned} W(x) &\geq \int_0^{t^*} e^{-\lambda s} f(y[x,u](s), u(s))ds + e^{-\lambda t^*} W(y[x,u](t^*)) \\ &\geq \int_0^{t^*} e^{-\lambda s} f(y[x,u](s), u(s))ds + e^{-\lambda t^*} h(y[x,u](t^*)) \\ &\geq V(x). \end{aligned}$$

Similarly we can prove (b).

(b′) Now suppose W satisfies the suboptimality in O and $W(x) \leq h(x)\ \forall x \in \partial E$. If $x \in \partial E$, then $W(x) \leq h(x) = V(x)$. If $x \in E$, then $\forall u \in \mathcal{U}$ we have

$$W(x) \leq \int_0^{t^*} e^{-\lambda s} f(y[x,u](s), u(s))ds + e^{-\lambda t^*} W(y[x,u](t^*)) \tag{10}$$

$$\leq \int_0^{t^*} e^{-\lambda s} f(y[x,u](s), u(s))ds + e^{-\lambda t^*} h(y[x,u](t^*)). \tag{11}$$

Hence $W(x) \leq V(x)$. □

Definition 3.5 (see [22, Definition 3.1]). *Suppose $\Omega \subset R^n$ is nonempty, $\Theta \subset R^n$ is open, and $\Gamma : R^n \Longrightarrow R^n$ is a set-valued map.*

(a) *Then (Γ, Ω) is weakly invariant in Θ provided that $\forall x \in \Omega \cap \Theta$, there exists an absolutely continuous function $y(\cdot)$ that satisfies $\dot{y}(s) \in \Gamma(y(s))$ a.e., $y(0) = x$, and*

$$y(s) \in \Omega \quad \forall s \in [0, \tau^*).$$

(b) *Then (Γ, Ω) is strongly invariant in Θ provided that $\forall x \in \Omega \cap \Theta$ and for any absolutely continuous function $y(\cdot)$ that satisfies $\dot{y}(s) \in \Gamma(y(s))$ a.e., $y(0) = x$ one has*

$$y(s) \in \Omega \quad \forall s \in [0, \tau^*)$$

where $\tau^* := \inf\{t > 0 : y(t) \notin \Theta\}$.

Define

$$\begin{aligned} F(s,x,r) &:= \{(g(x,u), -e^{-\lambda s} f(x,u)) : u \in U\}, \\ \widetilde{F}(s,x,r) &:= \{(g(x,u), e^{-\lambda s} f(x,u)) : u \in U\}. \end{aligned}$$

We write $\{1\} \times F$ for the set-valued map defined as

$$(\{1\} \times F)(s,x,r) := \{(1, g(x,u), -e^{-\lambda s} f(x,u)) : u \in U\}.$$

Similarly, we define $\{1\} \times \widetilde{F}$ and $\{-1\} \times \{-F\}$. Let $W : G \to R$. We denote the epigraph of the function $e^{-\lambda t} W(x)$ by X_W, i.e.,

$$X_W := \{(t,x,r) : r \geq e^{-\lambda t} W(x)\}.$$

The following results show that the optimality principles are equivalent to the invariance properties.

Proposition 3.6 (equivalence of optimality principles and invariances). *Let G be an open set in R^d.*

(a) *A function W satisfies the superoptimality principle in G if and only if $(\{1\}\times F, X_W)$ is weakly invariant in $R\times G\times R$;*

(b) *A function W satisfies the suboptimality principle in G if and only if either $(\{1\}\times\{\widetilde{F}\}, X_{-W})$ is strongly invariant in $R\times G\times R$ or $(\{-1\}\times\{-F\}, X_W)$ is strongly invariant in $R\times G\times R$.*

Proof. Since the proof is straightforward by using definitions, we prove only the second part of (b). Let $(t,x,r)\in X_W\cap R\times G\times R$. Then $x\in G$ and $r\ge e^{-\lambda t}W(x)$. By suboptimality principle, we have $\forall u\in\mathcal{U}$,

$$e^{\lambda\tau}W(y[x,u](-\tau))\le\int_{-\tau}^{0}e^{-\lambda s}f(y[x,u](s),u(s))ds+W(x)\quad\forall 0\le\tau\le\tau^*,$$

where $\tau^*:=\inf\{t>0: y[x,u](-\tau)\notin G\}$. Let $z_0(\tau)=-\tau+t$, $z(\tau)=y[x,u](-\tau)$, $z_{d+1}(\tau)=r-\int_0^{-\tau}e^{-\lambda(t+s)}f(y[x,u](s),u(s))ds$. Then $z_0(0)=t, z(0)=y[x,u](0)=x, z_{d+1}(0)=r$, $(\dot z_0,\dot z,\dot z_{d+1})(s)\in(\{-1\}\times\{-F\})(z_0(s),z(s))$, and

$$\begin{aligned}z_{d+1}(\tau)&=r-\int_0^{-\tau}e^{-\lambda(t+s)}f(y[x,u](s),u(s))ds\\&\ge e^{-\lambda t}W(x)-\int_0^{-\tau}e^{-\lambda(t+s)}f(y[x,u](s),u(s))ds\\&=e^{-\lambda t}(W(x)-\int_0^{-\tau}e^{-\lambda s}f(y[x,u](s),u(s))ds)\\&\ge e^{-\lambda(t-\tau)}W(y[x,u](-\tau))\\&=e^{-\lambda z_0(\tau)}W(z(\tau)).\end{aligned}$$

That is, $(z_0,z,z_{d+1})(\tau)\in X_W\ \forall 0\le\tau<\tau^*$. So $(\{-1\}\times\{-F\},X_W)$ is strongly invariant in $R\times G\times R$. Conversely, we can show that if $(\{-1\}\times\{-F\},X_W)$ is strongly invariant in $R\times G\times R$, then W satisfies the suboptimality principle in G. □

In the case when the function satisfying the optimality principles has semicontinuity properties, the invariances can be described by the H–J equations in the following way.

PROPOSITION 3.7 (equivalence of invariances and the H–J equations). *Let G be an open subset in R^d.*

(a) *Let $W: G\to R$ be a lower semicontinuous function. Then $(\{1\}\times F, X_W)$ is weakly invariant in $R\times G\times R$ if and only if*

$$\lambda W(x)+H(x,-\partial_p W(x))\ge 0\quad\forall x\in G.$$

(b) *Let $W: G\to R$ be an upper semicontinuous function. Then $(\{1\}\times\widetilde{F}, X_{-W})$ is strongly invariant in $R\times G\times R$ if and only if*

$$\lambda W(x)+H(x,-\partial^p W(x))\le 0\quad\forall x\in G.$$

(b′) *Let $W: G\to R$ be a lower semicontinuous function. Then $(\{-1\}\times\{-F\},X_W)$ is strongly invariant in $R\times G\times R$ if and only if*

$$\lambda W(x)+H(x,-\partial_p W(x))\le 0\quad\forall x\in G.$$

The proof is based on the following lemmas. We denote $N^p_\Omega(x) = \partial_p \delta_\Omega(x)$, where δ_Ω is the indicator function of a set Ω defined by

$$\delta_\Omega(x) = \begin{cases} 0 & \text{if } x \in \Omega, \\ +\infty & \text{if } x \notin \Omega. \end{cases}$$

LEMMA 3.8 (see, e.g., [22, Theorem 3.1]). *Suppose that for each $x \in R^n$, $\Gamma(x)$ is not empty, convex, and compact, and the graph $gph\Gamma := \{(x,v) : v \in \Gamma(x)\}$ is closed in R^{2n}. Let $\Omega \subseteq R^n$ be closed and $\Theta \subseteq R^n$ be open.*

(a) *Then (Γ, Ω) is weakly invariant in Θ if and only if*

$$\min\{\langle v, \xi\rangle : v \in \Gamma(x)\} \le 0 \quad \forall x \in \Omega \cap \Theta, \xi \in N^p_\Omega(x),$$

where $N^p_\Omega(x)$ is the proximal normal cone to Ω at $x \in \Omega$ defined by

$$N^p_\Omega(x) := \{\xi \in R^n : \exists M > 0 \text{ s.t. } \langle \xi, x' - x\rangle \le M\|x' - x\|^2 \ \ \forall x' \in \Omega\}.$$

(b) *In addition, assume that Γ is Lipschitz continuous; i.e., for each compact subset $C \subset R^n$, there exists $K > 0$ so that*

$$\Gamma(x) \subset \Gamma(y) + K\|x - y\|B \quad \forall x, y \in C.$$

Then (Γ, Ω) is strongly invariant in Θ if and only if

$$\max\{\langle v, \xi\rangle : v \in \Gamma(x)\} \le 0 \quad \forall x \in \Omega \cap \Theta, \xi \in N^p_\Omega(x).$$

LEMMA 3.9. *Suppose that for each $x \in R^n$, $\Gamma(x)$ is not empty, convex, and compact, and the graph $gph\Gamma := \{(x,v) : v \in \Gamma(x)\}$ is closed in R^{2n}. Let θ be a lower semicontinuous function and Θ be an open subset of R^n. Then*

(a)

$$\begin{aligned} &\min\{\langle v_1, \eta\rangle + v_2\rho : (v_1, v_2) \in \Gamma(z, r)\} \le 0 \quad \forall (z, r) \in epi\theta \cap \Theta, \\ &\qquad\qquad (\eta, \rho) \in N^p_{epi\theta}(z, r) \end{aligned} \tag{12}$$

if and only if

$$\min\{\langle v_1, \eta\rangle - v_2 : (v_1, v_2) \in \Gamma(z, \theta(z))\} \le 0 \quad \forall z \in \Theta, \eta \in \partial_p\theta(z). \tag{13}$$

(b) *In addition, assume that Γ is Lipschitz continuous. Then*

$$\begin{aligned} &\max\{\langle v_1, \eta\rangle + v_2\rho : (v_1, v_2) \in \Gamma(z, \theta(z))\} \le 0 \quad \forall (z, r) \in epi\theta \cap \Theta, \\ &\qquad\qquad (\eta, \rho) \in N^p_{epi\theta}(z, r) \end{aligned} \tag{14}$$

if and only if

$$\max\{\langle v_1, \eta\rangle - v_2 : (v_1, v_2) \in \Gamma(z, \theta(z)))\} \le 0 \quad \forall z \in \Theta, \eta \in \partial_p\theta(z). \tag{15}$$

Proof. Since an equivalent definition of the proximal subdifferential of ϕ at z is that

$$\eta \in \partial_p\theta(z) \text{ if and only if } (\eta, -1) \in N^p_{epi\theta}(z, \theta(z)),$$

(13) and (15) are (12) and (14) with $r = \theta(z)$ and $\rho = -1$, respectively. So it suffices to prove that (13) and (15) imply (12) and (14), respectively.

We first suppose that (13) holds. Let $(z,r) \in epi\theta \cap \Theta, (\eta,\rho) \in N^p_{epi\theta}(z,r)$. Then by the nature of epigraphs, we have $\rho \le 0$. Let us assume first that $\rho < 0$ from which it follows that $r = \theta(z)$. Since $N^p_{epi\theta}(z,\theta(z))$ is a cone, we have $(-\frac{\eta}{\rho}, -1) \in N^p_{epi\theta}(z,\theta(z))$ and consequently $-\frac{\eta}{\rho} \in \partial_p\theta(z)$. By (13), we have

$$\min\left\{-\left\langle v_1, \frac{\eta}{\rho}\right\rangle - v_2 : (v_1,v_2) \in \Gamma(z,\theta(z))\right\} \le 0.$$

Since $\rho < 0$, we have

$$\min\{\langle v_1,\eta\rangle + v_2\rho : (v_1,v_2) \in \Gamma(z,\theta(z))\} \le 0.$$

That is, (12) holds $\forall\ \rho < 0$.

We see that $(\eta,\rho) = 0$ trivially satisfies (12). Now suppose $\rho = 0$ and $\eta \ne 0$, from which it follows that $(\eta,0) \in N^p_{epi\theta}(z,\theta(z))$. By definition (cf. [16]), η is in the singular limiting subdifferential of θ at z. So there exists $\{z_i\}, \{\eta_i\}$, and $\{\rho_i\}$ so that $z_i \to z, \theta(z_i) \to \theta(z), \eta_i \to \eta, \rho_i < 0, \rho_i \uparrow 0$, and $-\frac{\eta_i}{\rho_i} \in \partial_p\theta(z_i)$. By (13), we have

$$\min\left\{-\left\langle v_1, \frac{\eta_i}{\rho_i}\right\rangle - v_2 : (v_1,v_2) \in \Gamma(z_i,\theta(z_i))\right\} \le 0.$$

So there exist $(v_1^i, v_2^i) \in \Gamma(z_i,\theta(z_i))$ such that

$$\langle v_1^i, \eta_i\rangle + v_2^i\rho_i \le 0.$$

Without loss of generality, assume that $v_1^i \to v_1, v_2^i \to v_2$. Then $(v_1,v_2) \in \Gamma(z,\theta(z))$ and

$$\langle v_1,\eta\rangle + v_2 \cdot 0 \le 0 \quad \forall (\eta,0) \in N_{epi\theta}(z,\theta(z)),$$

which is (12) when $\rho = 0$.

Now suppose (15) holds. Let $(\eta,\rho) \in N^p_{epi\theta}(z,r), (z,r) \in epi\theta \cap \Theta$. Then $\rho \le 0$. If $\rho < 0$, then the proof is similar to that in (a). If $\rho = 0$, then $r = \theta(z), (\eta,0) \in N^p_{epi\theta}(z,\theta(z))$. So there exist $\{z_i\}, \{\eta_i\}$, and $\{\rho_i\}$ such that $z_i \to z, \theta(z_i) \to \theta(z), \eta_i \to \eta, \rho_i < 0, \rho_i \uparrow 0$, and $-\frac{\eta_i}{\rho_i} \in \partial_p\theta(z_i)$. By (15), we have

$$\max\left\{-\left\langle v_1, \frac{\eta_i}{\rho_i}\right\rangle - v_2 : (v_1,v_2) \in \Gamma(z_i,\theta(z_i))\right\} \le 0.$$

That is,

$$\max\{\langle v_1,\eta_i\rangle + v_2\rho_i : (v_1,v_2) \in \Gamma(z_i,\theta(z_i))\} \le 0.$$

Since Γ is Lipschitz continuous, letting $(v_1,v_2) \in \Gamma(z_i,\theta(z_i))$, we have

$$\Gamma(z_i,\theta(z_i)) \subset \Gamma(z,\theta(z)) + K(\|z - z_i\|^2 + |\theta(z) - \theta(z_i)|^2)^{1/2}B.$$

Therefore there exists $(v_1^i, v_2^i) \in \Gamma(z,\theta(z))$ such that

$$(v_1,v_2) = (v_1^i, v_2^i) + K(\|x_1 - z_i\|^2 + |\theta(x_1) - \theta(z_i)|^2)^{1/2}e,$$

where $\|e\| \le 1$. Hence

$$\begin{aligned}\langle v_1,\eta_i\rangle + v_2\rho_i &= \langle v_1^i,\eta_i\rangle + v_2^i\rho_i + \langle\lambda_i e_1,\eta_i\rangle + \langle\lambda_i e_2,\rho_i\rangle\\ &\le \langle\lambda_i e_1,\eta_i\rangle + \langle\lambda_i e_2,\rho_i\rangle,\end{aligned}$$

where $\lambda_i = K(\|x_1 - z_i\|^2 + |\theta(x_1) - \theta(z_i)|^2)^{1/2} \to 0$ as $i \to \infty$. Taking limits, we have

$$\langle v_1, \eta \rangle + v_2 \cdot 0 \le 0 \quad \forall (\eta, 0) \in N_{epi\theta}(x_1, \theta(x_1)).$$

That is, (14) when $\rho = 0$. □

LEMMA 3.10 (see [24, Lemma 4.1]). *Let W be an extended-valued lower semicontinuous function. Then*

$$\partial_p(e^{-\lambda t}W(x)) = \{-\lambda e^{-\lambda t}W(x)\} \times \{e^{-\lambda t}\partial_p W(x)\}.$$

Proof of Proposition 3.7. By virtue of (a) in Lemmas 3.8 and 3.9, $(\{1\} \times F, X_W)$ is weakly invariant in $R \times G \times R$ if and only if

$$\min\{\xi_1 + \xi_2 \cdot g(x,u) + e^{-\lambda t}f(x,u) : u \in U\} \le 0 \quad \forall x \in G, \xi \in \partial_p(e^{-\lambda t}W(x)).$$

By Lemma 3.10, that is,

$$\min\{-\lambda W(x) + \eta \cdot g(x,u) + f(x,u) : u \in U\} \le 0 \quad \forall x \in G, \eta \in \partial_p W(x).$$

Hence

$$\lambda W(x) + H(x, -\eta) \ge 0 \quad \forall x \in G, \eta \in \partial_p W(x).$$

By virtue of (b) in Lemmas 3.8 and 3.9, $(\{1\} \times \widetilde{F}, X_{-W})$ is strongly invariant in $R \times G \times R$ if and only if

$$\max\{\xi_1 + \xi_2 \cdot g(x,u) - e^{-\lambda t}f(x,u) : u \in U\} \le 0 \quad \forall x \in G, \xi \in \partial_p(-e^{-\lambda t}W(x)).$$

By Lemma 3.10, that is,

$$\max\{\lambda W(x) - \eta \cdot g(x,u) - f(x,u) : u \in U\} \le 0 \quad \forall x \in G, \eta \in \partial_p(-W(x)).$$

Hence

$$\lambda W(x) + H(x, -\eta) \le 0 \quad \forall x \in G, \eta \in \partial^p W(x).$$

By virtue of (b) in Lemmas 3.8 and 3.9, $(\{-1\} \times \{-F\}, X_W)$ is strongly invariant in $R \times G \times R$ if and only if

$$\max\{-\xi_1 - \xi_2 \cdot g(x,u) - e^{-\lambda t}f(x,u) : u \in U\} \le 0 \quad \forall x \in G, \xi \in \partial_p(e^{-\lambda t}W(x)).$$

By Lemma 3.10, that is,

$$\max\{\lambda W(x) - \eta \cdot g(x,u) - f(x,u) : u \in U\} \le 0 \quad \forall x \in R^d, \eta \in \partial_p W(x).$$

Hence

$$\lambda W(x) + H(x, -\eta) \le 0 \quad \forall x \in G, \eta \in \partial_p W(x). \qquad \square$$

We now derive from Propositions 3.6 and 3.7 the equivalence between the optimality principles and the H–J equations.

PROPOSITION 3.11 (equivalence of optimality principles and the H–J equations). *Let G be an open subset of R^d.*

(a) *Let $W : G \to R$ be a lower semicontinuous function. Then it satisfies the superoptimality principle in G if and only if it is a supersolution of the H–J equation involving the proximal subdifferentials in G; i.e.,*

$$\lambda W(x) + H(x, -\partial_p W(x)) \geq 0 \quad \forall x \in G.$$

(b) *Let $W : G \to R$ be an upper semicontinuous function. Then it satisfies the suboptimality principle in G if and only if it is a subsolution of the H–J equation involving the proximal superdifferentials in G; i.e.,*

$$\lambda W(x) + H(x, -\partial^p W(x)) \leq 0 \quad \forall x \in G.$$

(b′) *Let $W : G \to R$ be a lower semicontinuous function. Then it satisfies the suboptimality principle in G if and only if it is a subsolution of the H–J equation involving the proximal subdifferentials in G; i.e.,*

$$\lambda W(x) + H(x, -\partial_p W(x)) \leq 0 \quad \forall x \in G.$$

4. Proof of main results.

Proof of Theorem 2.2. By Proposition 3.3, $V_*(x)$ satisfies the superoptimality principle in E. So by (a) of Proposition 3.11, it is a supersolution of the H–J equation involving the proximal subdifferentials in E.

We now prove that V_* satisfies the boundary condition

$$\max\{V_*(x) - h(x), \lambda V_*(x) + H(x, -\partial_p V_*(x))\} \geq 0 \quad \forall x \in \partial E. \tag{16}$$

If $V_*(x) - h(x) \geq 0 \ \forall x \in \partial E$, then the boundary condition (16) holds. Otherwise suppose that there exists $x \in \partial E$ such that $V_*(x) < h(x)$.

Let $x_n \to x, V(x_n) \to V_*(x)$. We may assume without loss of generality that $x_n \in E \ \forall n$. Indeed, if there exists a subsequence $\{x_p\}$ of $\{x_n\}$ such that $x_p \in \partial E \ \forall p$, then, by definition of the value function on the boundary of E and the lower semicontinuity of the exit cost h, we have

$$V_*(x) = \lim_{n\to\infty} V(x_n) = \lim_{p\to\infty} V(x_p) = \lim_{p\to\infty} h(x_p) \geq h(x),$$

which contradicts the assumption that $V_*(x) < h(x)$.

Now by the Bellman optimality principle, there exists a control $u_n \in \mathcal{U}$, $t_n^* := t^*[x_n, u_n] > 0$ such that

$$V(x_n) \geq \int_0^r e^{-\lambda s} f(y[x_n, u_n](s), u_n(s))ds + e^{-\lambda \tau_n} V(y[x_n, u_n](r)) \ \ \forall 0 \leq r \leq t_n^*.$$

Now let $\bar{r} = \liminf t_n^*$. We must have $\bar{r} > 0$, since otherwise we can find a subsequence of $\{t_n^*\}$ such that $t_n^* \to 0$ so that

$$\begin{aligned} V_*(x) &= \lim_{n\to\infty} V(x_n) \\ &\geq \liminf_{n\to\infty} \int_0^{t_n^*} e^{-\lambda s} f(y[x_n, u_n](s), u_n(s))ds + \liminf_{n\to\infty} e^{-\lambda t_n^*} h(y[x_n, u_n](t_n^*)) \\ &\geq h(x) \quad \text{since } h \text{ is lower semicontinuous,} \end{aligned}$$

which is a contradiction. Now by the compactness of relaxed controls on $[0, \bar{r}]$, there exists $u = \lim_{n\to\infty} u_n$ such that

$$V_*(x) \geq \int_0^r e^{-\lambda s} f(y[x, u](s), u(s))ds + e^{-\lambda r} V_*(y[x, u](r)) \ \ \forall r \in (0, \bar{r}].$$

Let $\xi \in \partial_p V_*(x)$. Then there exist $\sigma > 0, \delta > 0$ such that

$$V_*(x') - V_*(x) + \sigma\|x' - x\|^2 \geq \langle \xi, x' - x\rangle \quad \forall x' \in x + \delta B.$$

Let $x' = y[x,u](r)$ where $r \in [0, \bar{r}]$ is fixed. Then

$$\begin{aligned}\langle \xi, y[x,u](r) - x\rangle &\leq \sigma\|y[x,u](r) - x\|^2 + V_*(y[x,u](r)) - V_*(x)\\ &\leq \sigma\|y[x,u](r) - x\|^2 - e^{\lambda r}\int_0^r e^{-\lambda s} f(y[x,u](s), u(s))ds\\ &\quad + e^{\lambda r}V_*(x) - V_*(x).\end{aligned}$$

Since $y[x,u](r) - x = \int_0^r g(y[x,u](s), u(s))ds$, one has

$$\begin{aligned}&\int_0^r [\langle -\xi, g(y[x,u](s), u(s))\rangle - f(y[x,u](s), u(s))]ds\\ &+ \int_0^r [(1 - e^{\lambda(r-s)})f(y[x,u](s), u(s))]ds + (e^{\lambda r} - 1)V_*(x) \geq -\sigma\|y[x,u](r) - x\|^2.\end{aligned}$$

By virtue of the boundedness of g and the Lipschitz continuity of g, f uniformly in $u \in U$, one has

$$\begin{aligned}&\|y[x,u](r) - x\| \leq M_g r\\ &(\|\xi\|L_g + L_f)M_g s + \langle \xi, g(x, u(s))\rangle - f(x, u(s))\\ &\quad \geq \langle \xi, g(y[x,u](s), u(s))\rangle - f(y[x,u](s), u(s)),\end{aligned}$$

where M_g, L_g, L_f denote the bound of g and the Lipschitz constants of g, f, respectively. Therefore, one has

$$\begin{aligned}&\int_0^r [\langle -\xi, g(x, u(s))\rangle - f(x, u(s))]ds + (e^{\lambda r} - 1)V_*(x)\\ &\geq o(r) - \int_0^r [(1 - e^{\lambda(r-s)})f(y[x,u](s), u(s))]ds\\ &\geq o(r) - \int_0^r (1 - e^{\lambda(r-s)})M_f ds,\end{aligned}$$

where $o(r)$ indicates a function $g(r)$ such that $\lim_{t\to 0^+} |g(r)|/r = 0$ and M_f is the bound of f. Since the term in the square bracket in the first integral is bounded from above by

$$H(x, -\xi) = \max\{\langle -\xi, g(x,u)\rangle - f(x,u) : u \in U\},$$

(17) implies that

$$H(x, -\xi)r + (e^{\lambda r} - 1)V_*(x) \geq o(r) - \int_0^r (1 - e^{\lambda(r-s)})M_f ds.$$

Dividing the above inequality by r and letting $r \to 0$, we have

$$\lambda V_*(x) + H(x, -\partial_p V_*(x)) \geq 0.$$

Similarly by Proposition 3.3, $V^*(x)$ satisfies the suboptimality principle. So by Proposition 3.6, $(\{-1\} \times \{-F\}, X_{V^*})$ is strongly invariant in $R \times E \times R$. Hence, by

Proposition 3.7, V^* is a proximal subsolution of the H–J equation. The boundary condition can be proved similarly. □

Proof of Theorem 2.3. By Proposition 3.2, the value function $V(x)$ satisfies the suboptimality principle in E. Since $V(x)$ is upper semicontinuous, by (b) of Proposition 3.11, it is a subsolution of the H–J equation involving the proximal superdifferentials, i.e.,

$$\lambda V(x) + H(x, -\partial^p V(x)) \leq 0 \quad \forall x \in E.$$

Conversely, let $W(x)$ be an upper semicontinuous function such that

$$\begin{aligned} \lambda W(x) + H(x, -\partial^p W(x)) &\leq 0 \quad \forall x \in E \\ W(x) &\leq h(x) \quad \forall x \in \partial E. \end{aligned}$$

Then by (b) of Proposition 3.11, W satisfies the suboptimality principle in E. By (b) of Proposition 3.4, $W(x) \leq V(x)\ \forall x \in \bar{E}$. □

Proof of Theorem 2.4. By Proposition 3.2, the value function V satisfies both the superoptimality principle in E and the suboptimality principle in E. Since the value function is assumed to be lower semicontinuous, by the equivalence of the optimality principles and the H–J equations ((a) and (b′) of Proposition 3.11), the value function is both a supersolution and subsolution (hence a solution) of the H–J equation involving the proximal subdifferentials. Now if $W(x)$ is a lower semicontinuous solution of the H–J equation involving the proximal subdifferentials in E with the natural boundary condition $W(x) = h(x)\ \forall x \in \partial E$, then by (a) of Proposition 3.4, $W(x) \geq V(x)\ \forall x \in E$. □

Proof of Theorem 2.7. By Proposition 3.2, the value function V satisfies both the superoptimality principle in E and the suboptimality principle in E. Observing that $V(x) = h(x)\ \forall x \in E^c$ we have by assumption (H3) that the value function also satisfies the suboptimality principle in O which contains $\bar{E}$. Since by Proposition 2.5 the value function is lower semicontinuous, by (a) and (b′) of Proposition 3.11,

$$\lambda V(x) + H(x, -\partial_p V(x)) \geq 0 \quad \forall x \in E, \tag{17}$$

$$\lambda V(x) + H(x, -\partial_p V(x)) \leq 0 \quad \forall x \in O. \tag{18}$$

Now suppose W is a lower semicontinuous function that satisfies (17), (18), and the natural boundary condition $W(x) = h(x)\ \forall x \in O \backslash E$. Then by Proposition 3.11, W satisfies both the superoptimality principle in E and the suboptimality principle in O. Hence by (a) and (b′) of Proposition 3.4, $W(x) = V(x)\ \forall x \in \bar{E}$. □

REFERENCES

[1] M. Bardi and I. Capuzzo-Dolcetta, *Optimal Control and Viscosity Solutions of Hamilton-Jacobi-Bellman Equations*, Birkhäuser, Boston, 1997.

[2] M. Bardi and P. Soravia, *Hamilton-Jacobi equations with singular boundary conditions on a free boundary and applications to differential games*, Trans. Amer. Math. Soc., 325 (1991), pp. 205–229.

[3] E.N. Barron and R. Jensen, *Semicontinuous viscosity solutions for Hamilton-Jacobi equations with convex Hamiltonians*, Comm. Partial Differential Equations, 15 (1990), pp. 1713–1742.

[4] G. Barles and B. Perthame, *Discontinuous solutions of deterministic optimal stopping time problems*, Math. Model. Numer. Anal., 21 (1987), pp. 557–579.

[5] G. Barles and B. Perthame, *Exit time problems in optimal control and vanishing viscosity method*, SIAM J. Control Optim., 26 (1988), pp. 1133–1148.

[6] A.-P. Blanc, *Deterministic exit time control problems with discontinuous exit costs*, SIAM J. Control Optim., 35 (1997), pp. 399–434.
[7] F.H. Clarke, *Methods of Dynamic and Nonsmooth Optimization*, CBMS-NSF Regional Conf. Ser. in Appl. Math. 57, SIAM, Philadelphia, PA, 1989.
[8] F.H. Clarke, Yu.S. Ledyaev, R.J. Stern, and P.R. Wolenski, *Qualitative properties of trajectories of control system: A survey*, J. Dynam. Control Systems, 1 (1995), pp. 1–48.
[9] M.G. Crandall and P.-L. Lions, *Viscosity solutions of Hamilton-Jacobi equations*, Trans. Amer. Math. Soc., 277 (1983), pp. 1–42.
[10] M.A.H. Dempster and J.J. Ye, *Generalized Bellman-Hamilton-Jacobi optimality conditions for a control problem with a boundary condition*, Appl. Math. Optim., 33 (1996), pp. 211–225.
[11] H. Frankowska, *Lower semicontinuous solutions of Hamilton–Jacobi–Bellman equations*, SIAM J. Control Optim., 31 (1993), pp. 257–272.
[12] R. Gonzalez, *Sur la resolution de l'equation de Hamilton-Jacobi du contrôle deterministe*, Thèse de 3^o Cycle, Univ. Paris, 1980.
[13] R. Gonzalez and E. Rofman, *An algorithm to obtain the maximum solutions of the Hamilton-Jacobi equation*, in Optimization Techniques, J. Stoer, ed., Springer, Berlin, 1978, pp. 109–116.
[14] H. Ishii, *A boundary value problem of the Dirichlet type for Hamilton-Jacobi equations*, Ann. Sc. Norm. Sup. Pisa Cl. Sci., 16 (1989), pp. 105–135.
[15] H. Ishii, *Perron's method for Hamilton-Jacobi equations*, Duke Math. J., 55 (1987), pp. 369–384.
[16] P.D. Loewen, *Optimal Control via Nonsmooth Analysis*, CRM Proc. Lecture Notes 2, AMS, Providence, RI, 1993.
[17] P. Soravia, *Discontinuous viscosity solutions to Dirichlet problems for the Hamilton-Jacobi equations with convex Hamiltonians*, Comm. Partial Differential Equations, 18 (1993), pp. 1493–1514.
[18] P. Soravia, *Pursuit–evasion problems and viscosity solutions of Isaacs equations*, SIAM J. Control Optim., 31 (1993), pp. 604–623.
[19] A.I. Subbotin, *Generalization of the fundamental equation of the theory of differential games*, Dokl. Akad. Nauk SSSR, 254 (1980), pp. 293–297.
[20] A.I. Subbotin, *Generalized Solutions of First-Order PDEs, the Dynamical Optimization Perspective*, Birkhaüser, Boston, 1995.
[21] J. Warga, *Optimal Control of Differential and Functional Equations*, Academic Press, New York, 1972.
[22] P.R. Wolenski and Y. Zhuang, *Proximal analysis and the minimal time function*, SIAM J. Control Optim., 36 (1998), pp. 1048–1072.
[23] J.J. Ye, *Optimal Control of Piecewise Deterministic Markov Processes*, Ph.D. thesis, Dalhousie University, Halifax, NS, Canada, 1990.
[24] J.J. Ye and Q.J. Zhu, *Hamilton-Jacobi theory for a generalized optimal stopping time problem*, Nonlinear Anal., 34 (1998), pp. 1029–1053.

SIAM J. CONTROL OPTIM.
Vol. 38, No. 4, pp. 1086–1101

LONG TIME BEHAVIOR FOR SOME DYNAMICAL NOISE FREE NONLINEAR FILTERING PROBLEMS*

FRÉDÉRIC CÉROU†

Abstract. A possible approach to studying the accuracy of optimal nonlinear filtering is to look at the conditional law on an arbitrary neighborhood of the current (unknown) state as the time goes to infinity. This paper is devoted to a restricted class of systems with a deterministic state equation, and under additional assumptions, results concerning the concentration of the conditional law are shown, as is how estimating the current state could be different from estimating the initial condition. The assumptions on the system could seem quite restrictive; however, they concern only the "observability" of the system, without any reference to ergodicity as in previous works.

Key words. nonlinear filtering, long time behavior, measure concentration

AMS subject classifications. 93E11, 60G35

PII. S0363012995290124

1. Introduction.

The optimal filtering of nonlinear systems, in particular in the framework of stochastic differential equations of Itô type, is now a well-studied problem: the conditional law of the current state, given the past observations, is the solution of the Kushner–Stratonovitch equation. For numerical purposes, the Zakai equation, which gives an unnormalized conditional law, is also used. See, for instance, [14] for a review on nonlinear filtering. Nevertheless, a crucial question is still open concerning the "accuracy" of the filtering: that is, as time increases and as we obtain more and more observations, what can we say about the conditional law of the current state? What are the conditions to be imposed on the system in order to control in some sense the filtering error?

Some results have been proved in [8] and [9] by Kunita when the state process is ergodic. Let us also mention that the results presented in [8] have been extended to the noncompact case in [7] by Ji. If we look at this problem as an asymptotic stability problem of the Kushner–Stratonovitch equation (which can be considered as a stochastic dynamical system on the infinite dimensional space of probability measures on the state space), it becomes clear that it is closely related to the sensitivity of the filter to the prior distribution. That is, given an erroneous initial distribution, does the filter give results close to the ones obtained with the correct initialization, as the time goes to infinity? Using again ergodicity arguments, Ocone and Pardoux [13] showed some results concerning this topic. Note that this question about stability was already addressed in the linear case by Bucy and Joseph [2].

Nevertheless, it seems natural that without any ergodicity, only some good fitting between the dynamics and the observations should be sufficient (related to some "observability" notion) to control the filtering error, say, to make it asymptotically bounded for large time. This has been an open problem since the early developments of the theory and may be considered as both challenging for the scientist and crucial for practical purposes.

The aim of the present paper is to deal with such problems, in the restricted framework of noise free dynamics, that is, with identically zero diffusion coefficient

*Received by the editors August 9, 1995; accepted for publication (in revised form) July 1, 1999; published electronically April 4, 2000.

http://www.siam.org/journals/sicon/38-4/29012.html

†IRISA/INRIA, Campus de Beaulieu, 35042 Rennes Cedex, France (Frederic.Cerou@irisa.fr).

on the state equation. This is a much simpler but nontrivial model, used with success in some applications (such as passive target tracking), and for which well-adapted numerical algorithms have been developed (see [1] and [3] for details). In fact we can adopt a direct approach here: as we can write an explicit formula for the conditional unnormalized density, under some assumptions ensuring a uniform good "observability," we can directly estimate the long time behavior of the conditional law. More precisely we will show that for any ball of fixed radius centered at the current state (and also moving with it), the integral of the conditional density on the ball converges in probability to 1 as time goes to infinity. This criterion comes from practice—say tracking problems—when the output consists of confidence regions, and it will ensure that these regions will be included at large times in arbitrarily small neighborhoods of the current state. We will investigate the general case without special behavior of the dynamics, for which we need good observations on the whole space, and the linear Gaussian case, for which we can give the behavior of the conditional covariance matrix. Even though the assumptions imposed on the coefficients still seem too restrictive for practical use, the main purpose of this work is to illustrate that the ergodicity conditions are not needed to study the accuracy of the filter at large times.

Before studying the filtering problem we present a particular approach to an estimation problem, keeping in mind that the method developed will be useful in what follows.

2. Estimation. We consider the following estimation problem:

$$dY(t) = S\,(t, X_0)\,dt + dB_t,$$

where S is a measurable function from $\mathbb{R} \times \mathbb{R}^d$ into $\mathbb{R}^m$, $X_0 \in \mathbb{R}^d$ is a random vector, and B denotes standard m-dimensional Brownian motion. We denote by P_0 the law of X_0. For all $x \in \mathbb{R}^d$ we take the notation

$$g(t, x) = S(t, x) - S(t, X_0).$$

We assume that X_0 and B are independent. This section is devoted to the study of the asymptotic behavior of the conditional law of X_0, given the observations $\{Y_s,\, s \le t\}$, as $t \to +\infty$. This is a well-studied problem (see [6], for instance) but the standard approach is difficult to use for related filtering problems that we will study in the next sections, so we present here some methods that will be useful later. Assume that the following hypotheses are fulfilled, ensuring the existence of a conditional density with respect to P_0 given by a Bayes formula. Let $\{\mathcal{F}_t^Y,\, t \ge 0\}$ be the filtration of the process Y, and $\forall\ t \ge 0$

$$\forall x \in \mathbb{R}^d,\ \int_0^t \|S(s, x)\|\,ds < \infty, \tag{1}$$

$$P\left(\int_0^t \|S(s, X_0)\|^2\,ds < \infty\right) = 1, \tag{2}$$

and

$$P\left(\int_0^t \left\|E\left[S(s, X_0) \,|\, \mathcal{F}_s^Y\right]\right\|^2 ds < \infty\right) = 1. \tag{3}$$

Then by [12, Theorem 7.23, p. 289], the Bayes formula gives for all bounded continuous ψ

$$E\left[\psi(X_0)\,|\,\mathcal{F}_t^Y\right] = G\int_{\mathbb{R}^d}\psi(x)\,\exp\left[-\frac{1}{2}\int_0^t S^2(s,x)\,ds + \int_0^t S(s,x)\,dY_s\right]\,P_0(dx),$$

where G is a $\mathcal{F}_t^Y$ measurable normalizing factor. After dividing G and multiplying the integral term by $\exp\left[-\frac{1}{2}\int_0^t S^2(s,X_0)\,ds - \int_0^t S(s,X_0)\,.\,dB_s\right]$, we get the following form for the unnormalized a posteriori law ν_t^0:

$$\nu_t^0(A) = \int_A f_t(x)\,P_0(dx) = \int_A \exp\left[\int_0^t g(s,x).dB_s - \frac{1}{2}\int_0^t g^2(s,x)ds\right]P_0(dx) \tag{4}$$

for all Borel sets A. Let us denote by μ_t^0 the normalized conditional law. In the whole section we suppose that the following hypothesis is fulfilled.

(H1). There are two functions v and V from $\mathbb{R}_+$ into $\mathbb{R}_+$ and a positive constant C such that $1 \leq \frac{V(t)}{v(t)} \leq C$ and

$$\forall (x_1,x_2)\in\mathbb{R}^d\times\mathbb{R}^d,\quad v(t)\|x_1-x_2\|^2 \leq \int_0^t \|S(s,x_1)-S(s,x_2)\|^2 ds \leq V(t)\|x_1-x_2\|^2.$$

Notice that it is reasonable to assume that $\int_0^t \|S(s,x)\|^2\,ds < \infty$ for some x and that then assumption (H1) is much stronger than (1)–(3).

Main result. We have the following proposition concerning the concentration of the conditional law.

PROPOSITION 2.1. *Let us denote*

$$\mathcal{B}_a^0 = \{x\in\mathbb{R}^d,\ \|x-X_0\|\leq a\}.$$

Assume that (H1) *is fulfilled, and* $\lim_{t\to+\infty} V(t) = +\infty$. *Then* $\forall\ a>0$

$$\mu_t^0\left(\mathcal{B}_a^0\right) \xrightarrow{P} 1 \quad as \quad t\to+\infty.$$

Moreover, we have the following estimate concerning the rate of convergence:

$$\forall\ 0<\beta<\frac{a^2}{4C},\quad \exp(\beta V(t))\left(1-\mu_t^0\left(\mathcal{B}_a^0\right)\right)\xrightarrow{P} 0 \quad as \quad t\to+\infty.$$

To prove this proposition, we will need some auxiliary results to control the stochastic integral which appears in the likelihood function f_t.

PROPOSITION 2.2. *Assume that the upper estimate in* (H1) *is valid. Then there exists* $A_1>0$ *such that* $\forall\ a>0$,

$$E\left[\sup_{\|x-X_0\|\leq a}\left|\int_0^t g(s,x).dB_s\right|\right] \leq A_1 a\sqrt{V(t)}.$$

Proof. We note

$$\mathcal{B}_a^0 = \left\{x\in\mathbb{R}^d/\|x-X_0\|\leq a\right\}.$$

Since X_0 and B are independent, we can assume that $P = P_{X_0}\otimes P_B$ on some probability space $\Omega_{X_0}\times\Omega_B$. We denote by E_{X_0} (resp., E_B) the expectation over Ω_{X_0} (resp.,

Ω_B). Given X_0, the integral in the supremum is a Gaussian process, so [11, Theorem 11.17, p. 321] gives the estimate

$$E\left[\sup_{\|x-X_0\|\le a}\left|\int_0^t g(s,x).dB_s\right|\right] = 2\,E_{X_0}E_B\left[\sup_{\|x-X_0\|\le a}\int_0^t g(s,x).dB_s\right]$$
$$\le 48E_{X_0}\left[\int_0^\infty \left(\log N\left(\mathcal{B}_a^0, d_{U_t};\zeta\right)\right)^{\frac{1}{2}} d\zeta\right],$$

where d_{U_t} the pseudometric given by

$$d_{U_t}(x_1,x_2) = \|U_t(x_1,x_2)\|_{L^2(\Omega)}$$
$$= \left[\int_0^t |S(s,x_1)-S(s,x_2)|^2 ds\right]^{\frac{1}{2}}$$

with

$$U_t(x_1,x_2) = \int_0^t \left(S(s,x_1)-S(s,x_2)\right).dB_s$$

and $N\left(\mathcal{B}_a^0, d_{U_t};\zeta\right)$ is the smallest number of balls of radius ζ, for the pseudometric d_{U_t}, needed to cover $\mathcal{B}_a^0$. The hypothesis above gives

$$d_{U_t}(x_1,x_2) \le (V(t))^{\frac{1}{2}}\|x_1-x_2\| \le K(V(t))^{\frac{1}{2}}\|x_1-x_2\|_\infty,$$

where K comes from the equivalence of the norms in $\mathbb{R}^d$. So we get

$$N\left(\mathcal{B}_a^0, d_{U_t};\zeta\right) \le N\left(\mathcal{B}_a^0, K(V(t))^{\frac{1}{2}}\|.\|_\infty;\zeta\right)$$
$$= N\left(\mathcal{B}_a^0, \|.\|_\infty; \frac{\zeta}{K(V(t))^{\frac{1}{2}}}\right)$$
$$\le \left(\frac{2aK(V(t))^{\frac{1}{2}}}{\zeta}+1\right)^d.$$

From this follows

$$\int_0^\infty \left(\log N\left(\mathcal{B}_a^0, d_{U_t};\zeta\right)\right)^{\frac{1}{2}} d\zeta$$
$$\le \int_0^\infty \left(\log N\left(\mathcal{B}_a^0, \|.\|_\infty; \frac{\zeta}{K(V(t))^{\frac{1}{2}}}\right)\right)^{\frac{1}{2}} d\zeta$$
$$\le \int_0^{2aK(V(t))^{\frac{1}{2}}} d^{\frac{1}{2}}\left(\log\left(\frac{2aK(V(t))^{\frac{1}{2}}}{\zeta}+1\right)\right)^{\frac{1}{2}} d\zeta$$
$$= d^{\frac{1}{2}}2aK(V(t))^{\frac{1}{2}}\int_0^1 \left(\log\left(1+\frac{1}{u}\right)\right)^{\frac{1}{2}} du$$

with the change of variables

$$u = \frac{\zeta}{2aK(V(t))^{\frac{1}{2}}}.$$

As $0 \leq \left(\log\left(1+\frac{1}{u}\right)\right)^{\frac{1}{2}} \leq \frac{1}{\sqrt{u}}$, the last integral is convergent, and the proposition is proved. □

PROPOSITION 2.3. *Assume that the upper estimate in* (H1) *is valid. Then there exists* $A_2 > 0$ *such that* $\forall\ a > 0$,

$$E\left[\sup_{\|x-X_0\|\geq a}\left|\frac{\int_0^t g(s,x).dB_s}{\|x-X_0\|^2}\right|\right] \leq A_2 a^{-1}\sqrt{V(t)}.$$

Proof. Let n_0 be the greatest relative integer such that $2^{n_0} \leq a$. The triangular inequality gives

$$\begin{aligned}
E\left[\sup_{\|x-X_0\|\geq a}\frac{\left|\int_0^t g(s,x).dB_s\right|}{\|x-X_0\|^2}\right] &\leq \sum_{n\geq n_0}\frac{1}{2^{2n}}E\left[\sup_{2^n\leq\|x-X_0\|\leq 2^{n+1}}\left|\int_0^t g(s,x).dB_s\right|\right] \\
&\leq \sum_{n\geq n_0}\frac{1}{2^{2n}}E\left[\sup_{\|x-X_0\|\leq 2^{n+1}}\left|\int_0^t g(s,x).dB_s\right|\right] \\
&\leq A_1(V(t))^{\frac{1}{2}}\sum_{n\geq n_0}\frac{2^{n+1}}{2^{2n}} \quad \text{(Proposition 2.2)} \\
&\leq A_1(V(t))^{\frac{1}{2}}2^{-n_0}\sum_{p\geq 0}\frac{2^{p+1}}{2^{2p}} \\
&\leq A_1(V(t))^{\frac{1}{2}}\frac{2}{a}\sum_{p\geq 0}\frac{2^{p+1}}{2^{2p}}.
\end{aligned}$$

Because $2^{n_0} \leq a \leq 2^{n_0+1}$, the convergence of the series concludes the proof. □

Proof of Proposition 2.1. First note that it is sufficient to show that $\forall a > 0$ and $0 < \beta < \frac{a^2}{4C}$,

$$\exp(\beta V(t))\frac{\int_{\mathcal{B}_a^{0,c}} f_t(x)P_0(dx)}{\int_{\mathcal{B}_a^0} f_t(x)P_0(dx)} \xrightarrow{P} 0 \quad \text{as} \quad t \to +\infty. \tag{5}$$

We note that it is equivalent to show that convergence (5) holds for every increasing sequence of positive times t_k with $\lim_{k\to+\infty} t_k = +\infty$. We will use the characterization of the convergence in probability in terms of almost sure convergence. That is, from any subsequence t_ℓ we have to extract a subsubsequence t_n on which (5) holds almost surely (a.s.). From Propositions 2.2 and 2.3, as $\lim_{t\to+\infty} V(t) = +\infty$, we obviously have

$$\frac{1}{V(t)}\sup_{\|x-X_0\|\leq a}\left|\int_0^t g(s,x).dB_s\right| \xrightarrow{P} 0 \quad \text{as} \quad t \to +\infty$$

and

$$\frac{1}{V(t)} \sup_{\|x-X_0\|\geq a} \left| \frac{\int_0^t g(s,x).dB_s}{\|x-X_0\|^2} \right| \xrightarrow{P} 0 \quad \text{as} \quad t \to +\infty.$$

Thus, from any t_ℓ we can extract an increasing sequence of positive times t_n, with $\lim_{n\to+\infty} t_n = +\infty$, such that the two previous convergences hold a.s. Then we need to show only that on this sequence, convergence (5) holds a.s. as well. Using (H1), we get

$$\begin{aligned}
&\int_{\mathcal{B}_a^{0,c}} f_{t_n}(x) P_0(dx) \\
&\leq \int_{\mathcal{B}_a^{0,c}} \exp\left[-\frac{1}{2} v(t_n) \|x-X_0\|^2 + \int_0^{t_n} g(s,x).dB_s \right] P_0(dx) \\
&\leq \int_{\mathcal{B}_a^{0,c}} \exp\left[-V(t_n)\|x-X_0\|^2 \left(\frac{1}{2C} - \frac{1}{V(t_n)} \sup_{x\in\mathcal{B}_a^{0,c}} \left| \frac{\int_0^{t_n} g(s,x).dB_s}{\|x-X_0\|^2} \right| \right) \right] P_0(dx) \\
&\leq \int_{\mathcal{B}_a^{0,c}} \exp\left[-\frac{V(t_n)}{4C} \|x-X_0\|^2 \right] P_0(dx) \\
&\leq \int_{\mathcal{B}_a^{0,c}} \exp\left[-\frac{V(t_n)}{4C} a^2 \right] P_0(dx)
\end{aligned}$$

for n large enough. Thus we obtain

$$\int_{\mathcal{B}_a^{0,c}} f_{t_n}(x) P_0(dx) \leq \exp\left[-\frac{a^2 V(t_n)}{4C} \right].$$

On the other hand, $\forall\ 0 < a_0 \leq a$,

$$\begin{aligned}
\int_{\mathcal{B}_a^0} f_{t_n}(x) P_0(dx) &\geq \int_{\mathcal{B}_{a_0}^0} \exp\left[-\frac{1}{2} V(t_n) \|x-X_0\|^2 + \int_0^{t_n} g(s,x).dB_s \right] P_0(dx) \\
&\geq \int_{\mathcal{B}_{a_0}^0} \exp\left[-V(t_n) \left(\frac{a_0^2}{2} + \frac{1}{V(t_n)} \sup_{x\in\mathcal{B}_a^0} \left| \int_0^{t_n} g(s,x).dB_s \right| \right) \right] P_0(dx) \\
&\geq \exp\left[-a_0^2 V(t_n) \right] P_0(\mathcal{B}_{a_0}^0)
\end{aligned}$$

for n large enough. Note that X_0 is a.s. in the support of P_0, thus a.s. we have $P_0(\mathcal{B}_{a_0}^0) > 0$. Then we get

$$\exp\left(\beta V(t_n)\right) \frac{\int_{\mathcal{B}_a^{0,c}} f_{t_n}(x) P_0(dx)}{\int_{\mathcal{B}_a^0} f_{t_n}(x) P_0(dx)} \leq \frac{1}{P_0(\mathcal{B}_{a_0}^0)} \exp\left[V(t_n) \left(\beta + a_0^2 - \frac{a^2}{4C} \right) \right],$$

which goes to 0 provided we took $a_0 < \sqrt{\frac{a^2}{4C} - \beta}$. □

3. Dynamical noise free filtering: general case. We consider now a slightly different situation where we are given a dynamical noise free filtering problem:

$$\left\{ \begin{array}{rcl} dX_t &=& b(X_t)dt, \\ dY_t &=& h(X_t)dt + dB_t, \end{array} \right. \tag{6}$$

where X_t takes values in $\mathbb{R}^d$ and Y_t in $\mathbb{R}^m$, X_0 is a random vector with the given law P_{X_0}, and B_t is an m-dimensional Wiener process independent of X_0. From now on, we will assume that P_0 has a density with respect to Lebesgue measure denoted by p_0. In the following, $\Phi_t(x)$ will denote the (deterministic) flow of the state equation; it is the state reached by the system at time t, starting from x at $t = 0$. We will assume that it is well defined for every time t and every initial condition x_0. If we want to estimate the initial state, we have a particular case of the previous problem with

$$S(t, .) = h(\Phi_t(.)).$$

Thus, in order to estimate X_0, we can apply the previous section and study the estimation problem. But here the term "filtering" means that we are interested in the current state X_t. So the object to be estimated is also moving, and we will say that the filtering algorithm gives good results if the conditional law of X_t concentrates on arbitrary balls centered at X_t (and also moving with it). The purpose of this section is to find sufficient conditions imposed on the system (6) to have such a behavior of the conditional law. So, to transpose the previous results on the current state X_t, we will assume in addition that the flow verifies the next hypothesis.

(H2). $\|\Phi_t(x_1) - \Phi_t(x_2)\| \leq \alpha(t)\,\|x_1 - x_2\|\ \forall\ x_1,\ x_2$ in $\mathbb{R}^d$ and $\forall\ t > 0$ and for some monotonous function α from $\mathbb{R}_+$ into $\mathbb{R}_+$.

REMARK 3.1. *Note that the monotonous character of the function α is not restricting as we can always transform it in a nondecreasing function by taking* $\sup_{0\leq s\leq t} \alpha(s)$.

We are interested in the concentration of the a posteriori law μ_t of X_t on sets of the form

$$\mathcal{B}_a^t = \left\{x \in \mathbb{R}^d / \|x - X_t\| \leq a\right\}.$$

Let $\{\mathcal{F}_t^Y,\ t \geq 0\}$ be the filtration of the process Y. For all test functions ψ we have

$$E\left[\psi(X_t) \,|\, \mathcal{F}_t^Y\right] = E\left[\psi \circ \Phi_t(X_0) \,|\, \mathcal{F}_t^Y\right]$$

$$= \int_{\mathbb{R}^d} \psi \circ \Phi_t(\xi)\, \mu_t^0(d\xi)\,.$$

Thus

$$\mu_t\left(\mathcal{B}_a^t\right) = \int_{\mathbb{R}^d} \mathbf{1}_{\mathcal{B}_a^t}\left(\Phi_t^{-1}(\xi)\right)\, \mu_t^0(d\xi)$$

$$= \mu_t^0\left(\Phi_t^{-1}\left(\mathcal{B}_a^t\right)\right)\,,$$

which we rewrite as

$$\mu_t\left(\mathcal{B}_a^t\right) = \frac{\int_{\Phi_t^{-1}(\mathcal{B}_a^t)} f_t(x)\, p_0(x)\, dx}{\int_{\mathbb{R}^d} f_t(x)\, p_0(x)\, dx}. \tag{7}$$

This expression will converge to 1 in probability if and only if

$$\frac{\int_{\Phi_t^{-1}(\mathcal{B}_a^t)^c} f_t(x)\, p_0(x)\, dx}{\int_{\Phi_t^{-1}(\mathcal{B}_a^t)} f_t(x)\, p_0(x)\, dx} \to 0 \quad \text{in probability as} \quad t \to +\infty. \tag{8}$$

It is clear that in the case where the flow is contracting (i.e., $\lim_{t\to+\infty}\alpha(t)=0$), the a priori law of X_t concentrates on the current true position, as does the conditional law. We will come back to this case later; let us begin with the most interesting case where the flow is not contracting but the observations are "good enough" so that the filter gives good results.

THEOREM 3.2. *Assume that* (H1) *and* (H2) *are fulfilled and* p_0 *is continuous, bounded. Assume moreover that*

$$\frac{V(t)}{\alpha(t)^2}\longrightarrow+\infty \tag{9}$$

and α *is bounded away from zero. Then* $\forall\ a>0$

$$\mu_t\left(\mathcal{B}_a^t\right)\xrightarrow{P}1\quad as\quad t\to+\infty.$$

Moreover, we have the following estimate concerning the rate of convergence:

$$\forall\ 0<\beta<\frac{a^2}{4C},\quad \exp\left[\beta\frac{V(t)}{\alpha^2(t)}\right]\left(1-\mu_t\left(\mathcal{B}_a^t\right)\right)\xrightarrow{P}0\quad\text{as}\quad t\to+\infty.$$

In order to prove this theorem we will need the next lemma.

LEMMA 3.3. *Let* $K(t)$ *and* $a(t)$ *be two functions from* $\mathbb{R}_+$ *into* $\mathbb{R}_+$ *such that*

$$\lim_{t\to+\infty}K(t)a(t)^2=+\infty.$$

Then $\forall\ d\in\mathbb{N}^*$, *there exists* $T_d>0$ *such that* $\forall\ t>T_d$

$$a(t)^{-d}\int_{a(t)}^{+\infty}\exp(-K(t)r^2)r^{d-1}dr\le\frac{1}{K(t)a(t)^2}\exp(-K(t)a(t)^2).$$

Proof. A simple change of variables gives

$$a(t)^{-d}\int_{a(t)}^{+\infty}\exp\left[-K(t)r^2\right]r^{d-1}\,dr=\frac{1}{2}\frac{1}{\left(a(t)^2K(t)\right)^{\frac{d}{2}}}\int_{a(t)^2K(t)}^{+\infty}e^{-z}z^{\frac{d}{2}-1}\,dz.$$

Then we conclude by observing that for $m\le 0$,

$$\int_\lambda^{+\infty}e^{-z}z^m\,dz\le e^{-\lambda}\lambda^m\qquad\forall\lambda\in\mathbb{R}_+,$$

and for $m>0$

$$\lim_{\lambda\to+\infty}\frac{\int_\lambda^{+\infty}e^{-z}z^m\,dz}{e^{-\lambda}\lambda^m}=1.\qquad\square$$

Proof of Theorem 3.2. Assume first that $\alpha(t)$ is bounded, say $\forall t\ge0$, $\alpha(t)\le A$. Then obviously

$$\mu_t\left(\mathcal{B}_a^t\right)=\mu_t^0\left(\Phi_t^{-1}\left(\mathcal{B}_a^t\right)\right)\ge\mu_t^0\left(B^0_{\frac{a}{\alpha(t)}}\right)\ge\mu_t^0\left(B^0_{\frac{a}{A}}\right),$$

and in this case the result is a direct consequence of Proposition 2.1.

Thus from now on we consider that $\alpha(t) \to +\infty$. From Propositions 2.2 and 2.3 and (H2) we have $\forall\ a > 0$

$$\frac{\alpha(t)^2}{V(t)} \sup_{\|x-X_0\| \leq \frac{a}{\alpha(t)}} \left| \int_0^t g(s,x)\, dB_s \right| \xrightarrow{P} 0$$

and

$$\frac{1}{V(t)} \sup_{\|x-X_0\| \geq \frac{a}{\alpha(t)}} \left| \frac{\int_0^t g(s,x)\, dB_s}{\|x-X_0\|^2} \right| \xrightarrow{P} 0$$

as $t \to +\infty$. We use the same kind of argument as we used in Proposition 2.1. Again, let t_n be an increasing sequence of times such that $\lim_{n\to+\infty} t_n = +\infty$, and for which the previous convergences hold a.s., and note that it is sufficient to show that $\forall a > 0$, $0 < \beta < \frac{a^2}{4C}$,

$$\exp\left[\beta \frac{V(t_n)}{\alpha(t_n)^2}\right] \frac{\int_{\Phi_{t_n}^{-1}\left(\mathcal{B}_a^{t_n,c}\right)} f_{t_n}(x)\, p_0(x)\, dx}{\int_{\Phi_{t_n}^{-1}\left(\mathcal{B}_a^{t_n}\right)} f_{t_n}(x)\, p_0(x)\, dx} \xrightarrow{\text{a.s.}} 0 \quad \text{as } n \to +\infty. \tag{10}$$

By (H2) we have $\forall\ t \geq 0$ and $a > 0$

$$\Phi_t^{-1}\left(\mathcal{B}_a^{t,c}\right) \subset \mathcal{B}^{0,c}_{\frac{a}{\alpha(t)}}$$

and

$$B^0_{\frac{a}{\alpha(t)}} \subset \Phi_t^{-1}\left(\mathcal{B}_a^t\right).$$

Then, using (H1) and (H2) and the fact that p_0 is bounded, say, by $K > 0$,

$$\begin{aligned}
&\int_{\Phi_t^{-1}\left(B_a^{t_n,c}\right)} f_{t_n}(x)\, p_0(x)\, dx \\
&\leq \int_{B^{0,c}_{\frac{a}{\alpha(t_n)}}} \exp\left[-\frac{V(t_n)}{C}\|x-X_0\|^2 \left(\frac{1}{2} \right.\right. \\
&\quad \left.\left. -\frac{C}{V(t_n)} \sup_{\|x'-X_0\| \geq \frac{a}{\alpha(t_n)}} \left| \frac{\int_0^{t_n} g(s,x')\, dB_s}{\|x'-X_0\|^2} \right| \right) \right] p_0(x)\, dx \Bigg] p_0(x)\, dx \\
&\leq K \int_{B^{0,c}_{\frac{a}{\alpha(t_n)}}} \exp\left[-\frac{V(t_n)}{4C}\|x-X_0\|^2 \right] dx
\end{aligned}$$

for n large enough, by definition of t_n. Using polar coordinates we obtain

$$\int_{\Phi_t^{-1}\left(B_a^{t_n,c}\right)} f_{t_n}(x)\, p_0(x)\, dx \leq K S_d \int_{\frac{a}{\alpha(t_n)}}^{+\infty} \exp\left[-\frac{V(t_n)}{4C} r^2 \right] r^{d-1}\, dr,$$

where S_d is the surface of the unit sphere in $\mathbb{R}^d$, with the convention $S_1 = 2$.

On the other hand, for any a_1 such that $0 < a_1 \leq a$,

$$\int_{\Phi_{t_n}^{-1}(\mathcal{B}_a^t)} f_{t_n}(x)\, p_0(x)\, dx$$

$$\geq \int_{B^0_{\frac{a_1}{\alpha(t_n)}}} \exp\left[-\frac{V(t_n)}{\alpha(t_n)^2}\left(\frac{a_1^2}{2} + \frac{\alpha(t_n)^2}{V(t_n)} \sup_{\|x'-X_0\|\leq \frac{a_1}{\alpha(t_n)}} \left|\int_0^{t_n} g(s,x')\, dB_s\right|\right)\right] p_0(x)\, dx$$

$$\geq \exp\left[-\frac{V(t_n)}{\alpha(t_n)^2}\, a_1^2\right] \int_{B^0_{\frac{a_1}{\alpha(t_n)}}} p_0(x)\, dx$$

for n large enough. As P_0 has a continuous density p_0, $p_0(X_0)$ is a.s. nonzero, and there is a small $a_2 > 0$ such that $\forall\, x \in \mathcal{B}^0_{a_2}$, $p_0(x) \geq \frac{1}{2} p_0(X_0)$. As $\alpha(t_n) \to +\infty$, for n large enough we have also that $\mathcal{B}^0_{\frac{a_1}{\alpha(t_n)}} \subset \mathcal{B}^0_{a_2}$, and thus

$$\int_{\mathcal{B}^0_{\frac{a_1}{\alpha(t_n)}}} p_0(x)\, dx \geq \frac{1}{2} p_0(X_0) \int_{\mathcal{B}^0_{\frac{a_1}{\alpha(t_n)}}} dx \geq \frac{1}{2} p_0(X_0)\, K_d \left(\frac{a_1}{\alpha(t_n)}\right)^d,$$

where K_d is the Lebesgue's measure of the unit ball in $\mathbb{R}^d$. Finally, using Lemma 3.3, we get for n large enough

$$\exp\left[\beta \frac{V(t_n)}{\alpha(t_n)^2}\right] \frac{\int_{\Phi_{t_n}^{-1}(\mathcal{B}_a^{t_n,c})} f_{t_n}(x)\, p_0(x)\, dx}{\int_{\Phi_{t_n}^{-1}(\mathcal{B}_a^{t_n})} f_{t_n}(x)\, p_0(x)\, dx}$$

$$\leq \exp\left[(\beta + a_1^2)\frac{V(t_n)}{\alpha(t_n)^2}\right] \frac{2KS_d}{K_d p_0(X_0)} \left[\frac{a}{a_1}\right]^d \left[\frac{\alpha(t_n)}{a}\right]^d \int_{\frac{a}{\alpha(t_n)}}^{+\infty} \exp\left[-\frac{V(t_n)}{4C} r^2\right] r^{d-1}\, dr$$

$$\leq \frac{2KS_d}{K_d p_0(X_0)} \left[\frac{a}{a_1}\right]^d \frac{4C\alpha(t_n)^2}{a^2 V(t_n)} \exp\left[\frac{V(t_n)}{\alpha(t_n)^2}\left(-\frac{a^2}{4C} + a_1^2 + \beta\right)\right],$$

which concludes the proof, provided we chose $a_1 < \sqrt{\frac{a^2}{4C} - \beta}$. □

Some remarks on other cases. Now let us come back to the case $\alpha(t) \to 0$. With no observation the concentration speed around the true position follows the ratio

$$\frac{\int_{\Phi_t^{-1}(\mathcal{B}_a^t)^c} p_0(x)\, dx}{\int_{\Phi_t^{-1}(\mathcal{B}_a^t)} p_0(x)\, dx} \simeq \int_{\Phi_t^{-1}(\mathcal{B}_a^t)^c} p_0(x)\, dx$$

because the denominator converges to 1 a.s. If we have observations and (H1) is fulfilled, we can show a faster convergence in probability. Let t_n be a sequence with the same properties as in the previous proof. Choose a_2 such that

$$a > \frac{a}{2\sqrt{C}} > a_2 > 0.$$

Using the same notations and arguments as in the proof of the previous theorem and using the following notations:

$$N_n = \int_{\Phi_{t_n}^{-1}(\mathcal{B}_a^{t_n})^c} f_{t_n}(x)\, dx\,,$$

$$D_n = \int_{\Phi_{t_n}^{-1}(\mathcal{B}_a^{t_n})} f_{t_n}(x)\, dx\,,$$

we get, for n large enough,

$$N_n \leq \int_{\Phi_{t_n}^{-1}(\mathcal{B}_a^t)^c} \exp\left[\frac{v(t_n)\alpha(t_n)}{a}\|x-X_0\|^2\left(-\frac{1}{2}\frac{a}{\alpha(t_n)}\right.\right.$$
$$\left.\left.+\frac{a}{\alpha(t_n)}\frac{1}{v(t_n)}\sup_{x'\in\mathcal{B}^{0,c}_{\frac{a}{\alpha(t_n)}}}\frac{\left|\int_0^{t_n} g(s,x').dB_s\right|}{\|x'-X_0\|^2}\right)\right] p_0(x)\,dx$$
$$\leq \exp\left[-\frac{1}{4}v(t_n)\frac{a^2}{\alpha(t_n)^2}\right]\int_{\Phi_{t_n}^{-1}(\mathcal{B}_a^t)^c} p_0(x)\,dx$$

and

$$D_n \geq \int_{\mathcal{B}^0_{\frac{a_2}{\alpha(t_n)}}} \exp\left[-\frac{1}{2}\frac{V(t_n)a_2^2}{\alpha(t_n)^2} - \sup_{x'\in\mathcal{B}^0_{\frac{a_2}{\alpha(t_n)}}}\left|\int_0^{t_n} g(s,x').dB_s\right|\right] p_0(x)\,dx$$
$$\geq \int_{\mathcal{B}^0_{\frac{a_2}{\alpha(t_n)}}} \exp\left[\frac{V(t_n)a_2}{\alpha(t_n)}\left(-\frac{1}{2}\frac{a_2}{\alpha(t_n)}\right.\right.$$
$$\left.\left.-\frac{\alpha(t_n)}{a_2V(t_n)}\sup_{x'\in\mathcal{B}^0_{\frac{a_2}{\alpha(t_n)}}}\left|\int_0^{t_n} g(s,x').dB_s\right|\right)\right] p_0(x)\,dx$$
$$\geq \exp\left[-C\frac{v(t_n)a_2^2}{\alpha(t_n)^2}\right]\int_{\mathcal{B}^0_{\frac{a_2}{\alpha(t_n)}}} p_0(x)\,dx$$
$$\geq \frac{1}{2}\exp\left[-C\frac{v(t_n)a_2^2}{\alpha(t_n)^2}\right].$$

By taking the two last inequalities we have

$$\frac{N_n}{D_n} \leq 2\exp\left[\frac{v(t_n)}{\alpha(t_n)^2}\left(-\frac{a^2}{4}+Ca_2^2\right)\right]\int_{\Phi_{t_n}^{-1}(\mathcal{B}_a^t)^c} p_0(x)\,dx.$$

So the convergence, this time in probability, is sped up by the exponential term.

Finally, consider some cases when the filter does not give good results. More precisely, consider the opposite of (H2).

$$(\overline{H2}). \qquad \|\Phi_t(x_1)-\Phi_t(x_2)\| \geq \alpha(t)\,\|x_1-x_2\|$$

$\forall\ x_1,\ x_2$ in $\mathbb{R}^d$ and $\forall\ t>0$ and for some nondecreasing function α from $\mathbb{R}_+$ into $\mathbb{R}_+$. We can show the following proposition (opposite of Theorem 3.2).

PROPOSITION 3.4. *Assume that* (H1) *and* ($\overline{\text{H2}}$) *are fulfilled,* p_0 *is continuous, bounded. Moreover assume that*

$$\frac{V(t)}{\alpha(t)^2} \longrightarrow 0 \tag{11}$$

and

$$\alpha(t) \longrightarrow +\infty \tag{12}$$

as $t \to +\infty$. *Then* $\forall\ a > 0$

$$\mu_t\left(\mathcal{B}_a^t\right) \xrightarrow{P} 0 \quad as \quad t \to +\infty.$$

Proof. We will show equivalently that

$$\frac{\int_{\Phi_t^{-1}(\mathcal{B}_a^t)^c} f_t(x)dx}{\int_{\Phi_t^{-1}(\mathcal{B}_a^t)} f_t(x)dx} \xrightarrow{P} +\infty \quad \text{as} \quad t \to +\infty.$$

Note that in this case it is reasonable to suppose (12); otherwise (11) would imply that $V(t) \to 0$, that is, by (H1), h is constant. From Propositions 2.2 and 2.3, using (11), we have

$$\frac{1}{\alpha(t)^2} \sup_{\|x'-X_0\| \geq \frac{a}{\alpha(t)}} \frac{\left|\int_0^t g(s,x').dB_s\right|}{\|x'-X_0\|^2} \xrightarrow{P} 0$$

and

$$\sup_{\|x'-X_0\| \leq \frac{a}{\alpha(t)}} \left|\int_0^t g(s,x').dB_s\right| \xrightarrow{P} 0\ .$$

Let t_n be a sequence of positive real numbers, increasing to $+\infty$, and such that the above convergences take place a.s. For any $b > a$ let

$$\mathcal{A}_n = \left\{x \in \mathbb{R}^d\ :\ \frac{a}{\alpha(t_n)} \leq \|x - X_0\| \leq \frac{b}{\alpha(t_n)}\right\}\ .$$

We obtain $\forall\ b > a$

$$\begin{aligned}
N_n &= \int_{\Phi_{t_n}^{-1}(B_a^{t_n,c})} f_{t_n}(x)\,dx \\
&\geq \int_{\mathcal{A}_n} \exp\left[\alpha(t_n)^2\|x-X_0\|^2\left(-\frac{V(t_n)}{2\alpha(t_n)^2}\right.\right. \\
&\qquad \left.\left. -\frac{1}{\alpha(t_n)^2}\sup_{\|x'-X_0\| \geq \frac{a}{\alpha(t_n)}} \frac{\left|\int_0^{t_n} g(s,x').dB_s\right|}{\|x'-X_0\|^2}\right)\right] p_0(x)\,dx \\
&\geq \int_{\mathcal{A}_n} \exp\left[b^2\left(-\frac{V(t_n)}{2\alpha(t_n)^2} - \frac{1}{\alpha(t_n)^2}\sup_{\|x'-X_0\| \geq \frac{a}{\alpha(t_n)}} \frac{\left|\int_0^{t_n} g(s,x').dB_s\right|}{\|x'-X_0\|^2}\right)\right] p_0(x)\,dx \\
&\geq \frac{1}{2}\int_{\mathcal{A}_n} p_0(x)\,dx
\end{aligned}$$

for n large enough. Using (12) and the continuity of p_0 we get

$$N_n \geq \frac{p_0(X_0)}{4} K_d \frac{b^d - a^d}{\alpha(t_n)^d} .$$

On the other hand,

$$\begin{aligned} D_n &= \int_{\Phi_{t_n}^{-1}(B_a^{t_n})} f_{t_n}(x)\, dx \\ &\leq \int_{\left\{ \|x - X_0\| \leq \frac{a}{\alpha(t_n)} \right\}} \exp\left[\sup_{\|x' - X_0\| \leq \frac{a}{\alpha(t_n)}} \left| \int_0^{t_n} g(s, x').dB_s \right| \right] p_0(x)\, dx \\ &\leq 2 p_0(X_0)\, K_d \frac{a^d}{\alpha(t_n)^d} , \end{aligned}$$

again for n large. Then, for arbitrary large $M > 0$, we have that for n large enough,

$$\frac{N_n}{D_n} \geq \frac{1}{8} \frac{b^d - a^d}{a^d} \geq M ,$$

provided we chose

$$b \geq a\, (8M + 1)^{\frac{1}{d}} . \qquad \square$$

4. Linear Gaussian case. The solution of the Riccati equation below and some ideas of this small section are taken from [4]. We consider throughout this section the following linear system:

$$\left\{ \begin{array}{rcll} dX_t &=& A X_t\, dt, & \\ dY_t &=& C\, X_t\, dt + dB_t, & \text{with } X_0 \simeq N(m_0, P_0), \end{array} \right.$$

where A and C are matrices of respective dimensions $d \times d$ and $d \times m$, and X_0 and B are independent. We will investigate conditions ensuring that the conditional covariance given by the Kalman–Bucy filter tends to 0 as $t \to +\infty$. It is well known that in this case the conditional law is Gaussian, so the above convergence is equivalent to the concentration of the law around the true position. See [10] or [14] for a general presentation of the Kalman–Bucy filter.

Denote by P_t the conditional covariance. In our particular case we can write the Riccati equation for P_t as

$$\dot{P}_t = A\, P_t + P_t\, A^* - P_t\, C^*\, C\, P_t, \tag{13}$$

where $*$ denotes the transpose and the dot the differential with respect to t, P_0 being the initial condition. We can solve this equation explicitly. First we state the following lemma.

LEMMA 4.1. *Let M and N be two symmetric positive semidefinite $h \times h$ matrices (i.e., $\langle N\, x,\, x \rangle \geq 0$ and $\langle M\, x,\, x \rangle \geq 0\ \forall\ x$). Then if M or N is invertible, $I + MN$ is also invertible.*

Proof. Suppose N is invertible (the case M invertible is similar). Then $N > 0$ and so is N^{-1}. Thus $N^{-1} + M > 0$ and we conclude that $(N^{-1} + M)N = I + MN$ is also invertible. $\square$

Now we can solve (13). This is the purpose of the next lemma.

LEMMA 4.2. *If P_0 is invertible, then P_t is also and*

$$P_t = e^{tA}\ \left(P_0^{-1} + Q(t)\right)^{-1}\ e^{tA^*}$$

with

$$Q(t) = \int_0^t e^{rA^*}\, C^*\, C\, e^{rA}\, dr.$$

Proof. The solution is unique because $P \mapsto F(P) = AP + PA^* - PC^*CP$ is locally Lipschitz. As P_0 is invertible, by Lemma 4.1, $I + Q(t)P_0$ is also invertible and so is P_t rewritten as

$$P_t = e^{tA}\, P_0\ (I + Q(t)\, P_0)^{-1}\ e^{tA^*}\ .$$

Then a simple computation shows that P_t is a solution of (13). ∎

Now we can state the result.

PROPOSITION 4.3. *Assume that P_0 is invertible. Then $P_t \to 0$ as $t \to +\infty$ if and only if the pair (A, C) is detectable, and for any eigenvalue λ of A, $\Re\mathrm{e}(\lambda) \leq 0$.*

Proof. First recall that P_t is symmetric positive definite. As P_t is diagonalizable, with real positive eigenvalues, showing the convergence of P_t is equivalent to the following: $\forall x \in \mathbb{R}^d$, $x^* P_t^{-1} x \to +\infty$, or more simply for x in a basis of $\mathbb{R}^d$. Note that it is also equivalent to consider complex vectors: take $x = x_1 + ix_2$ with x_1 and x_2 in $\mathbb{R}^d$, and write

$$x^*\, P_t^{-1}\, x = (x_1^* - ix_2^*)\, P_t^{-1}\, (x_1 + ix_2) = x_1^*\, P_t^{-1}\, x_1 + x_2^*\, P_t^{-1}\, x_2.$$

So throughout the proof we will work in $\mathbb{C}^d$.

Consider the Jordan form of the matrix A, say J, and call T the corresponding basis transformation. Then $x^* P_t^{-1} x \to +\infty\ \forall x \in \mathbb{C}^d$ if and only if $u_i^* P_t^{-1} u_i \to +\infty\ \forall u_i$ in the Jordan basis.

We have two cases. First assume that u_i is in the stable subspace of A, i.e., associated with a submatrix J_k of the Jordan form and an eigenvalue λ_k with strictly negative real part. Then obviously

$$u_i^*\, P_t^{-1}\, u_i \geq u_i^*\, e^{-tA^*}\, P_0^{-1}\, e^{-tA}\, u_i = \left\|e^{-tA}\, u_i\right\|^2_{P_0^{-1}} \to +\infty.$$

Second case: u_i is in the unstable subspace of A, i.e., associated with a Jordan submatrix with pure imaginary eigenvalue. In this case we write

$$\begin{aligned} u_i^*\, P_t^{-1}\, u_i &\geq \int_0^t u_i^*\, e^{-(t-s)A^*}\, C^*\, C\, e^{-(t-s)A}\, u_i\, ds \\ &= \int_0^t \left\|C\, e^{-(t-s)A}\, u_i\right\|^2 ds \\ &= \int_0^t \left\|C\, T\, e^{-(t-s)J}\, T^{-1}\, u_i\right\|^2 ds. \end{aligned} \tag{14}$$

Let m_k be the dimension of the subblock corresponding to u_i, k_0 the rank of the first term of the subblock, and ℓ the rank of u_i in the corresponding basis of the eigenspace.

J and $e^{-(t-s)J}$ are block diagonal (see [10] or [5] for precisions on the Jordan form). Denoting by J_k the corresponding subblock, we have

$$e^{-(t-s)J_k} = e^{-\lambda_k(t-s)} \begin{bmatrix} 1 & \dots & \frac{-(t-s)^{m_k-1}}{(m_k-1)!} \\ & \ddots & \vdots \\ 0 & & 1 \end{bmatrix}.$$

Then observe that, M being the number of subblocks,

$$\begin{aligned} &\|C\,T\,e^{-(t-s)J}\,(T^{-1}\,u_i)\|^2 \\ &= \left\| C\,T \begin{bmatrix} e^{-(t-s)J_1} & & & & \\ & \ddots & & 0 & \\ & & e^{-(t-s)J_k} & & \\ & 0 & & \ddots & \\ & & & & e^{-(t-s)J_M} \end{bmatrix} \cdot \begin{bmatrix} 0 \\ \vdots \\ 0 \\ 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix} \right\|^2 \\ &\geq \|C\,u_{k_0}\|^2 \left(\frac{(t-s)^{\ell-1}}{(\ell-1)!} \right)^2 - \sum_{q=1}^{l-1} \|C\,u_{k_0+q}\|^2 \left(\frac{(t-s)^{q-1}}{(q-1)!} \right)^2. \end{aligned}$$

Remark that u_{k_0} is an eigenvector of A, so, by the dual version of the Hautus lemma[1] (see [15, pp. 93 and 200]) and the assumption that (A, C) is detectable, $u_{k_0} \notin \ker C$. Thus $\|Cu_{k_0}\| > 0$ and for t large enough, (14) is bigger than, say, $K_i \int_0^t (t-s)^{2\ell-2}\,ds$ for some $K_i > 0$. This gives the divergence to $+\infty$.

Now we prove the converse. We again have two cases. First assume that there is some eigenvalue of A, say λ_i, with $\Re\mathrm{e}(\lambda_i) > 0$. Let u_i be a normalized associated eigenvector. Then

$$u_i^*\,e^{-tA^*}\,P_0^{-1}\,e^{-tA}\,u_i \longrightarrow 0$$

and

$$\int_0^t \left\| C\,e^{-(t-s)A}\,u_i \right\|^2 ds \leq \|C\|^2 \int_0^t e^{-2(t-s)\Re\mathrm{e}(\lambda_i)}\,ds < +\infty,$$

which proves that $P_t \not\to 0$.

Finally, consider the second case: assume that all the eigenvalues of A have non-positive real parts but that the pair (A, C) is not detectable. Then, again using the Hautus lemma, there exists an eigenvector u_i of A associated to some eigenvalue λ_i such that $u_i \in \ker C$ and $\Re\mathrm{e}(\lambda_i) = 0$. Then we have

$$u_i^*\,e^{-tA^*}\,P_0^{-1}\,e^{-tA}\,u_i \leq \|u_i\|^2_{P_0^{-1}}$$

and

$$\int_0^t \left\| C\,e^{-(t-s)A} u_i \right\|^2 ds = 0,$$

[1] Which tells us that (A, C) is observable if and only if $\begin{bmatrix} \lambda_i I - A \\ C \end{bmatrix}$ is full rank $\forall\ \lambda_i$ eigenvalues of A. Here we apply this to the unstable part of the system.

which concludes the proof. □

REMARK 4.4. *Keeping in mind that P_t is diagonalizable, the last proof shows that when P_t goes to 0 and the unstable subspace of A is nontrivial, the smallest eigenvalue of P_t^{-1} is at least of order t (take $\ell = 1$), that is, the largest eigenvalue of P_t is at most of order $1/t$, which gives an estimate on the order of convergence.*

Acknowledgments. I would like to thank Étienne Pardoux and Fabien Campillo for their support and interest during this work and Alexander Davie and Jean-Baptiste Pomet for valuable discussions about possible generalizations. I am also very grateful to Daniel Ocone for suggesting some improvements to this paper. Finally let me thank the reviewers for their interest in this work.

REFERENCES

[1] H. BERNIER, F. CAMPILLO, F. CÉROU, AND F. LE GLAND, *Parallélisme de données et filtrage non linéaire—analyse de performance*, Rapport Technique 167, INRIA, 1994.

[2] R. S. BUCY AND P. D. JOSEPH, *Filtering for Stochastic Processes with Application to Guidance*, Interscience, New York, 1968.

[3] F. CAMPILLO, F. CÉROU, F. LE GLAND, AND R. RAKOTOZAFY, *Particle and cell approximations for nonlinear filtering*, Rapport de Recherche 2567, INRIA, 1994.

[4] D. FELLAH, *Étude du comportement asymptotique de la matrice de covariance dans le filtre de Kalman–Bucy*, Mémoire de DEA, Université de Provence, Marseille, France, 1988.

[5] M. W. HIRSCH AND S. SMALE, *Differential Equations, Dynamical Systems, and Linear Algebra*, Academic Press, New York, 1974.

[6] I. A. IBRAGIMOV AND R. Z. KHASMINSKII, *Statistical Estimation. Asymptotic Theory*, Appl. Math. 16, Springer-Verlag, New York, Berlin, 1981.

[7] D. JI, *Asymptotic Analysis of Nonlinear Filtering Problems*, Ph.D. thesis, Brown University, Providence, RI, 1988.

[8] H. KUNITA, *Asymptotic behavior of the nonlinear filtering errors of Markov processes*, J. Multivariate Anal., 1 (1971), pp. 365–393.

[9] H. KUNITA, *Ergodic properties of nonlinear filtering processes*, in Spatial Stochastic Processes: Festschrift in honor of T. E. Harris, K. L. Alexander and J. C. Watkins, eds., Progr. in Probab. 19, Birkhäuser, Boston, 1991, pp. 233–256.

[10] H. KWAKERNAAK AND R. SIVAN, *Linear Optimal Control Systems*, John Wiley, New York, 1972.

[11] M. LEDOUX AND M. TALAGRAND, *Probability in Banach Spaces. Isoperimetry and Processes*, Ergeb. Math. Grenzgeb. (3) 23, Springer-Verlag, Berlin, 1991.

[12] R. SH. LIPTSER AND A. N. SHIRYAYEV, *Statistics of Random Processes* I. *General Theory*, Appl. Math. 5, Springer-Verlag, New York, Heidelberg, 1977.

[13] D. OCONE AND E. PARDOUX, *Asymptotic stability of the optimal filter with respect to its initial condition*, SIAM J. Control Optim., 34 (1996), pp. 226–243.

[14] E. PARDOUX, *Filtrage non linéaire et équations aux dérivées partielles stochastiques associées*, in École d'Été de Probabilités de Saint-Flour 19—1989, P. L. Hennequin, ed., Lecture Notes in Math. 1464, Springer-Verlag, Berlin, 1991, pp. 67–163.

[15] E. D. SONTAG, *Mathematical Control Theory*, Texts in Appl. Math. 6, Springer-Verlag, New York, 1990.

SIAM J. CONTROL OPTIM.
Vol. 38, No. 4, pp. 1102–1119

ON THE MINIMIZING PROPERTY OF A SECOND ORDER DISSIPATIVE SYSTEM IN HILBERT SPACES*

FELIPE ALVAREZ†

Abstract. We study the asymptotic behavior at infinity of solutions of a second order evolution equation with linear damping and convex potential. The differential system is defined in a real Hilbert space. It is proved that if the potential is bounded from below, then the solution trajectories are minimizing for it and converge weakly towards a minimizer of Φ if one exists; this convergence is strong when Φ is even or when the optimal set has a nonempty interior. We introduce a second order proximal-like iterative algorithm for the minimization of a convex function. It is defined by an implicit discretization of the continuous evolution problem and is valid for any closed proper convex function. We find conditions on some parameters of the algorithm in order to have a convergence result similar to the continuous case.

Key words. dissipative system, linear damping, asymptotic behavior, weak convergence, convexity, implicit discretization, iterative-variational algorithm

AMS subject classifications. 34G20, 34A12, 34D05, 90C25

PII. S0363012998335802

1. Introduction.

Consider the following differential system defined in a real Hilbert space H:

$$u'' + \gamma u' + \nabla\Phi(u) = 0, \tag{1.1}$$

where $\gamma > 0$ and $\Phi : H \to \mathbb{R}$ is differentiable. It is customary to call this equation *nonlinear oscillator with damping*. Here, the damping or friction has a linear dependence on the velocity. This is a particular case of the so-called dissipative systems. In fact, given u solution of (1.1) define $E(t) := \frac{1}{2}|u'|^2 + \Phi(u)$; it is direct to check that $E' = -\gamma|u'|^2$. Thus, the *energy* of the system is dissipated as t increases. Although (1.1) appears in various contexts with different physical interpretations, the motivation for this work comes from the dynamical approach to optimization problems.

Roughly speaking, any iterative algorithm generating a sequence $\{x_k\}_{k\in\mathbb{N}}$ may be considered as a discrete dynamical system. If it is possible to find a continuous version for the discrete procedure, one expects that the properties of the corresponding continuous dynamical system are close to those of the discrete one. This occurs, for instance, for the now classical proximal method for convex minimization: given $x_0 \in H$, solve the iterative scheme

$$\frac{x_{k+1} - x_k}{\lambda_k} + \partial f(x_{k+1}) \ni 0, \tag{Prox}$$

where $\lambda_k > 0$, $f : H \to \mathbb{R} \cup \{\infty\}$ is a closed proper convex function and ∂f denotes the usual subdifferential in convex analysis. (Prox) is an implicit discretization for the steepest descent method, which consists of solving the following differential inclusion:

$$x' + \partial f(x) \ni 0. \tag{SD}$$

*Received by the editors April 20, 1998; accepted for publication (in revised form) September 17, 1999; published electronically April 4, 2000. This work was partially supported by FONDECYT grants 1961131 and 1990884.

http://www.siam.org/journals/sicon/38-4/33580.html

†Departamento de Ingeniería Matemática, Universidad de Chile, Beauchef 850, Santiago, Chile (falvarez@dim.uchile.cl).

Under suitable conditions, both the trajectory $\{x(t) : t \to \infty\}$ defined by (SD) and the sequence $\{x_k\}$ generated by (Prox) converge toward a particular minimizer of f (see [5, 6, 7] for (SD) and [18] for (Prox); see also [12] for a survey on these and new results). The dynamical approach to iterative methods in optimization has many advantages. It provides a deep insight into the expected behavior of the method, and sometimes the techniques used in the continuous case can be adapted to obtain results for the discrete algorithm. On the other hand, a continuous dynamical system satisfying nice properties may suggest new iterative methods.

This viewpoint has motivated increasing attention in recent years; see, e.g., [1, 2, 3, 4, 8, 13, 14]. In [3], Attouch, Goudou, and Redont deal with nonconvex functions that have, a priori, many local minima. The idea is to exploit the dynamics defined by (1.1) to explore critical points of Φ (i.e., solutions of $\nabla\Phi(x) = 0$). If Φ is coercive (bounded level sets) and of class C^1 with a locally Lipschitz gradient, then it is possible to prove that for any u solution of (1.1) we have $\nabla\Phi(u(t)) \to 0$ as $t \to \infty$. The convergence of the trajectory $\{u(t) : t \to \infty\}$ is a more delicate problem. When Φ is coercive, an obvious sufficient condition for the convergence of the trajectory is that the critical points, also known as equilibrium points, are isolated. Certainly, this is not necessary. In one dimension ($H = \mathbb{R}$) and without additional conditions, the solution always converges toward an equilibrium (see, e.g., [10]). The proof relies on topological arguments that are not generalizable to higher dimensions. Indeed, this is no longer true even in two dimensions: it is possible to construct a coercive C^1 function defined on $\mathbb{R}^2$ whose gradient is locally Lipschitz and for which at least one solution of (1.1) does not converge as $t \to \infty$ (see [3]). Thus, a natural question is to find general conditions under which the trajectory converges in the degenerate case, that is, when the set of equilibrium points of Φ contains a nontrivial connected component. A positive result in this direction has recently been given by Haraux and Jendoubi [11], where convergence to an equilibrium is established when Φ is *analytic*. However, this assumption is very restrictive from the optimization point of view.

Motivated by the previous considerations, in this work we focus our attention on the asymptotic behavior as $t \to \infty$ of the solutions of (1.1) when Φ is assumed to be convex. The paper is organized as follows. In section 2 we prove that if Φ is convex and bounded from below, then the trajectory $\{u(t) : t \to \infty\}$ is minimizing for Φ. If the infimum of Φ on H is attained, then $u(t)$ converges weakly towards a minimizer of Φ. The convergence is strong when Φ is even or when the optimal set has a nonempty interior. In section 2.2 we give a localization result for the limit point, analogous to the corresponding result for the steepest descent method [13]. In section 2.4 we generalize the convergence result to cover the equation $u'' + \Gamma u' + \nabla\Phi(u) = 0$, where $\Gamma : H \to H$ is a bounded self-adjoint linear operator which we assume to be elliptic: there is $\gamma > 0$ such that for any $x \in H$, $\langle \Gamma x, x\rangle \geq \gamma|x|^2$. We refer to this equation as nonlinear oscillator with *anisotropic damping*. This equation appears to be useful to diminish oscillations or even eliminate them, and also to accelerate the convergence of the trajectory. In section 2.3 we give an heuristic motivation of the above mentioned facts, which is based on an analysis of a quadratic function. Still under the convexity condition on Φ, section 3 deals with the discretization of (1.1). Here, we consider the *implicit* scheme

$$\frac{u_{k+1} - 2u_k + u_{k-1}}{h^2} + \gamma\frac{u_{k+1} - u_k}{h} + \nabla\Phi(u_{k+1}) = 0,$$

where $h > 0$. Since Φ is convex, the latter is equivalent to the following variational problem:

$$u_{k+1} = \operatorname{argmin}\left\{\Phi(x) + \frac{1+\gamma h}{2h^2}\,|x - z_k|^2 : x \in H\right\},$$

where $z_k = u_k + \frac{1}{1+\gamma h}(u_k - u_{k-1})$. This procedure does not require Φ to be differentiable and allows us to introduce the following more general iterative-variational algorithm:

$$\frac{1}{\lambda_k}(u_{k+1} - (1+\alpha_k)u_k + \alpha_k u_{k-1}) + \partial_{\epsilon_k} f(u_{k+1}) \ni 0, \tag{1.2}$$

where $\epsilon_k, \lambda_k > 0$, $\alpha_k \in [0,1[$, $f : H \to \mathbb{R} \cup \{\infty\}$ is a closed proper convex function and $\partial_\epsilon f$ is the ϵ-approximate subdifferential in convex analysis. We call (1.2) the *inertial proximal method*. We find conditions on the parameters α_k, ϵ_k, and λ_k in order to have a convergence result similar to the continuous case. Finally, in section 4 we state some of the questions opened by this work. Let us mention that the first to consider (1.1) for finite dimensional optimization problems was B. T. Polyack [16]. He studied a two-step discrete algorithm called the "heavy-ball with friction" method, which may be interpreted as an *explicit* discretization of (1.1). Both approaches are complementary; however, the analysis and the type of results in the implicit and explicit cases are different.

2. Dissipative differential system. Throughout this paper, H is a real Hilbert space, $\langle\cdot,\cdot\rangle$ denotes the associated inner product, and $|\cdot|$ stands for the corresponding norm. We are interested in the behavior at infinity of $u : [0,\infty[\to H$, a solution of the following abstract evolution equation:

$$(E_\gamma; u_0, v_0) \qquad \begin{cases} u'' + \gamma u' + \nabla\Phi(u) = 0, \\ u(0) = u_0,\ u'(0) = v_0, \end{cases}$$

where $\gamma > 0$, $\Phi : H \to \mathbb{R}$, and $u_0, v_0 \in H$ are given. Note that if we assume that the gradient $\nabla\Phi$ is locally Lipschitz, then the existence and uniqueness of a local solution for $(E_\gamma; u_0, v_0)$ follow from standard results of differential equations theory. In that case, to prove that u is infinitely extendible to the right, it suffices to show that its derivative u' is bounded. Set

$$E(t) := \frac{1}{2}|u'(t)|^2 + \Phi(u(t)).$$

Since $E'(t) = -\gamma|u'(t)|^2$, the function E is nonincreasing. If we suppose that Φ is bounded from below, then u' is bounded.

2.1. Asymptotic convergence. In that which follows, we suppose the existence of a global solution of $(E_\gamma; u_0, v_0)$. We write $\inf\Phi$ for the infimum value of Φ on H; thus, $\inf\Phi > -\infty$ will mean that Φ is bounded from below. We denote by $\operatorname{Argmin}\Phi$ the set $\{x \in H : \Phi(x) = \inf\Phi\}$. On the nonlinearity we shall assume

$$(h_\Phi) \qquad \Phi \in C^1(H;\mathbb{R}) \text{ is convex and } \inf\Phi > -\infty.$$

THEOREM 2.1. *Suppose that (h_Φ) holds. If $u \in C^2([0,\infty[;H)$ is a solution of $(E_\gamma; u_0, v_0)$, then $u' \in L^2([0,\infty[;H)$, $u'(t) \to 0$ as $t \to \infty$, and*

$$\lim_{t\to\infty} \Phi(u(t)) = \inf\Phi. \tag{2.1}$$

Furthermore, if $\operatorname{Argmin}\Phi \neq \emptyset$, then there exists $\widehat{u} \in \operatorname{Argmin}\Phi$ such that $u(t) \rightharpoonup \widehat{u}$ weakly in H as $t \to \infty$.

We begin by noticing that u' is bounded (see the argument above). In order to prove the minimizing property (2.1), it suffices to prove that

$$\limsup_{t\to\infty} \Phi(u(t)) \le \Phi(x)$$

for any $x \in H$. Fix $x \in H$ and define the auxiliary function $\varphi(t) := \frac{1}{2}|u(t)-x|^2$. Since u is a solution of (E_γ), it follows that

$$\varphi'' + \gamma\varphi' = \langle \nabla\Phi(u), x-u\rangle + |u'|^2,$$

which together with the convexity inequality $\Phi(u) + \langle \nabla\Phi(u), x-u\rangle \le \Phi(x)$ yields

$$\varphi'' + \gamma\varphi' \le \Phi(x) - \Phi(u) + |u'|^2. \tag{2.2}$$

We do not have information on the behavior of $\Phi(u(t))$ but we know that $E(t)$ is nonincreasing. Thus, we rewrite (2.2) as

$$\varphi'' + \gamma\varphi' \le \Phi(x) - E(t) + \frac{3}{2}|u'|^2.$$

Given $t > 0$, for all $\tau \in [0,t]$ we have

$$\varphi''(\tau) + \gamma\varphi'(\tau) \le \Phi(x) - E(t) + \frac{3}{2}|u'(\tau)|^2.$$

After multiplication by $e^{\gamma\tau}$ and integration we obtain

$$\varphi'(t) \le e^{-\gamma t}\varphi'(0) + \frac{1}{\gamma}(1-e^{-\gamma t})[\Phi(x) - E(t)] + \frac{3}{2}\int_0^t e^{-\gamma(t-\tau)}|u'(\tau)|^2 d\tau.$$

We write this equation with t replaced by θ, and use the fact that $E(t)$ decreases and integrate once more to obtain

$$\varphi(t) \le \varphi(0) + \frac{1}{\gamma}(1-e^{-\gamma t})\varphi'(0) + \frac{1}{\gamma^2}(\gamma t - 1 + e^{-\gamma t})[\Phi(x) - E(t)] + h(t), \tag{2.3}$$

where

$$h(t) := \frac{3}{2}\int_0^t \int_0^\theta e^{-\gamma(\theta-\tau)}|u'(\tau)|^2 d\tau d\theta.$$

Since $E(t) \ge \Phi(u(t))$, (2.3) gives

$$\frac{1}{\gamma^2}(\gamma t - 1 + e^{-\gamma t})\Phi(u(t)) \le \varphi(0) + \frac{1}{\gamma}(1-e^{-\gamma t})\varphi'(0) + \frac{1}{\gamma^2}(\gamma t - 1 + e^{-\gamma t})\Phi(x) + h(t).$$

Dividing this inequality by $\frac{1}{\gamma^2}(\gamma t - 1 + e^{-\gamma t})$ and letting $t \to \infty$ we get

$$\limsup_{t\to\infty} \Phi(u(t)) \le \Phi(x) + \limsup_{t\to\infty} \frac{\gamma}{t} h(t).$$

It suffices to show that $h(t)$ remains bounded as $t \to \infty$. By Fubini's theorem

$$h(t) = \frac{3}{2}\int_0^t \int_\tau^t e^{-\gamma(\theta-\tau)}|u'(\tau)|^2 d\theta d\tau = \frac{3}{2\gamma}\int_0^t |u'(\tau)|^2(1-e^{-\gamma(t-\tau)})d\tau.$$

Note that from the equality $E' = -\gamma|u'|^2$ it follows that

$$\frac{1}{2}|u'|^2 + \Phi(u) + \gamma \int_0^t |u'(\tau)|^2 d\tau = E_0,$$

and in particular,

$$\int_0^t |u'(\tau)|^2 d\tau \leq \frac{E_0 - \inf \Phi}{\gamma} < \infty.$$

Then $u' \in L^2([0,\infty[; H)$, and

$$h(t) \leq \frac{3}{2\gamma} \int_0^t |u'(\tau)|^2 d\tau \leq \frac{3}{2\gamma} \int_0^\infty |u'(\tau)|^2 d\tau < \infty.$$

On the other hand, since $E(\cdot)$ is nonincreasing and bounded from below by $\inf \Phi$, it converges as $t \to \infty$. If $\lim_{t\to\infty} E(t) > \inf \Phi$, then $\lim_{t\to\infty} |u'(t)| > 0$ because of (2.1). This contradicts the fact that $u' \in L^2$. Therefore, $\lim_{t\to\infty} E(t) = \inf \Phi$, hence $u'(t) \to 0$ as $t \to \infty$.

The task now is to establish the weak convergence of $u(t)$ when $\operatorname{Argmin} \Phi \neq \emptyset$. For this purpose, we shall apply the Opial lemma [15], which holds interest in that it allows one to prove convergence without knowing the limit point. We state it as follows.

LEMMA (Opial). *Let H be a Hilbert space, let $\{u(t) : t \to \infty\} \subset H$ be a trajectory, and denote by W the set of its weak limit points*

$$W := \{y \in H : \exists t_k \to \infty \ s.t. \ u(t_k) \rightharpoonup y\}.$$

If there exists $\emptyset \neq S \subset H$ such that

$$\forall z \in S, \ \lim_{t\to\infty} |u(t) - z| \ exists, \tag{2.4}$$

then $W \neq \emptyset$. Moreover, if $W \subset S$, then $u(t)$ converges weakly toward $\widehat{u} \in S$ as $t \to \infty$.

In order to apply the above result, we must find an adequate set S. Suppose that there exists $\widehat{u} \in H$ such that $u(t_k) \rightharpoonup \widehat{u}$ for a suitable sequence $t_k \to \infty$. The function Φ is weak lower-semicontinuous, because Φ is convex and continuous; hence

$$\Phi(\widehat{u}) \leq \liminf_{k\to\infty} \Phi(u(t_k)) = \lim_{t\to\infty} \Phi(u(t)) = \inf \Phi,$$

and therefore $\widehat{u} \in \operatorname{Argmin} \Phi$. According to the Opial lemma, we have only to prove that

$$\forall z \in \operatorname{Argmin} \Phi, \ \lim_{t\to\infty} |u(t) - z| \ \text{ exists.}$$

For this, fix $z \in \operatorname{Argmin} \Phi$ and define $\varphi(t) := \frac{1}{2}|u(t) - z|^2$. The following lemma provides a sufficient condition on $[\varphi']_+$, the positive part of the derivative, in order to ensure convergence for φ.

LEMMA 2.2. *Let $\theta \in C^1([0,\infty[; \mathbb{R})$ be bounded from below. If $[\theta']_+ \in L^1([0,\infty[; \mathbb{R})$, then $\theta(t)$ converges as $t \to \infty$.*

Proof. Set

$$w(t) := \theta(t) - \int_0^t [\theta'(\tau)]_+ d\tau.$$

Since $w(t)$ is bounded from below and $w'(t) \leq 0$, then $w(t)$ converges as $t \to \infty$, and consequently $\theta(t)$ converges as $t \to \infty$. □

On account of this result, it suffices to prove that $[\varphi']_+$ belongs to $L^1(0,\infty)$. Of course, to obtain information on φ' we shall use the fact that $u(t)$ is solution of (E_γ). Due to the optimality of z, it follows from (2.2) that

$$\varphi'' + \gamma\varphi' \leq |u'|^2. \tag{2.5}$$

LEMMA 2.3. *If $\omega \in C^1([0,\infty[;\mathbb{R})$ satisfies the differential inequality*

$$\omega' + \gamma\omega \leq g(t) \tag{2.6}$$

with $\gamma > 0$ and $g \in L^1([0,\infty[;\mathbb{R})$, then $[\omega]_+ \in L^1([0,\infty[;\mathbb{R})$.

Proof. We can certainly assume that $g \geq 0$, for if not, we replace g by $|g|$. Multiplying (2.6) by $e^{\gamma t}$ and integrating we get

$$\omega(t) \leq e^{-\gamma t}\omega(0) + \int_0^t e^{-\gamma(t-\tau)} g(\tau)\, d\tau.$$

Thus

$$[\omega(t)]_+ \leq e^{-\gamma t}[\omega(0)]_+ + \int_0^t e^{-\gamma(t-\tau)} g(\tau)\, d\tau,$$

and Fubini's theorem gives $\int_0^\infty \int_0^t e^{-\gamma(t-\tau)} g(\tau)\, d\tau dt = \frac{1}{\gamma}\int_0^\infty g(\tau)\, d\tau < \infty$. □

Recalling that $|u'|^2 \in L^1([0,\infty[;\mathbb{R})$, the proof of the theorem is completed by applying Lemma 2.3 to (2.5). □

We say that $\nabla\Phi$ is *strongly monotone* if there exists $\beta > 0$ such that for any $x, y \in H$ we have

$$\langle \nabla\Phi(x) - \nabla\Phi(y), x - y\rangle \geq \beta |x - y|^2.$$

A weaker condition is the strong monotonicity over bounded sets, that is to say, for all $K > 0$ there exists $\beta_K > 0$ such that for any $x, y \in B[0,K]$ we have

$$\langle \nabla\Phi(x) - \nabla\Phi(y), x - y\rangle \geq \beta_K |x - y|^2. \tag{2.7}$$

If the latter property holds, then we have strong convergence for $u(t)$ when the infimum of Φ is attained. The argument is standard: let $\widehat{u}$ be the (unique) minimum point for Φ and set $K := \max\{\sup_{t\geq 0} |u(t)|, |\widehat{u}|\}$; then from (2.7) we deduce

$$\Phi(\widehat{u}) + \frac{\beta_K}{2}|u(t) - \widehat{u}|^2 \leq \Phi(u(t)). \tag{2.8}$$

Since we have proven that $\lim_{t\to\infty} \Phi(u(t)) = \inf \Phi = \Phi(\widehat{u})$, estimate (2.8) implies $u(t) \to \widehat{u}$ strongly in H. Note that we do not need to apply the Opial lemma.

The latter is the case of a nondegenerate minimum point. When Φ admits multiple minima, it is not possible to obtain strong convergence without additional assumptions on Φ or the space H. For instance, we have the following.

THEOREM 2.4. *Under the hypotheses of Theorem* 2.1, *if either*

(i) $\operatorname{Argmin}\Phi \neq \emptyset$ *and* Φ *is even*

or

(ii) $\operatorname{int}(\operatorname{Argmin}\Phi) \neq \emptyset$,
then

$$u(t) \to \widehat{u} \text{ strongly in } H \text{ as } t \to \infty,$$

where $\widehat{u} \in \operatorname{Argmin}\Phi$.

Proof. The proof is adapted from the corresponding results for the steepest descent method; see [7] for the analogous hypothesis of (i) and [6] for (ii).

(i) Fix $t_0 > 0$ and define $g : [0, t_0] \to \mathbb{R}$ by

$$g(t) := |u(t)|^2 - |u(t_0)|^2 - \frac{1}{2}|u(t) - u(t_0)|^2.$$

Then $g'(t) = \langle u'(t), u(t) + u(t_0)\rangle$ and $g''(t) = \langle u''(t), u(t) + u(t_0)\rangle + |u'(t)|^2$. Consequently

$$g''(t) + \gamma g'(t) = \langle -\nabla\Phi(u(t)), u(t) + u(t_0)\rangle + |u'(t)|^2.$$

Since $E(t) = \frac{1}{2}|u'(t)|^2 + \Phi(u(t))$ is decreasing and Φ is even, we deduce that

$$E(t) \geq \frac{1}{2}|u'(t_0)|^2 + \Phi(-u(t_0))$$

for all $t \in [0, t_0]$. By the convexity of Φ we conclude that

$$E(t) \geq \frac{1}{2}|u'(t_0)|^2 + \Phi(u(t)) + \langle \nabla\Phi(u(t)), -u(t) - u(t_0)\rangle$$

and hence that

$$\frac{1}{2}|u'(t)|^2 \geq \langle -\nabla\Phi(u(t)), u(t) + u(t_0)\rangle.$$

Thus

$$g''(t) + \gamma g'(t) \leq \frac{3}{2}|u'(t)|^2.$$

The standard integration procedure yields

$$g(t_0) - g(t) \leq \frac{1}{\gamma}(e^{-\gamma t} - e^{-\gamma t_0})g'(0) + \frac{3}{2}\int_t^{t_0}\int_0^{\theta} e^{-\gamma(\theta-\tau)}|u'(\tau)|^2 d\tau d\theta.$$

Therefore, for all $t \in [0, t_0]$ we have that

$$\frac{1}{2}|u(t) - u(t_0)|^2 \leq |u(t)|^2 - |u(t_0)|^2 + \frac{1}{\gamma}(e^{-\gamma t} - e^{-\gamma t_0})g'(0) + h(t_0) - h(t), \tag{2.9}$$

where

$$h(t) = \frac{3}{2}\int_0^t \int_0^{\theta} e^{-\gamma(\theta-\tau)}|u'(\tau)|^2 d\tau d\theta.$$

On the other hand, in the proof of Theorem 2.1 we have shown that $h(t)$ is convergent as $t \to \infty$. We also proved that for all $z \in \operatorname{Argmin}\Phi$ the $\lim_{t\to\infty}|u(t) - z|$ exists. Since Φ is convex and even, we have $0 \in \operatorname{Argmin}\Phi$ whenever the infimum is realized.

In that case, $|u(t)|$ is convergent as $t \to \infty$ and we infer from (2.9) that $\{u(t) : t \to \infty\}$ is a Cauchy net. Hence $u(t)$ converges strongly as $t \to \infty$ and, by Theorem 2.1, the limit belongs to $\operatorname{Argmin} \Phi$.

(ii) Let $z_0 \in \operatorname{int}(\operatorname{Argmin} \Phi)$. There exists $\rho > 0$ such that for every $z \in H$ with $|z - z_0| \leq \rho$, then $z \in \operatorname{int}(\operatorname{Argmin} \Phi)$. In particular, if $|z - z_0| \leq \rho$, then $\nabla\Phi(z) = 0$. Consequently,

$$\langle \nabla\Phi(x), x - z_0 \rangle \geq \langle \nabla\Phi(x), z - z_0 \rangle$$

for every $x \in H$ and z with $|z - z_0| \leq \rho$. Hence,

$$\langle \nabla\Phi(x), x - z_0 \rangle \geq \rho |\nabla\Phi(x)|$$

for every $x \in H$. Applying this inequality to $x = u(t)$ we deduce that

$$-\langle u'' + \gamma u', u - z_0 \rangle \geq \rho |u'' + \gamma u'|.$$

Set $\varphi(t) := \frac{1}{2}|u(t) - z_0|^2$. We thus obtain

$$-\varphi'' + |u'|^2 - \gamma\varphi' \geq \rho |u'' + \gamma u'|.$$

Integrating this inequality yields

$$\varphi'(0) - \varphi'(t) + \int_0^t |u'(\tau)|^2 d\tau + \gamma(\varphi(0) - \varphi(t)) \geq \rho \int_0^t |u''(\tau) + \gamma u'(\tau)| d\tau.$$

We have already proved that the $\lim_{t\to\infty} \varphi(t)$ exists and $\lim_{t\to\infty} \varphi'(t) = 0$. Moreover, $u' \in L^2(0, \infty; H)$. As a conclusion, $u'' + \gamma u' \in L^1(0, \infty; H)$. We deduce that the $\lim_{t\to\infty} u'(t) + \gamma u(t)$ exists, which finishes the proof because $u'(t) \to 0$ as $t \to \infty$. □

2.2. Localization of the limit point. In the proof of Theorem 2.1 we have used the differential inequality (2.2), which in some sense measures the evolution of the system. A simpler but analogous inequality appears in the asymptotic analysis for the steepest descent inclusion (SD). This was used by B. Lemaire in [13] to *locate* the limit point of the trajectories of (SD). Following this approach, in this section we give a localization result of the limit point of the solutions of (E_γ). For simplicity of notation, set $S := \operatorname{Argmin} \Phi$ and we denote by $\operatorname{proj}_S : H \to S$ the projection operator onto the closed convex set S.

PROPOSITION 2.5. *Let u be solution of $(E_\gamma; u_0, v_0)$ and $\widehat{u} \in S$ be such that $u(t) \rightharpoonup \widehat{u}$ weakly as $t \to \infty$. Then, for all $x \in S$*

$$|\widehat{u} - x| \leq |u_0 + \frac{1}{\gamma} v_0 - x| + \frac{1}{\gamma} \delta(u_0), \tag{2.10}$$

where $\delta(u_0) = \sqrt{2}[\Phi(u_0) - \inf \Phi]^{1/2}$. Consequently

$$\text{(i)} \qquad |\widehat{u} - \operatorname{proj}_S(u_0 + \frac{1}{\gamma} v_0)| \leq d(u_0 + \frac{1}{\gamma} v_0, S) + \frac{1}{\gamma} \delta(u_0),$$

where $d(u_0, S)$ is the distance between u_0 and the set S.

(ii) *If S is an affine subspace of H, then*

$$|\widehat{u} - \operatorname{proj}_S(u_0 + \frac{1}{\gamma} v_0)| \leq \frac{1}{\gamma} \delta(u_0).$$

If, moreover, Φ is a quadratic form, then

$$u(t) \to \operatorname{proj}_S(u_0 + \frac{1}{\gamma}v_0) \text{ strongly in } H \text{ as } t \to \infty.$$

Proof. Let $x \in S$ and set $\varphi(t) := \frac{1}{2}|u(t) - x|^2$. The inequality (2.2) and the optimality of x give $\varphi'' + \gamma\varphi' \le |u'|^2$. Hence

$$\varphi(t) \le \varphi(0) + \frac{1}{\gamma}(1 - e^{-\gamma t})\varphi'(0) + \int_0^t \int_0^\theta e^{-\gamma(\theta-\tau)}|u'(\tau)|^2 d\tau d\theta.$$

Due to the weak lower-semicontinuity of the norm and Fubini's theorem, we can let $t \to \infty$ to obtain

$$\frac{1}{2}|\widehat{u} - x|^2 \le \frac{1}{2}|u_0 - x|^2 + \frac{1}{\gamma}\langle v_0, u_0 - x\rangle + \frac{1}{\gamma}\int_0^\infty |u'(\tau)|^2 d\tau. \tag{2.11}$$

On the other hand, from the energy equation

$$\frac{1}{2}|u'|^2 + \Phi(u) + \gamma\int_0^t |u'(\tau)|^2 d\tau = \frac{1}{2}|v_0|^2 + \Phi(u_0),$$

it follows that

$$\int_0^\infty |u'(\tau)|^2 d\tau \le \frac{1}{\gamma}\Big[\frac{1}{2}|v_0|^2 + \Phi(u_0) - \inf\Phi\Big].$$

Replacing the last estimate in (2.11), it easy to show that (2.10) holds.

For (i), it suffices to take $x = \operatorname{proj}_S(u_0 + \frac{1}{\gamma}v_0)$ in (2.10).

For (ii), let $e := \widehat{u} - \operatorname{proj}_S(u_0 + \frac{1}{\gamma}v_0)$. If $e \neq 0$, then set

$$x_r := \operatorname{proj}_S\Big(u_0 + \frac{1}{\gamma}v_0\Big) - rd\Big(u_0 + \frac{1}{\gamma}v_0, S\Big)\frac{e}{|e|},$$

which belongs to S. An easy computation shows that

$$\Big|u_0 + \frac{1}{\gamma}v_0 - x_r\Big| - |\widehat{u} - x_r| = \Big(\sqrt{1+r^2} - r\Big)d\Big(u_0 + \frac{1}{\gamma}v_0 - x, S\Big) - |e|,$$

which together with (2.10) yields

$$|e| \le \Big(\sqrt{1+r^2} - r\Big)d\Big(u_0 + \frac{1}{\gamma}v_0 - x, S\Big) + \frac{1}{\gamma}\delta(u_0).$$

Letting $r \to \infty$ we get the result.

Finally, suppose that $\Phi(x) = \frac{1}{2}\langle Ax, x\rangle$ where $A : H \to H$ is a positive and self-adjoint bounded linear operator. Then $S = \{x \in H \mid Ax = 0\}$ the null space of A. Let $z \in S$; for all $t \ge 0$ we have that

$$\begin{aligned}\langle u'(t) - v_0, z\rangle + \gamma\langle u(t) - u_0, z\rangle &= \int_0^t \langle u''(\tau) + \gamma u'(\tau), z\rangle d\tau \\ &= \int_0^t \langle -Au(\tau), z\rangle d\tau \\ &= \int_0^t -\langle u(\tau), Az\rangle d\tau = 0.\end{aligned}$$

Since $u'(t) \to 0$ and $u(t) \to \widehat{u} \in S$ strongly (Φ is even) as $t \to \infty$, we can deduce that

$$\Big\langle \widehat{u} - \Big(u_0 + \frac{1}{\gamma}v_0\Big), z\Big\rangle = 0$$

for all $z \in S$, which completes the proof. ∎

2.3. Linear system: Heuristic comparison. Before proceeding further, it is interesting from the optimization viewpoint to compare the behavior of the trajectories defined by

$$(E_\gamma) \qquad u'' + \gamma u' + \nabla\Phi(u) = 0,$$

with the steepest descent equation

$$(\mathrm{SD}) \qquad u' + \nabla\Phi(u) = 0,$$

and with the continuous Newton's method

$$(\mathrm{N}) \qquad u' + \nabla^2\Phi(u)^{-1}\nabla\Phi(u) = 0.$$

For simplicity, in this section we restrict ourselves to the associated linearized systems in a finite dimensional space. We shall consider $H = \mathbb{R}^N$ and assume that $\Phi \in C^2(\mathbb{R}^N;\mathbb{R})$. Related to (SD), we have the linearized system around some $x_0 \in \mathbb{R}^N$, which is defined by

$$(\mathrm{LSD}) \qquad x' + \nabla^2\Phi(x_0)(x - x_0) + \nabla\Phi(x_0) = 0.$$

We assume that the Hessian matrix $\nabla^2\Phi(x_0)$ is positive definite. An explicit computation shows that $x(t) \to \widehat{x} := x_0 - \nabla^2\Phi(x_0)^{-1}\nabla\Phi(x_0)$ as $t \to \infty$. In fact, the solutions of (LSD) are of the form $y(t) = \widehat{x} + \eta(t)$, where η solves the homogeneous equation $\eta' + \nabla^2\Phi(x_0)\eta = 0$. Take a matrix P such that $P^{-1}\nabla^2\Phi(x_0)P = \mathrm{diag}(\lambda_1, \ldots, \lambda_N)$, where $\lambda_i > 0$, and set $P\xi = \eta$. We obtain the system $\xi_i' + \lambda_i\xi_i = 0$, whose solutions are $\xi_i(t) = C_i e^{-\lambda_i t}$. Generally speaking, if there is a $\lambda_i << 1$, we will have a relative slow convergence towards the solution; on the other hand, when dealing with large λ_i's the numerical integration by an approximate method will present stability problems. Thus we see that the numerical performance of (SD) is strongly determined by the *local geometry* of the function Φ.

We turn now to the linearized version of (N), given by

$$(\mathrm{LN}) \qquad y' + y - \widehat{x} = 0.$$

The solutions are of the form $y(t) = \widehat{x} + e^{-t}y(0)$, which are much better than the previous ones. The major properties are (1) the straight-line geometry of the trajectories; (2) that the rate of convergence is independent of the quadratic function to be minimized. Certainly, this is just a local approximation of the original function and the global behavior of the trajectory may be complicated. Nevertheless, this outstanding *normalization* property of Newton's system makes it effective in practice, due to the fact that the associated trajectories are easy to follow by a discretization method. Of course, an important disadvantage of (N) is the computation of the inverse of the Hessian matrix, which may be involved for a numerical algorithm.

Finally, we consider

$$(LE_\gamma) \qquad z'' + \gamma z' + \nabla^2\Phi(x_0)(z - x_0) + \nabla\Phi(x_0) = 0.$$

For this equation we have $z(t) = \widehat{x} + \epsilon(t)$, where ϵ solves the homogeneous problem $\epsilon'' + \gamma\epsilon' + \nabla^2\Phi(x_0)\epsilon = 0$. Setting $P\delta = \epsilon$ with P as above, then δ_i satisfies $\delta_i'' + \gamma\delta_i' + \lambda_i\delta_i = 0$. It is a simple matter to show that $|\delta_i(t)| \le C_i e^{-\mu_i(\gamma)t}$ with $\mu_i :]0,\infty[\to]0,\infty[$ continuous and C_i a constant independent of γ. In fact, $\mu_i(\gamma) = \frac{\gamma}{2}$ if $\gamma \in]0, 2\sqrt{\lambda_i}]$ and

$\mu_i(\cdot)$ is nonincreasing on $]2\sqrt{\lambda_i},\infty[$. Moreover, if $\gamma \geq 2\sqrt{\lambda_i}$, then the corresponding $\delta_i(t)$ does not present oscillations. Thus the choice $\gamma = 2\sqrt{\lambda_i}$ gives $\mu_i = \sqrt{\lambda_i}$, the greatest rate that can be obtained. But we can get any value in the interval $]0,\sqrt{\lambda_i}]$; for instance, when $\lambda_i > 1$ we obtain $\mu_i = 1$ either with $\gamma = 2$ or $\gamma = \lambda_i + 1$. The last choice has the advantage that the associated trajectory is not oscillatory, which is interesting by numerical reasons. Note that we should take a different parameter γ according to the corresponding eigenvalue λ_i.

Therefore, the presence of the damping parameter γ gives us a control on the behavior of the solutions of (E_γ) and, in particular, on some qualitative properties of the associated trajectories. For a general Φ we must take into account that (a) a careful selection of the damping parameter γ should depend on the local geometry of the function Φ, leading to a nonautonomous damping; (b) this selection could give a different value of γ for some particular directions, leading to an anisotropic damping. No attempt has been made here to develop a theory in order to guide these choices.

2.4. Linear and anisotropic damping. In the preceding section we have seen that it may be of interest to consider an anisotropic damping. With the aim of contributing to this issue, in this section we establish the asymptotic convergence for the solutions of the following system:

$$(E_\Gamma; u_0, v_0) \qquad \begin{cases} u'' + \Gamma u' + \nabla\Phi(u) = 0, \\ u(0) = u_0,\ u'(0) = v_0, \end{cases}$$

where $\Gamma : H \to H$ is a bounded self-adjoint linear operator, which we assume to be elliptic:

$$(h_\Gamma) \qquad \text{there exists } \gamma > 0 \text{ such that for any } x \in H,\ \langle \Gamma x, x\rangle \geq \gamma |x|^2.$$

THEOREM 2.6. *Suppose (h_Φ) and (h_Γ) hold. If $u \in C^2([0,\infty[;H)$ is a solution of $(E_\Gamma; u_0, v_0)$, then it satisfies $u' \in L^2([0,\infty[;H)$, $u'(t) \to 0$ as $t \to \infty$, and*

$$\lim_{t\to\infty} \Phi(u(t)) = \inf \Phi. \tag{2.12}$$

Furthermore, if $\operatorname{Argmin}\Phi \neq \emptyset$, *then there exists* $\widehat{u} \in \operatorname{Argmin}\Phi$ *such that* $u(t) \rightharpoonup \widehat{u}$ *weakly in* H *as* $t \to \infty$.

Proof. We only need to adapt the proof of Theorem 2.1. First, note that the properties of existence, uniqueness, and infinite extendibility to the right of the solution follow by similar arguments. Likewise, the energy $E(t) := \frac{1}{2}|u'(t)|^2 + \Phi(u(t))$ satisfies $E' = -\langle \Gamma u', u'\rangle$, and we can deduce that $u' \in L^2$.

Next, define the operator $A : H \to H$ by $Ax := \Gamma x - \gamma x$, with $\gamma > 0$ given by (h_Γ). Fix $x \in H$ and set $\varphi(t) := \frac{1}{2}|u(t) - x|^2$ and $\rho(t) := \frac{1}{2}\langle A(u(t) - x), u(t) - x\rangle$, in such a way that

$$\varphi'' + \gamma\varphi' + \rho' = \langle \nabla\Phi(u), x - u\rangle + |u'|^2. \tag{2.13}$$

As in the proof of Theorem 2.1, (2.13) gives

$$\varphi'(t) + \int_0^t e^{-\gamma(t-\tau)}\rho'(\tau)d\tau \leq e^{-\gamma t}\varphi'(0) + \frac{1}{\gamma}(1 - e^{-\gamma t})[\Phi(x) - E(t)] + r(t), \tag{2.14}$$

with

$$r(t) := \frac{3}{2}\int_0^t e^{-\gamma(t-\tau)}|u'(\tau)|^2 d\tau,$$

the only difference being the term $\int_0^t e^{-\gamma(t-\tau)}\rho'(\tau)d\tau$. An integration by parts yields

$$\int_0^t e^{-\gamma(t-\tau)}\rho'(\tau)d\tau = \rho(t) - e^{-\gamma t}\rho(0) - \gamma\int_0^t e^{-\gamma(t-\tau)}\rho(\tau)d\tau.$$

Setting $f(t) := \int_0^t e^{-\gamma(t-\tau)}\rho(\tau)d\tau$, we have

$$\int_0^t e^{-\gamma(t-\tau)}\rho'(\tau)d\tau = f'(t) - e^{-\gamma t}\rho(0).$$

Thus, we can rewrite (2.14) as

$$\varphi'(t) + f'(t) \le e^{-\gamma t}(\varphi'(0) + \rho(0)) + \frac{1}{\gamma}(1 - e^{-\gamma t})[\Phi(x) - E(t)] + r(t).$$

We leave it to the reader to verify that the minimizing property (2.12) can now be established as in Theorem 2.1. The proof of $u'(t) \to 0$ as $t \to \infty$ is analogous.

When $\operatorname{Argmin}\Phi \neq \emptyset$, we fix $z \in \operatorname{Argmin}\Phi$ and consider the corresponding functions φ and ρ as above (with x replaced by z). Using the optimality of z, it follows that

$$\varphi'(t) + f'(t) \le e^{-\gamma t}(\varphi'(0) + \rho(0)) + \int_0^t e^{-\gamma(t-\tau)}|u'(\tau)|^2 d\tau, \tag{2.15}$$

with f associated with ρ as above. Integrating this inequality we conclude that $\varphi(t)$ stays bounded as $t \to \infty$, but we cannot deduce its convergence. Then, we rewrite (2.15) in the form

$$\varphi'(t) + \int_0^t e^{-\gamma(t-\tau)}\rho'(\tau)d\tau \le e^{-\gamma t}\varphi'(0) + \int_0^t e^{-\gamma(t-\tau)}|u'(\tau)|^2 d\tau,$$

and we conclude that $[\varphi'(t) + \int_0^t e^{-\gamma(t-\tau)}\rho'(\tau)d\tau]_+ \in L^1([0,\infty[;\mathbb{R})$. We note that

$$\varphi'(t) + \int_0^t e^{-\gamma(t-\tau)}\rho'(\tau)d\tau = \mu'(t) + \xi'(t),$$

where

$$\mu(t) := \frac{1}{2\gamma}\langle \Gamma(u(t) - z), u(t) - z\rangle,$$

and

$$\xi(t) := -\frac{1}{\gamma}\int_0^t e^{-\gamma(t-\tau)}\rho'(\tau)d\tau.$$

By virtue of Lemma 2.2, if we show that $\xi(t)$ is bounded from below, then $\mu(t) + \xi(t)$ converges as $t \to \infty$. Since $\rho'(t) = \langle Au'(t), u(t) - z\rangle$, there exists a constant $M > 0$ independent of t such that $|\rho'(t)| \le M|u'(t)|\sqrt{\varphi(t)}$ for any $t > 0$. We conclude that $\rho'(t) \to 0$ as $t \to \infty$. From this fact it follows easily that $\xi(t) \to 0$ as $t \to \infty$. Therefore, $\mu(t) + \xi(t)$ converges as $t \to \infty$, hence $\mu(t)$ converges as well.

The proof is completed by applying the Opial lemma to the trajectory $\{u(t) : t \to \infty\}$, where the Hilbert space H is endowed with the inner product $\langle\langle \cdot, \cdot\rangle\rangle : H \times H \to \mathbb{R}$ defined by $\langle\langle x, y\rangle\rangle := \frac{1}{\gamma}\langle \Gamma x, y\rangle$ and its associated norm. □

Remark 1. In Theorems 2.1, 2.4, and 2.6 we do not require any coerciveness assumption on Φ. When $\operatorname{Argmin}\Phi \neq \emptyset$, the dissipativeness in the dynamics suffices for the convergence of the solutions. If the infimum value is not realized, the trajectory may be unbounded as in the one-dimensional equation $u'' + \gamma u' + e^u = 0$, whose solutions $u \in C^2([0,\infty[;\mathbb{R})$ are so that $u(t) \to -\infty$ and $u'(t) \to 0$ as $t \to \infty$. In any case, our results assert that the dynamical system defined by (E_γ) (or more generally by (E_Γ)) is dissipative in the sense that every trajectory evolves towards a minimum of the energy. Certainly, there is a strong connection with the concept of point dissipativeness or ultimately boundedness in the theory of dynamical systems, where the Lyapunov function associated with the semigroup is usually supposed to be coercive (cf. [9, gradient systems]).

Remark 2. To ensure local existence and uniqueness of a classical solution for the differential equation, it suffices to require a local Lipschitz property on $\nabla\Phi$. Actually, in some situations this hypothesis is not necessary and the existence may be established by other arguments. For instance, that is the case of the Hille–Yosida theorem for evolution equations governed by monotone operators and the theory of linear and nonlinear semigroups for partial differential equations. Note that such a Lipschitz condition on the gradient is not used in the asymptotic analysis of the trajectories. Therefore, the previous asymptotic results remain valid for other classes of infinite dimensional dissipative systems provided the existence of a global solution. It is not our purpose to develop this point here for the continuous system because it exceeds the scope of this paper. However, in the next section we consider an implicit discretization of the continuous system. As we will see, the existence of the discrete trajectory is ensured by variational arguments. This will allow us to apply the discrete scheme to nonsmooth convex functions and to adapt the asymptotic analysis to this case.

3. Discrete approximation method. Once we have established the existence of a solution of an initial value problem, we are interested in its numerical values. We must accept that most differential equations cannot be solved explicitly; we are thus led to work with approximate methods. An important class of these methods is based on the approximation of the exact solution over a discrete set $\{t_n\}$: associated with each point t_n we compute a value u_n, which approximates $u(t_n)$ the exact solution at t_n. Generally speaking, these procedures have the disadvantage that a large number of calculations has to be done in order to keep the discretization error $e_n := u_n - u(t_n)$ sufficiently small. In addition to this, the estimates for the errors strongly depend on the length of the discretization range for the t variable. It turns out that these methods are not well adapted to the approximation of the exact solution on an unbounded domain.

Nevertheless, there is an important point to note here. If our objective is the asymptotic behavior of the solutions as t goes to ∞, then the accurate approximation of the whole trajectory becomes immaterial. We present a discrete method whose feature is that no attempt is made to approximate the exact solution over a set of points but that the discrete values are sought only to preserve the asymptotic behavior of the solutions.

3.1. Implicit iterative scheme. Dealing with the discretization of a first order differential equation $y' = F(y)$, it is classical to consider the implicit iterative scheme

$$\frac{y_{k+1} - y_k}{h} = F(y_{k+1}), \tag{3.1}$$

where $h > 0$ is a parameter called step size. In the case of equation (E_γ), or more precisely its first order equivalent system, (3.1) corresponds to recursively solve

$$\frac{u_{k+1} - 2u_k + u_{k-1}}{h^2} + \gamma \frac{u_{k+1} - u_k}{h} + \nabla\Phi(u_{k+1}) = 0. \tag{3.2}$$

Since Φ is convex, (3.2) is equivalent to the following variational problem:

$$u_{k+1} = \operatorname{argmin}\left\{\Phi(x) + \frac{1+\gamma h}{2h^2}\,|x - z_k|^2 : x \in H\right\},$$

where $z_k = u_k + \frac{1}{1+\gamma h}(u_k - u_{k-1})$. This motivates the introduction of the more general iterative procedure

$$u_{k+1} = \operatorname{argmin}\left\{\Phi(x) + \frac{1}{2\lambda}\,|x - z_k|^2 : x \in H\right\},$$

where $z_k = u_k + \alpha(u_k - u_{k-1})$, λ and α are positive. Note that when $\alpha = 0$, we recover the standard (Prox) iteration. If $\alpha > 0$, the starting point for the next iteration is computed as a development in terms of the *velocity* of the already generated sequence. Therefore, this iterative scheme defines a second order dynamics, while (Prox) is actually of a first order nature.

We have been working under the assumption that Φ is differentiable. However, for the above iterative variational method this regularity is no longer necessary. Thus, in that which follows $f : H \to \mathbb{R} \cup \{\infty\}$ denotes a closed proper convex function (see [17]), which eventually realizes the value ∞, and we consider

$$u_{k+1} = \operatorname{argmin}\left\{f(x) + \frac{1}{2\lambda}\,|x - z_k|^2 : x \in H\right\}, \tag{3.3}$$

where $z_k = u_k + \alpha(u_k - u_{k-1})$. In terms of the stationary condition, (3.3) is equivalent to

$$\frac{1}{\lambda}(u_{k+1} - (1+\alpha)u_k + \alpha u_{k-1}) + \partial f(u_{k+1}) \ni 0,$$

where ∂f is the standard convex subdifferential [17].

3.2. Convergence for the variational algorithm. By numerical reasons, it is natural to consider the following approximate iterative scheme:

$$\frac{1}{\lambda_k}(u_{k+1} - (1+\alpha_k)u_k + \alpha_k u_{k-1}) + \partial_{\epsilon_k} f(u_{k+1}) \ni 0, \tag{3.4}$$

where α_k is nonnegative, λ_k is positive, and $\partial_\epsilon f$ is the ϵ-subdifferential. Note that a sequence $\{u_k\} \subset H$ satisfying (3.4) always exists. Indeed, given $u_{k-1}, u_k \in H$, we can take u_{k+1} as the unique solution of the strongly convex problem $\min\{f(x) + \frac{1}{2\lambda_k}|x - z_k|^2 : x \in H\}$ with z_k as above.

THEOREM 3.1. *Assume that f is closed proper convex and bounded from below. Let $\{u_k\} \subset H$ be a sequence generated by* (3.4)*, where*

(i) $0 \le \alpha_k \le 1$ *and* $\{\lambda_k\}$ *is bounded from below by a positive constant,*

(ii) *the sequence* $\{\alpha_k/\lambda_k\}$ *is nonincreasing and* $\sum \lambda_k \epsilon_k < \infty$.

Then

$$\lim_{k\to\infty} \frac{1}{\lambda_k}(u_{k+1} - (1+\alpha_k)u_k + \alpha_k u_{k-1}) = 0, \tag{3.5}$$

and in particular $\lim_{k\to\infty} d(0,\partial_{\epsilon_k} f(u_{k+1})) = 0$.

When $\operatorname{Argmin} f \neq \emptyset$, *assume in addition that*

(iii) *there exists* $\bar{\alpha} \in]0,1[$ *such that* $0 \leq \alpha_k \leq \bar{\alpha}$, *and* $\{\lambda_k\}$ *is bounded from above if there is at least one* $\alpha_k > 0$.

Then, there exists $\widehat{u} \in \operatorname{Argmin} f$ *such that* $u_k \rightharpoonup \widehat{u}$ *weakly as* $k \to \infty$.

Proof. The proof consists of adapting the analysis done for the differential equation (E_γ). We begin by defining the discrete energy by

$$E_{k+1} = \frac{\alpha_k}{2\lambda_k}|u_{k+1} - u_k|^2 + f(u_{k+1}),$$

and we study the successive difference $E_{k+1} - E_k$. Since $\alpha_k/\lambda_k \leq \alpha_{k-1}/\lambda_{k-1}$,

$$E_{k+1} - E_k \leq \frac{\alpha_k}{2\lambda_k}\left(|u_{k+1} - u_k|^2 - |u_k - u_{k-1}|^2\right) + f(u_{k+1}) - f(u_k).$$

By definition of $\partial_{\epsilon_k} f$, (3.4) yields

$$f(u_{k+1}) - f(u_k) \leq -\frac{1}{\lambda_k}\langle u_{k+1} - (1+\alpha_k)u_k + \alpha_k u_{k-1}, u_{k+1} - u_k\rangle + \epsilon_k.$$

As we can write

$$\langle u_{k+1} - (1+\alpha_k)u_k + \alpha_k u_{k-1}, u_{k+1} - u_k\rangle = |u_{k+1} - u_k|^2 - \alpha_k\langle u_k - u_{k-1}, u_{k+1} - u_k\rangle,$$

we have

$$E_{k+1} - E_k \leq -\frac{\alpha_k}{2\lambda_k}|u_{k+1} - 2u_k + u_{k-1}|^2 - \frac{1-\alpha_k}{\lambda_k}|u_{k+1} - u_k|^2 + \epsilon_k,$$

and consequently

$$\sum_{k=1}^{N}\left[\frac{\alpha_k}{2\lambda_k}|u_{k+1} - 2u_k + u_{k-1}|^2 + \frac{1-\alpha_k}{\lambda_k}|u_{k+1} - u_k|^2\right] \leq E_1 - E_{N+1} + \sum_{k=1}^{N}\epsilon_k.$$

Noting that

$$E_1 - E_{N+1} + \sum_{k=1}^{N}\epsilon_k \leq E_1 - \inf f + \sum \epsilon_k < \infty,$$

and because $0 \leq \alpha_k \leq 1$, we deduce that

$$\sum \frac{\alpha_k}{2\lambda_k}|u_{k+1} - 2u_k + u_{k-1}|^2 < \infty$$

and

$$\sum \frac{1-\alpha_k}{\lambda_k}|u_{k+1} - u_k|^2 < \infty. \tag{3.6}$$

As $0 \leq \alpha_k \leq 1$ and λ_k is bounded from below by a positive constant, we have

$$\lim_{k\to\infty}\frac{\alpha_k}{\lambda_k}|u_{k+1} - 2u_k + u_{k-1}| = \lim_{k\to\infty}\frac{(1-\alpha_k)}{\lambda_k}|u_{k+1} - u_k| = 0.$$

Writing

$$u_{k+1} - (1+\alpha_k)u_k + \alpha_k u_{k-1} = \alpha_k(u_{k+1} - 2u_k + u_{k-1}) + (1-\alpha_k)(u_{k+1} - u_k),$$

we conclude that (3.5) holds.

Suppose now that $\operatorname{Argmin} f \neq \emptyset$. We apply the Opial lemma to prove the weak convergence of $\{u_k\}$. On account of (3.5), it is sufficient to show that for any $z \in \operatorname{Argmin} f$, the sequence of positive numbers $\{|u_k - z|\}$ is convergent. Fix $z \in \operatorname{Argmin} f$; since u_{k+1} satisfies (3.4), we have

$$f(u_{k+1}) - \frac{1}{\lambda_k}\langle u_{k+1} - (1+\alpha_k)u_k + \alpha_k u_{k-1}, z - u_{k+1}\rangle \leq f(z) + \epsilon_k,$$

and by the optimality of z

$$\langle u_{k+1} - u_k, u_{k+1} - z\rangle - \alpha_k\langle u_k - u_{k-1}, u_{k+1} - z\rangle \leq \lambda_k\epsilon_k. \tag{3.7}$$

Set $\varphi_k := \frac{1}{2}|u_k - z|^2$. It is direct to check that for any $k \in \mathbb{N}$

$$\varphi_{k+1} = \varphi_k + \langle u_{k+1} - u_k, u_{k+1} - z\rangle - \frac{1}{2}|u_{k+1} - u_k|^2.$$

Since $\langle u_k - u_{k-1}, u_{k+1} - z\rangle = \langle u_k - u_{k-1}, u_k - z\rangle + \langle u_k - u_{k-1}, u_{k+1} - u_k\rangle$, (3.7) shows that

$$\varphi_{k+1} - \varphi_k - \alpha_k\left(\varphi_k - \varphi_{k-1} + \frac{1}{2}|u_k - u_{k-1}|^2 + \langle u_k - u_{k-1}, u_{k+1} - u_k\rangle\right) \leq \lambda_k\epsilon_k,$$

and therefore

$$\varphi_{k+1} - (1+\alpha_k)\varphi_k + \alpha_k\varphi_{k-1} \leq \delta_k,$$

where $\delta_k = \alpha_k|u_k - u_{k-1}|^2 + \frac{\alpha_k}{2}|u_{k+1} - u_k|^2 + \lambda_k\epsilon_k$. Using (iii) and (3.6) it follows that $\sum |u_{k+1} - u_k|^2 < \infty$, thus $\sum \delta_k < \infty$. Set $\theta_k := \varphi_k - \varphi_{k-1}$; the above inequality implies

$$[\theta_{k+1}]_+ \leq \bar{\alpha}[\theta_k]_+ + \delta_k.$$

Thus

$$[\theta_{k+1}]_+ \leq \bar{\alpha}^k[\theta_1]_+ + \sum_{j=0}^{k-1}\bar{\alpha}^j\delta_{k-j},$$

which yields

$$\sum_{k=0}^{\infty}[\theta_{k+1}]_+ \leq \frac{1}{1-\bar{\alpha}}\left([\theta_1]_+ + \sum_{k=1}^{\infty}\delta_k\right) < \infty.$$

Set $w_k := \varphi_k - \sum_{j=1}^{k}[\theta_j]_+$. Since $\varphi_k \geq 0$ and $\sum[\theta_j]_+ < \infty$, w_k is bounded from below. As $\{w_k\}$ is nonincreasing we have that it converges. Hence $\{\varphi_k\}$ converges, which completes the proof of the theorem. □

For simplicity, we have considered in this section the isotropic damping system. However, a similar analysis can be done for the anisotropic damping associated with

an elliptic self-adjoint linear operator $\Gamma : H \to H$. The variational problem associated with the implicit discretization is

$$u_{k+1} = \operatorname{argmin}\left\{\Phi(x) + \frac{1}{2h^2}\,|x - z_k|^2_{(I+h\Gamma)} : x \in H\right\},$$

where $z_k = u_k + (I + h\Gamma)^{-1}(u_k - u_{k-1})$ and for any $y \in H$,

$$|y|_{(I+h\Gamma)} := \sqrt{\langle (I + h\Gamma)y, y\rangle}.$$

For a function $f : H \to \mathbb{R} \cup \{\infty\}$ closed proper and convex, the latter motivates the scheme

$$R(u_{k+1} - (I + S)u_k + Su_{k-1}) + \partial f(u_{k+1}) \ni 0,$$

where $R : H \to H$ is a linear positive definite operator and $S : H \to H$ is linear and positive semidefinite. If we assume both R and $I - S$ are elliptic, it is possible to obtain a convergence result like the previous one. It suffices to adapt the main arguments. Since the basic ideas are contained in the proof of Theorems 2.6 and 3.1, we shall go no further in this matter.

4. Some open problems. In the case of multiple optimal solutions, our convergence results do not provide additional information on the point attained in the limit. A possible approach to overcome this disadvantage may be to couple the dissipative system with approximation techniques such as regularization, interior-barrier or globally defined penalizations, and viscosity methods. In the continuous case, this alternative has been considered with success for the steepest descent equation in [2] and for Newton's method in [1], giving a characterization for the limit point under suitable assumptions on the approximate scheme. On account of these results, one may conjecture that this can be done for the equations considered in the present work.

On the other hand, we have seen that the behavior of the trajectories depends on a relation between the damping and the local geometry of the function we wish to minimize. This remark leads us to the obvious problem of the choice of the damping parameter, made in order to have a better control on the trajectory. This is also a problem in the discrete algorithm. Usually we have an incomplete knowledge of the objective function, which makes the question more difficult. We think that a first step in this direction may be the study of more general damped equations, with nonlinear and/or nonautonomous damping.

Acknowledgments. I wish to express my gratitude to the Laboratoire d'Analyse Convexe de l'Université Montpellier II for their hospitality and support, and especially to Professor Hedy Attouch. I gratefully acknowledge financial support through a French Foreign Scholarship grant from the French Ministry of Education and a Chilean National Scholarship grant from the CONICYT of Chile. I wish to thank a referee for helpful comments concerning Theorems 2.2 and 3.1.

REFERENCES

[1] F. Alvarez and J. M. Pérez, *A dynamical system associated with Newton's method for parametric approximations of convex minimization problems*, Appl. Math. Optim., 38 (1998), pp. 193–217.

[2] H. Attouch and R. Cominetti, *A dynamical approach to convex minimization coupling approximation with the steepest descent method*, J. Differential Equations, 128 (1996), pp. 519–540.

[3] H. Attouch, X. Goudou, and P. Redont, *A dynamical method for the global exploration of stationary points of a real-valued mapping: The heavy ball method*, Commun. Contemp. Math., to appear.
[4] D. Bayer and J. C. Lagarias, *The nonlinear geometry of linear programming* (*parts* I *and* II), Trans. Amer. Math. Soc., 314 (1989), pp. 527–581.
[5] H. Brezis, *Monotonicity methods in Hilbert spaces and some applications to nonlinear partial differential equations*, in Contributions to Nonlinear Analysis, E. H. Zarantonello, ed., Academic Press, New York, 1971, pp. 101–151.
[6] H. Brezis, *Opérateurs Maximaux Monotones*, North-Holland Math. Stud. 5, North-Holland, Amsterdam, 1973.
[7] R. E. Bruck, *Asymptotic convergence of nonlinear contraction semi-groups in Hilbert spaces*, J. Funct. Anal., 18 (1975), pp. 15–26.
[8] R. Cominetti, *Asymptotic convergence of the steepest descent method for the exponential penalty in linear programming*, J. Convex Anal., 2 (1995), pp. 145–152.
[9] J. K. Hale, *Asymptotic Behavior of Dissipative Systems*, Math. Surveys Monogr. 25, American Mathematical Society, Providence, RI, 1988.
[10] A. Haraux, *Systèmes dynamiques dissipatifs et applications*, Rech. Math. Appl. 17, Masson, Paris, 1991.
[11] A. Haraux and M. A. Jendoubi, *Convergence of solutions of second-order gradient-like systems with analytic nonlinearities*, J. Differential Equations, 144 (1998), pp. 313–320.
[12] B. Lemaire, *About the convergence of the proximal method*, in Advances in Optimization, Proceedings Lambrecht, 1991, Lecture Notes in Econ. and Math. Systems 382, Springer-Verlag, New York, 1992, pp. 39–51.
[13] B. Lemaire, *An asymptotical variational principle associated with the steepest descent method for a convex function*, J. Convex Anal., 3 1996, pp. 63–70.
[14] G. P. McCormick, *The projective SUMT method for convex programming*, Math. Oper. Res., 14 (1989), pp. 203–223.
[15] Z. Opial, *Weak convergence of the sequence of successive approximations for nonexpansive mappings*, Bull. Amer. Math. Soc., 73 (1967), pp. 591–597.
[16] B. T. Polyack, *Some methods of speeding up the convergence of iterative methods*, Zh. Vychisl. Mat. Mat. Fiz., 4 (1964), pp. 1–17.
[17] R. T. Rockafellar, *Convex Analysis*, Princeton University Press, Princeton, NJ, 1970.
[18] R. T. Rockafellar, *Monotone operators and the proximal point algorithm*, SIAM J. Control Optim., 14 (1976), pp. 877–898.

SIAM J. CONTROL OPTIM.
Vol. 38, No. 4, pp. 1120–1144

TIME-VARYING AND ADAPTIVE INTEGRAL CONTROL OF INFINITE-DIMENSIONAL REGULAR LINEAR SYSTEMS WITH INPUT NONLINEARITIES*

H. LOGEMANN† AND E. P. RYAN†

Abstract. Closing the loop around an exponentially stable, single-input, single-output, regular linear system—subject to a globally Lipschitz, nondecreasing actuator nonlinearity and compensated by an integral controller with time-dependent gain $k(t)$—is shown to ensure asymptotic tracking of a constant reference signal r, provided that (a) the steady-state gain of the linear part of the system is positive, (b) the reference value r is feasible in an entirely natural sense, and (c) the function $t \mapsto k(t)$ monotonically decreases to zero at a sufficiently slow rate. This result forms the basis of a simple adaptive control strategy that ensures asymptotic tracking under conditions (a) and (b).

Key words. adaptive control, infinite-dimensional regular systems, input nonlinearities, integral control, saturation, robust tracking

AMS subject classifications. 34D05, 93C25, 93D05, 93D09, 93D15, 93D21

PII. S0363012998339228

1. Introduction.

The paper has, as a precursor, the article [9] which contains an extension, to infinite-dimensional systems with input nonlinearities, of the well-known principle (see, for example, [5], [13], and [17]) that closing the loop around a stable, linear, finite-dimensional, continuous-time, single-input, single-output plant, with transfer function $\mathbf{G}(s)$ compensated by a pure integral controller $\mathbf{C}(s) = k/s$, will result in a stable closed-loop system that achieves asymptotic tracking of arbitrary constant reference signals, provided that $|k|$ is sufficiently small and $\mathbf{G}(0)k > 0$. In particular, in [9] it is shown that the above principle may remain valid if the plant to be controlled is a single-input, single-output, continuous-time, infinite-dimensional, regular (as defined in section 2 below) linear system subject to an input nonlinearity ϕ. More precisely, if ϕ is globally Lipschitz and nondecreasing, if $\mathbf{G}(0) > 0$, and if the constant reference signal r is feasible (in the sense that $[\mathbf{G}(0)]^{-1}r$ is in the closure of the image of ϕ), then there exists $k^* > 0$ such that, $\forall k \in (0, k^*)$, the output $y(t)$ of the closed-loop system (shown in Figure 1) converges to r as $t \to \infty$. Therefore, if a (regular) plant is known to be stable, if the input nonlinearity is of the above class, if $\mathbf{G}(0) \neq 0$, and if the sign of $\mathbf{G}(0)$ is known (in principle, the latter information can be obtained from plant step response data), then the problem of tracking feasible signals r by low-gain integral control reduces to that of tuning the gain parameter k. In a nonadaptive, linear, finite-dimensional context, one such controller design approach ("tuning regulator theory"[5]) has been successfully applied in process control (see, for example, [4] and [14]). Furthermore, the problem of tuning the integrator gain adaptively has been addressed recently in a number of papers: see, for example, [3] and [15], [16] for the finite-dimensional case (with input constraints treated in [15]), and [10], [11], [12] for the linear infinite-dimensional case.

The present paper addresses aspects of adaptive tuning of the integrator gain

*Received by the editors May 22, 1998; accepted for publication (in revised form) December 29, 1998; published electronically April 4, 2000. This work was supported by UK Engineering and Physical Sciences Research Council grant GR/L78086.
http://www.siam.org/journals/sicon/38-4/33922.html
†Department of Mathematical Sciences, University of Bath, Bath BA2 7AY, United Kingdom (hl@maths.bath.ac.uk, epr@maths.bath.ac.uk).

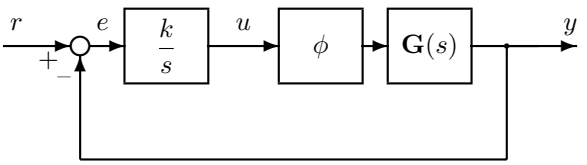


FIG. 1. *Low-gain control with input nonlinearity.*

for infinite-dimensional regular linear systems (with transfer function $\mathbf{G}$), subject to input nonlinearities ϕ of the same class as considered in its precursor [9]. In [9], the constant-gain case is treated; there, the existence of a value $k^* > 0$, with the property that asymptotic tracking of feasible reference signals r is ensured for every *fixed* gain $k \in (0, k^*)$, is established. Let k^{**} denote the supremum of all such k^*. In [9], it is shown that $k^{**} \geq \kappa^*/\lambda$, where $\lambda > 0$ is a Lipschitz constant for ϕ and κ^* denotes the supremum of all numbers $\kappa > 0$ such that

$$1 + \kappa \operatorname{Re} \frac{\mathbf{G}(s)}{s} \geq 0 \quad \forall\ s \text{ with } \operatorname{Re} s > 0.$$

For lower bounds and formulae for κ^* in terms of plant data, we refer to [8]. In general, k^{**} is a function of the plant data and so, in the presence of uncertainty, may fail to be computable. In such cases, it is natural to consider time-dependent gain strategies $t \mapsto k(t) > 0$ capable of attaining sufficiently small values. Theorem 3.8 has the following flavor: if $k(\cdot)$ monotonically decreases to zero *sufficiently slowly*, then asymptotic tracking of feasible reference signals is achieved. The practical utility of this result is limited insofar as the gain function is selected a priori: no use is made of the instantaneous output information $y(t)$ from the plant to update the gain. Utilizing the available output information, Theorem 3.13 establishes the efficacy of the simple adaptive gain strategy

$$k(t) = \frac{1}{l(t)}, \qquad \text{where} \quad \dot{l}(t) = |r - y(t)|,\ \ l(0) = l_0 > 0,$$

and shows that, if the reference signal r is such that $[\mathbf{G}(0)]^{-1} r \in \operatorname{im} \phi$ is not a critical value of ϕ, then the monotone function $t \mapsto k(t) > 0$ converges to a positive limit as $t \to \infty$.

2. Preliminaries on regular linear systems. We assemble some fundamental facts pertaining to regular linear systems and tailored to later requirements; the reader is referred to [20], [21], [22], [23], [24], and [9] for full details. This section is prefaced with the remark that the class of regular linear infinite-dimensional systems is rather general: it includes most distributed parameter systems and all time-delay systems (retarded and neutral) which are of interest in applications. Although there exist abstract examples of well-posed, infinite-dimensional systems that fail to be regular, the authors are of the opinion that any physically motivated, well-posed, linear, continuous-time, autonomous control system is regular.

First, some notation: for a Hilbert space H and any $\tau \geq 0$, $\mathbf{R}_\tau$ denotes the operator of right-shift by τ on $L^p_{\mathrm{loc}}(\mathbb{R}_+, H)$, where $\mathbb{R}_+ := [0, \infty)$; the truncation operator $\mathbf{P}_\tau : L^p_{\mathrm{loc}}(\mathbb{R}_+, H) \to L^p(\mathbb{R}_+, H)$ is given by $(\mathbf{P}_\tau u)(t) = u(t)$ if $t \in [0, \tau]$ and $(\mathbf{P}_\tau u)(t) = 0$ otherwise; for $\alpha \in \mathbb{R}$, we define the exponentially weighted L^p-space $L^p_\alpha(\mathbb{R}_+, H) := \{f \in L^p_{\mathrm{loc}}(\mathbb{R}_+, H) |\ f(\cdot) \exp(-\alpha\, \cdot) \in L^p(\mathbb{R}_+, H)\}$; $\mathcal{B}(H_1, H_2)$ denotes

the space of bounded linear operators from a Hilbert space H_1 to a Hilbert space H_2; for $\alpha \in \mathbb{R}$, $\mathbb{C}_\alpha := \{s \in \mathbb{C} \,|\, \operatorname{Re} s > \alpha\}$; the Laplace transform is denoted by $\mathcal{L}$.

Well-posed systems. The concept of a well-posed linear system was introduced by Weiss [24]. An equivalent definition can be found in [19].

DEFINITION 2.1. *Let U, X, and Y be real Hilbert spaces. A well-posed linear system with state space X, input space U, and output space Y is a quadruple $\boldsymbol{\Sigma} = (\mathbf{T}, \boldsymbol{\Phi}, \boldsymbol{\Psi}, \mathbf{F})$, where $\mathbf{T} = (\mathbf{T}_t)_{t\geq 0}$ is a C_0-semigroup of bounded linear operators on X; $\boldsymbol{\Phi} = (\boldsymbol{\Phi}_t)_{t\geq 0}$ is a family of bounded linear operators from $L^2(\mathbb{R}_+, U)$ to X such that, $\forall\, \tau, t \geq 0$,*

$$\boldsymbol{\Phi}_{\tau+t}(\mathbf{P}_\tau u + \mathbf{R}_\tau v) = \mathbf{T}_t \boldsymbol{\Phi}_\tau u + \boldsymbol{\Phi}_t v \quad \forall\, u, v \in L^2(\mathbb{R}_+, U)\,;$$

$\boldsymbol{\Psi} = (\boldsymbol{\Psi}_t)_{t\geq 0}$ is a family of bounded linear operators from X to $L^2(\mathbb{R}_+, Y)$ such that $\boldsymbol{\Psi}_0 = 0$ and, $\forall\, \tau, t \geq 0$,

$$\boldsymbol{\Psi}_{\tau+t} x_0 = \mathbf{P}_\tau \boldsymbol{\Psi}_\tau x_0 + \mathbf{R}_\tau \boldsymbol{\Psi}_t \mathbf{T}_\tau x_0 \quad \forall\, x_0 \in X;$$

and $\mathbf{F} = (\mathbf{F}_t)_{t\geq 0}$ is a family of bounded linear operators from $L^2(\mathbb{R}_+, U)$ to $L^2(\mathbb{R}_+, Y)$ such that $\mathbf{F}_0 = 0$ and, $\forall\, \tau, t \geq 0$,

$$\mathbf{F}_{\tau+t}(\mathbf{P}_\tau u + \mathbf{R}_\tau v) = \mathbf{P}_\tau \mathbf{F}_\tau u + \mathbf{R}_\tau(\boldsymbol{\Psi}_t \boldsymbol{\Phi}_\tau u + \mathbf{F}_t v) \quad \forall\, u, v \in L^2(\mathbb{R}_+, U).$$

For an input $u \in L^2_{\mathrm{loc}}(\mathbb{R}_+, U)$ and initial state $x_0 \in X$, the associated state function $x \in C(\mathbb{R}_+, X)$ and output function $y \in L^2_{\mathrm{loc}}(\mathbb{R}_+, Y)$ of $\boldsymbol{\Sigma}$ are given by

$$x(t) = \mathbf{T}_t x_0 + \boldsymbol{\Phi}_t \mathbf{P}_t u, \tag{1a}$$

$$\mathbf{P}_t y = \boldsymbol{\Psi}_t x_0 + \mathbf{F}_t \mathbf{P}_t u. \tag{1b}$$

$\boldsymbol{\Sigma}$ is said to be *exponentially stable* if the semigroup $\mathbf{T}$ is exponentially stable:

$$\omega(\mathbf{T}) := \lim_{t\to\infty} \frac{1}{t} \ln \|\mathbf{T}_t\| < 0.$$

$\boldsymbol{\Psi}_\infty$ and $\mathbf{F}_\infty$ will denote the unique operators $X \to L^2_{\mathrm{loc}}(\mathbb{R}_+, Y)$ and $L^2_{\mathrm{loc}}(\mathbb{R}_+, U) \to L^2_{\mathrm{loc}}(\mathbb{R}_+, Y)$, respectively, satisfying

$$\boldsymbol{\Psi}_\tau = \mathbf{P}_\tau \boldsymbol{\Psi}_\infty, \quad \mathbf{F}_\tau = \mathbf{P}_\tau \mathbf{F}_\infty \quad \forall\, \tau \geq 0. \tag{2}$$

If $\boldsymbol{\Sigma}$ is exponentially stable, then the operators $\boldsymbol{\Phi}_t$ and $\boldsymbol{\Psi}_t$ are uniformly bounded; $\boldsymbol{\Psi}_\infty$ is a bounded operator from X into $L^2(\mathbb{R}_+, Y)$, and $\mathbf{F}_\infty$ maps $L^2(\mathbb{R}_+, U)$ boundedly into $L^2(\mathbb{R}_+, Y)$. Since $\mathbf{P}_\tau \mathbf{F}_\infty = \mathbf{P}_\tau \mathbf{F}_\infty \mathbf{P}_\tau\, \forall\, \tau \geq 0$, $\mathbf{F}_\infty$ is a *causal* operator.

Regularity. Weiss [20] has established that, if $\alpha > \omega(\mathbf{T})$ and $u \in L^2_\alpha(\mathbb{R}_+, U)$, then $\mathbf{F}_\infty u \in L^2_\alpha(\mathbb{R}_+, Y)$ and there exists a unique holomorphic $\mathbf{G} : \mathbb{C}_{\omega(\mathbf{T})} \to \mathcal{B}(U, Y)$ such that

$$\mathbf{G}(s)(\mathcal{L}u)(s) = [\mathcal{L}(\mathbf{F}_\infty u)](s) \quad \forall\, s \in \mathbb{C}_\alpha,$$

where $\mathcal{L}$ denotes Laplace transform. In particular, $\mathbf{G}$ is bounded on $\mathbb{C}_\alpha\ \forall\ \alpha > \omega(\mathbf{T})$. The function $\mathbf{G}$ is called the *transfer function* of $\boldsymbol{\Sigma}$.

$\boldsymbol{\Sigma}$ and its transfer function $\mathbf{G}$ are said to be *regular* if there exists a linear operator D such that

$$\lim_{s\to\infty,\, s\in\mathbb{R}} \mathbf{G}(s)u = Du \quad \forall\, u \in U,$$

in which case, by the principle of uniform boundedness, it follows that $D \in \mathcal{B}(U, Y)$. The operator D is called the *feedthrough operator* of $\mathbf{\Sigma}$.

Generating operators. The generator of $\mathbf{T}$ is denoted by A with domain $\mathrm{dom}(A)$. Let X_1 be the space $\mathrm{dom}(A)$ endowed with the graph norm. The norm on X is denoted by $\|\cdot\|$, whilst $\|\cdot\|_1$ denotes the graph norm. Let X_{-1} be the completion of X with respect to the norm $\|x\|_{-1} = \|(\lambda I - A)^{-1}x\|$, where $\lambda \in \varrho(A)$ is any fixed element of the resolvent set $\varrho(A)$ of A. Then $X_1 \subset X \subset X_{-1}$ and the canonical injections are bounded and dense. The semigroup $\mathbf{T}$ can be restricted to a C_0-semigroup on X_1 and extended to a C_0-semigroup on X_{-1}. The exponential growth constant is the same on all three spaces. The generator on X_{-1} is an extension of A to X (which is bounded as an operator from X to X_{-1}). We shall use the same symbol $\mathbf{T}$ (respectively, A) for the original semigroup (respectively, its generator) and the associated restrictions and extensions. With this convention, we may write $A \in \mathcal{B}(X, X_{-1})$. Considered as a generator on X_{-1}, the domain of A is X.

By a representation theorem due to Salamon [19] (see also Weiss [22, 23]), there exist unique operators $B \in \mathcal{B}(U, X_{-1})$ and $C \in \mathcal{B}(X_1, Y)$ (the *control operator* and the *observation operator* of $\mathbf{\Sigma}$, respectively) such that, $\forall\, t \geq 0$, $u \in L^2_{\mathrm{loc}}(\mathbb{R}_+, U)$ and $x_0 \in X_1$,

$$\mathbf{\Phi}_t \mathbf{P}_t u = \int_0^t \mathbf{T}_{t-\tau} B u(\tau)\, d\tau \quad \text{and} \quad (\mathbf{\Psi}_\infty x_0)(\mathrm{t}) = C\mathbf{T}_{\mathrm{t}} x_0.$$

B is said to be *bounded* if it is so as a map from the input space U to the state space X; otherwise, B is said to be *unbounded*. C is said to be *bounded* if it can be extended continuously to X; otherwise, C is said to be *unbounded*. If $\mathbf{T}$ is exponentially stable, then there exist constants β, $\gamma > 0$ such that, $\forall\, t \geq 0$, $u \in L^2(\mathbb{R}_+, U)$, and $x_0 \in X_1$,

$$\|\mathbf{\Phi}_t \mathbf{P}_t u\| = \left\| \int_0^t \mathbf{T}_{t-\tau} B u(\tau)\, d\tau \right\| \leq \beta \|u\|_{L^2(0,t;U)}, \tag{3}$$

$$\|\mathbf{\Psi}_\infty x_0\|_{L^2(0,t;Y)} = \left(\int_0^t \|C\mathbf{T}_\tau x_0\|^2 d\tau \right)^{1/2} \leq \gamma \|x_0\|, \tag{4}$$

wherein, with slight abuse of notation, we write $\|u\|_{L^2(0,t;U)}$ and $\|\mathbf{\Psi}_\infty x_0\|_{L^2(0,t;Y)}$ to denote $\|\mathbf{P}_t u\|_{L^2(\mathbb{R}_+;U)}$ and $\|\mathbf{P}_t \mathbf{\Psi}_\infty x_0\|_{L^2(\mathbb{R}_+;U)}$, respectively. The *Lebesgue extension* of C was adopted in [23] and is defined by

$$C_L x_0 = \lim_{t \to 0} C \frac{1}{t} \int_0^t \mathbf{T}_\tau x_0\, d\tau,$$

where $\mathrm{dom}(C_L)$ is the set of all those $x_0 \in X$ for which the above limit exists. Clearly $X_1 \subset \mathrm{dom}(C_L) \subset X$. Furthermore, for any $x_0 \in X$, we have that $\mathbf{T}_t x_0 \in \mathrm{dom}(C_L)$ for almost all (a.a.) $t \geq 0$ and

$$(\mathbf{\Psi}_\infty x_0)(t) = C_L \mathbf{T}_t x_0 \quad \text{a.a. } t \geq 0. \tag{5}$$

If $\mathbf{\Sigma}$ is regular, then for any $x_0 \in X$ and $u \in L^2_{\mathrm{loc}}(\mathbb{R}_+, U)$, the functions $x(\cdot)$ and $y(\cdot)$, defined by (1a), satisfy the equations

$$\dot{x}(t) = Ax(t) + Bu(t), \quad x(0) = x_0, \tag{6a}$$

$$y(t) = C_L x(t) + Du(t) \tag{6b}$$

for a.a. $t \geq 0$ (in particular $x(t) \in \mathrm{dom}(C_L)$ for a.a. $t \geq 0$). The derivative on the left-hand side of (6a) has, of course, to be understood in X_{-1}. In other words, if we consider the initial-value problem (6a) in the space X_{-1}, then for any $x_0 \in X$ and $u \in L^2_{\mathrm{loc}}(\mathbb{R}_+, U)$, (6a) has unique strong solution (in the sense of Pazy [18, p. 109]) given by the variation of parameters formula

$$t \mapsto x(t) = \mathbf{T}_t x_0 + \int_0^t \mathbf{T}_{t-\tau} B u(\tau)\, d\tau. \tag{7}$$

It has been demonstrated in [20] that if $\mathbf{\Sigma}$ is regular, then $(sI - A)^{-1}BU \subset \mathrm{dom}(C_L)$ $\forall\, s \in \varrho(A)$, and the transfer function $\mathbf{G}$ can be expressed as

$$\mathbf{G}(s) = C_L(sI - A)^{-1}B + D \qquad \forall\, s \in \mathbb{C}_{\omega(\mathbf{T})},$$

which is familiar from finite-dimensional systems theory. The operators A, B, C, and D are called the *generating operators* of $\mathbf{\Sigma}$.

Two technical lemmas. In essence, part (a) of the following lemma provides an estimate, in the $L^2(\mathbb{R}_+, X)$ topology, for the solution of the initial-value problem (6a) with initial data $x_0 \in X$ and input $u \in L^2(\mathbb{R}_+, U)$, part (b) asserts that the solution is in $L^\infty(\mathbb{R}_+, X)$ whenever $x_0 \in X$ and $u \in L^\infty(\mathbb{R}_+, U)$, whilst part (c) establishes that the initial-value problem, again with initial data $x_0 \in X$, has a convergent-input, convergent-state property. Parts (a) and (c) constitute Lemma 2.2 of [9]; proof of part (b) is implicit in the argument establishing Lemma 2.2 of [9].

LEMMA 2.2. *Let (A, B, C, D) be the generating operators of an exponentially stable regular system $\mathbf{\Sigma} = (\mathbf{T}, \mathbf{\Phi}, \mathbf{\Psi}, \mathbf{F})$.*

(a) *There exist constants $\alpha_0, \alpha_1 > 0$ such that $\forall\, (x_0, u) \in X \times L^2(\mathbb{R}_+, U)$, the solution $x(\cdot)$ of the initial-value problem* (6a) *satisfies*

$$\|x\|_{L^2(\mathbb{R}_+, X)} \leq \alpha_0 \|x_0\| + \alpha_1 \|u\|_{L^2(\mathbb{R}_+, U)}.$$

(b) *$\forall\, (x_0, u) \in X \times L^\infty(\mathbb{R}_+, U)$, the solution $x(\cdot)$ of the initial-value problem* (6a) *satisfies*

$$x \in L^\infty(\mathbb{R}_+, X).$$

(c) *If $u \in L^\infty(\mathbb{R}_+, U)$ and $\lim_{t\to\infty} u(t) = u_\infty$ exists, then, $\forall\, x_0 \in X$, the solution $x(\cdot)$ of the initial-value problem* (6a) *satisfies*

$$\lim_{t\to\infty} \|x(t) + A^{-1}Bu_\infty\| = 0.$$

The next lemma shows that, for finite-dimensional U and Y, the impulse response of an exponentially stable regular system with bounded B (or bounded C) is the sum of a weighted L^1-function and a point mass at 0.

LEMMA 2.3. *Let (A, B, C, D) be the generating operators of an exponentially stable regular system $\mathbf{\Sigma} = (\mathbf{T}, \mathbf{\Phi}, \mathbf{\Psi}, \mathbf{F})$. Assume that either B or C is bounded.*

(a) *There exists $\alpha < 0$ such that, $\forall\, u \in U$,*

$$\mathcal{L}^{-1}(\mathbf{G}(\cdot)u - Du) \in L^1_\alpha(\mathbb{R}_+, Y).$$

(b) *If $U = \mathbb{R}^m$ and $Y = \mathbb{R}^p$, then*

$$\mathcal{L}^{-1}(\mathbf{G}) \in L^1_\alpha(\mathbb{R}_+, \mathbb{R}^{p\times m}) + (\mathbb{R}^{p\times m})\, \delta_0,$$

where δ_0 denotes the unit point mass at 0.

Proof. Suppose that B is bounded, and set $\mathbf{G}_0(s) := \mathbf{G}(s) - D = C(sI - A)^{-1}B$. Fix $u \in U$ and choose $(b_n) \subset X_1$ such that $\lim_{n\to\infty} \|Bu - b_n\| = 0$ (such a sequence exists by denseness of X_1 in X). Consequently, $\mathbf{\Psi}_\infty Bu$, $\mathbf{\Psi}_\infty b_n \in L^2(\mathbb{R}_+, Y)$, and

$$\lim_{n\to\infty} \|\mathbf{\Psi}_\infty Bu - \mathbf{\Psi}_\infty b_n\|_{L^2(\mathbb{R}_+,Y)} = 0.$$

Hence, $\forall\ s \in \mathbb{C}_0$,

$$\mathbf{G}_0(s)u = \lim_{n\to\infty} C(sI - A)^{-1}b_n = \lim_{n\to\infty} \int_0^\infty (\mathbf{\Psi}_\infty b_n)(t)e^{-st}\,dt = [\mathcal{L}(\mathbf{\Psi}_\infty Bu)](s).$$

Note that, by exponential stability, $\mathbf{\Psi}_\infty Bu \in L^2_\beta(\mathbb{R}_+, Y)$ for some $\beta < 0$, and hence $\mathbf{\Psi}_\infty Bu \in L^1_\alpha(\mathbb{R}_+, Y)\ \forall\ \alpha \in (\beta, 0)$, which yields part (a) in the case of bounded B. This result together with the duality between admissible control and observation operators (see [23]) yields part (a) in the case of bounded C. Part (b) is an immediate consequence of part (a). ∎

3. Integral control of regular systems with input nonlinearities. The problem of tracking constant reference signals r will be addressed in a context of uncertain single-input ($u(t) \in \mathbb{R}$), single-output ($y(t) \in \mathbb{R}$) linear systems, having a nonlinearity ϕ in the input channel:

$$\begin{aligned} \dot{x}(t) &= Ax(t) + B\phi(u(t)), \quad x(0) = x_0 \in X, \\ y(t) &= C_L x(t) + D\phi(u(t)). \end{aligned} \tag{8}$$

We consider quadruples (A, B, C, D), with C having Lebesgue extension C_L, of class $\mathcal{R}$ defined below.

DEFINITION 3.1. *Let $\mathcal{R}$ denote the class of quadruples (A, B, C, D) which are the generating operators of a regular linear system $\mathbf{\Sigma}$, with state space X, input space $\mathbb{R}$, output space $\mathbb{R}$, and transfer function $\mathbf{G}$, satisfying*

$$\text{(a)}\quad \mathbf{\Sigma} \text{ is exponentially stable}; \qquad \text{(b)}\quad \mathbf{G}(0) > 0.$$

The following property of $\mathcal{R}$ is readily verified.

PROPOSITION 3.2. *If $(A, B, C, D) \in \mathcal{R}$, then $(A + \varepsilon I, B, C, D) \in \mathcal{R}\ \ \forall \varepsilon > 0$ sufficiently small.*

Admissible input nonlinearities are those functions ϕ of class $\mathcal{N}$ defined below.

DEFINITION 3.3. *Let $\mathcal{N}$ be the class of functions $\phi : \mathbb{R} \to \mathbb{R}$ with the properties* (a) *ϕ is monotone nondecreasing;* (b) *ϕ satisfies a global Lipschitz condition (with Lipschitz constant λ), that is, for some λ, $|\phi(u) - \phi(v)| \le \lambda|u - v|\,\forall\, u, v \in \mathbb{R}$.*

As an example of $\phi \in \mathcal{N}$, consider the input nonlinearity in Figure 2 below.

Let $(\varepsilon_n) \subset (0, \infty)$ be a sequence with $\varepsilon_n \downarrow 0$ as $n \to \infty$. For each $\phi \in \mathcal{N}$, define $\phi^\diamond : \mathbb{R} \to \mathbb{R}$ by

$$\phi^\diamond(\xi) = \limsup_{n\to\infty} \frac{\phi(\xi + \varepsilon_n) - \phi(\xi)}{\varepsilon_n}.$$

Note that $0 \le \phi^\diamond(\xi) \le \lambda\ \ \forall\ \xi \in \mathbb{R}$, where λ is a Lipschitz constant for ϕ. Moreover, as the (pointwise) upper limit of a sequence of continuous functions, $\phi^\diamond$ is Borel measurable, and so its composition $\phi^\diamond \circ u$ with a Lebesgue measurable function u is Lebesgue measurable; furthermore, by the same argument as used in proving Lemma 3.5 of [9], a chain rule applies to such compositions, which we now record.

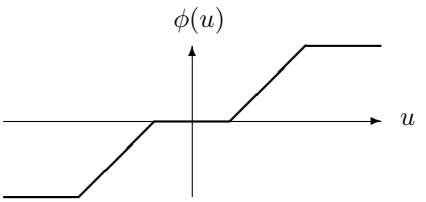


FIG. 2. *Nonlinearity with saturation and dead zone.*

PROPOSITION 3.4. *Let $\phi \in \mathcal{N}$ and let $u : \mathbb{R}_+ \to \mathbb{R}$ be absolutely continuous. Then, $\phi \circ u$ is absolutely continuous and*

$$(\phi \circ u)'(t) = \phi^\diamond(u(t))\dot{u}(t) \quad \text{for a.a. } t \in \mathbb{R}_+.$$

The Clarke [2] directional derivative $\phi^o(u; v)$ of $\phi \in \mathcal{N}$ at u in direction v is given by

$$\limsup_{\substack{w \to u \\ h \downarrow 0}} \frac{\phi(w + hv) - \phi(w)}{h}.$$

Define $\phi^-(\cdot) := -\phi^o(\cdot; -1)$. By upper semicontinuity of ϕ^o, ϕ^- is lower semicontinuous. By definition of $\phi^\diamond$ and monotonicity of $\phi \in \mathcal{N}$ (with Lipschitz constant λ), we have

$$0 \le \phi^-(u) \le \phi^\diamond(u) \le \lambda \quad \forall\, u. \tag{9}$$

$u \in \mathbb{R}$ is said to be a *critical point* (and $\phi(u)$ is said to be a *critical value*) of ϕ if $\phi^-(u) = 0$.

3.1. Integral control with time-varying gain. Let $(A, B, C, D) \in \mathcal{R}$, $\phi \in \mathcal{N}$, and $k \in L^\infty(\mathbb{R}_+, \mathbb{R})$. We denote, by $r \in \mathbb{R}$, the value of the constant reference signal to be tracked by the output $y(t)$. In Proposition 3.6 of [9], it is shown that the following condition is necessary for solvability of the tracking problem: $[\mathbf{G}(0)]^{-1} r \in \operatorname{clos}(\operatorname{im} \phi)$. Reference values r satisfying this condition are referred to as *feasible*.

We will investigate integral control action

$$u(t) = u_0 + \int_0^t k(\tau)[r - C_L x(\tau) - D\phi(u(\tau))]\, d\tau,$$

with time-varying gain $k(\cdot)$, leading to the following nonlinear system of differential equations:

$$\dot{x}(t) = Ax(t) + B\phi(u(t)), \quad x(0) = x_0 \in X, \tag{10a}$$

$$\dot{u}(t) = k(t)[r - C_L x(t) - D\phi(u(t))], \quad u(0) = u_0 \in \mathbb{R}. \tag{10b}$$

For $a \in (0, \infty]$, a continuous function

$$[0, a) \to X \times \mathbb{R}, \quad t \mapsto (x(t), u(t))$$

is a *solution* of (10) if $(x(\cdot), u(\cdot))$ is absolutely continuous as a $(X_{-1} \times \mathbb{R})$-valued function, $x(t) \in \operatorname{dom}(C_L)$ for a.a. $t \in [0, a)$, $(x(0), u(0)) = (x_0, u_0)$, and the differential

equations in (10) are satisfied almost everywhere (a.e.) on $[0, a)$, where the derivative in (10a) should be interpreted in the space X_{-1}.[1]

An application of a well-known result on abstract Cauchy problems (see Pazy [18, Theorem 2.4, p. 107]) shows that a continuous $(X \times \mathbb{R})$-valued function $(x(\cdot), u(\cdot))$ is a solution of (10) if and only if it satisfies the following integrated version of (10):

$$x(t) = \mathbf{T}_t x_0 + \int_0^t \mathbf{T}_{t-\tau} B\phi(u(\tau))\, d\tau, \tag{11a}$$

$$u(t) = u_0 + \int_0^t k(\tau)[r - C_L x(\tau) - D\phi(u(\tau))]\, d\tau. \tag{11b}$$

The next result asserts that (10) has a unique solution: the proof is contained in the appendix.

LEMMA 3.5. *Let $(A, B, C, D) \in \mathcal{R}$, $\phi \in \mathcal{N}$, $k \in L^\infty(\mathbb{R}_+)$, and $r \in \mathbb{R}$. For each $(x_0, u_0) \in X \times \mathbb{R}$, there exists a unique solution $(x(\cdot), u(\cdot))$ of (10) defined on $\mathbb{R}_+$.*

In [9], the constant-gain case is considered in the context of systems $(A, B, C, D) \in \mathcal{R}$ with input nonlinearities $\phi \in \mathcal{N}$: there, the existence of a value $k^* > 0$, with the property that asymptotic tracking of feasible reference signals r is ensured for all fixed gains $k \in (0, k^*)$, is established. However, k^* is, in general, a function of the plant data and so, in the presence of plant uncertainty, may fail to be computable in practice. In such circumstances, one might be led naïvely to consider a time-dependent gain strategy $t \mapsto k(t) > 0$ with $k(t)$ approaching zero as t tends to infinity.

The main result of this section is contained in the following two theorems which confirm the validity of the above naïvety provided that the gain approaches zero sufficiently slowly. In particular, Theorem 3.6 proves that if $t \mapsto k(t) > 0$ is chosen to be bounded and monotone decreasing to zero, then the unique solution of (10) is such that both $x(\cdot)$ and $\phi(u(\cdot))$ converge. The essence of Theorem 3.8 is the assertion that if, in addition, r is feasible and $k(\cdot)$ approaches zero sufficiently slowly, then $\phi(u(\cdot))$ converges to the value $\phi_r := [\mathbf{G}(0)]^{-1} r$, thereby ensuring asymptotic tracking of r.

THEOREM 3.6. *Let $(A, B, C, D) \in \mathcal{R}$, $\phi \in \mathcal{N}$, and $r \in \mathbb{R}$. Let $k : \mathbb{R}_+ \to (0, \infty)$ be a bounded, monotone function with $k(t) \downarrow 0$ as $t \to \infty$. $\forall\ (x_0, u_0) \in X \times \mathbb{R}$, the unique solution $(x(\cdot), u(\cdot))$ of (10) satisfies*

(a) $\lim_{t\to\infty} \phi(u(t))$ *exists and is finite, and*
(b) $\lim_{t\to\infty} \|x(t) + A^{-1} B\phi^*\| = 0$, *where* $\phi^* := \lim_{t\to\infty} \phi(u(t))$.

Proof. Let $(x_0, u_0) \in X \times \mathbb{R}$ be arbitrary. Let λ be a Lipschitz constant for $\phi \in \mathcal{N}$ (and so $0 \le \phi^\diamond(u) \le \lambda \ \ \forall\ u \in \mathbb{R}$). By Lemma 3.5, there exists a unique solution of (10) on $\mathbb{R}_+$. We denote this solution by $(x(\cdot), u(\cdot))$ and introduce new variables by writing $\phi_r = [\mathbf{G}(0)]^{-1} r$ and defining

$$z(t) := x(t) + A^{-1} B\phi(u(t)), \quad v(t) := \phi(u(t)) - \phi_r \quad \forall\ t \in \mathbb{R}_+. \tag{12}$$

By regularity, it follows that $z(t) \in \mathrm{dom}(C_L)$ for a.a. $t \in \mathbb{R}_+$. Moreover, by Proposition 3.4, $\dot{v}(t) = \phi^\diamond(u(t))\dot{u}(t)$ for a.a. $t \in \mathbb{R}_+$. Since (z, v) is absolutely continuous as

[1] Being a Hilbert space, $X_{-1} \times \mathbb{R}$ is reflexive, and hence any absolutely continuous $(X_{-1} \times \mathbb{R})$-valued function is a.e. differentiable and can be recovered from its derivative by integration; see [1, Theorem 3.1, p. 10].

an $(X_{-1} \times \mathbb{R})$-valued function, we obtain by direct calculation

$$\dot{z} = Az - k\phi^\diamond(u)A^{-1}B(C_L z + \mathbf{G}(0)v), \tag{13a}$$
$$z(0) = x_0 + A^{-1}B\phi(u_0),$$
$$\dot{v} = -k\phi^\diamond(u)(C_L z + \mathbf{G}(0)v), \tag{13b}$$
$$v(0) = \phi(u_0) - \phi_r.$$

We claim that there exist positive constants γ_1, γ_2, and σ_1 such that, $\forall\ t, s$ with $\sigma_1 \le s \le t$,

$$\int_s^t |C_L z|\,|k\phi^\diamond(u)v| \le \gamma_1 \|z(s)\| \left(\int_s^t k^2\phi^\diamond(u)v^2\right)^{1/2} + \gamma_2 \int_s^t k^2\phi^\diamond(u)v^2. \tag{14}$$

In order to prove (14), let us first estimate $\int_s^t |C_L z|^2$. For notational convenience, write $w = \phi^\diamond(u)\,[C_L z + \mathbf{G}(0)v]$. As a solution of (13$a$), $z(\cdot)$ satisfies

$$z(\tau) = \mathbf{T}_{\tau - s}z(s) - A^{-1}\int_s^\tau \mathbf{T}_{\tau-\xi}Bk(\xi)w(\xi)\,d\xi$$

$\forall\ s$ with $0 \le s \le \tau$. Invoking (4) and (5) and noting that $C_L A^{-1}$ maps X boundedly into $\mathbb{R}$, there exist constants $\alpha_0, \alpha_1 > 0$ such that

$$\int_s^t |C_L z(\tau)|^2 d\tau \le \alpha_0 \|z(s)\|^2 + \alpha_1 \int_s^t \left\| \int_s^\tau \mathbf{T}_{\tau-\xi}Bk(\xi)w(\xi)d\xi \right\|^2 d\tau \tag{15}$$

$\forall\, 0 \le s \le t$. By Lemma 2.2 (part (a)), interpreted in the context of the initial-value problem

$$\dot{\zeta} = A\zeta + Bkw, \quad \zeta(s) = 0,$$

we have

$$\left(\int_s^t \left\| \int_s^\tau \mathbf{T}_{\tau-\xi}Bk(\xi)w(\xi)d\xi \right\|^2 d\tau\right)^{1/2} \le \alpha_2 \left(\int_s^t |kw|^2\right)^{1/2}$$

for some constant α_2. Therefore, by (15) and monotonicity of k, it follows that, for some constants $\alpha_3, \alpha_4 > 0$,

$$\left(\int_s^t |C_L z|^2\right)^{1/2} \le \alpha_3\|z(s)\| + k(s)\alpha_4 \left(\int_s^t |\phi^\diamond(u)|^2 |C_L z|^2\right)^{1/2}$$
$$+ \alpha_4 \mathbf{G}(0) \left(\int_s^t |k\phi^\diamond(u)v|^2\right)^{1/2} \quad \forall\ 0 \le s \le t. \tag{16}$$

Fix $\sigma_1 > 0$ such that $\delta := k(\sigma_1)\alpha_4\lambda < 1$. Then,

$$k(s)\alpha_4 \left(\int_s^t |\phi^\diamond(u)|^2 |C_L z|^2\right)^{1/2} \le \delta \left(\int_s^t |C_L z|^2\right)^{1/2} \quad \forall\ \sigma_1 \le s \le t,$$

and so, by (16),

$$\left(\int_s^t |C_L z|^2\right)^{1/2} \le \beta_1\|z(s)\| + \beta_2 \left(\int_s^t k^2\phi^\diamond(u)v^2\right)^{1/2} \quad \forall\ \sigma_1 \le s \le t, \tag{17}$$

with $\beta_1 = \alpha_3/(1-\delta)$ and $\beta_2 = \alpha_4 \mathbf{G}(0)\sqrt{\lambda}/(1-\delta)$. We may now deduce that, $\forall\ t, s$ with $\sigma_1 \le s \le t$,

$$\int_s^t |C_L z|\,|k\phi^\diamond(u)v| \le \left(\int_s^t |C_L z|^2\right)^{1/2} \left(\int_s^t |k\phi^\diamond(u)v|^2\right)^{1/2}$$
$$\le \beta_1\sqrt{\lambda}\|z(s)\| \left(\int_s^t k^2\phi^\diamond(u)v^2\right)^{1/2} + \beta_2\sqrt{\lambda}\int_s^t k^2\phi^\diamond(u)v^2,$$

which is (14) with $\gamma_1 = \beta_1\sqrt{\lambda}$ and $\gamma_2 = \beta_2\sqrt{\lambda}$. By (13b), for a.a. $t \ge 0$,

$$v(t)\dot{v}(t) = -k(t)\mathbf{G}(0)\phi^\diamond(u(t))v^2(t) - k(t)\phi^\diamond(u(t))v(t)C_L z(t), \tag{18}$$

and hence

$$v(t)\dot{v}(t) \le -k(t)\mathbf{G}(0)\phi^\diamond(u(t))v^2(t) + |C_L z(t)|\,|k(t)\phi^\diamond(u(t))v(t)|.$$

Integrating this inequality and using (14) and monotonicity of k yields, $\forall\ t, s$ with $\sigma_1 \le s \le t$,

$$v^2(t) \le v^2(s) + 2\gamma_1\sqrt{k(s)}\|z(s)\| \left(\int_s^t k\phi^\diamond(u)v^2\right)^{1/2}$$
$$+ 2\int_s^t (k\gamma_2 - \mathbf{G}(0))k\phi^\diamond(u)v^2. \tag{19}$$

By positivity of $\mathbf{G}(0)$ and monotonicity of $k(\cdot)$, there exists $\sigma \ge \sigma_1$ such that, $\forall\,\tau \ge \sigma$, $(k(\tau)\gamma_2 - \mathbf{G}(0)) \le -\frac{1}{2}\mathbf{G}(0) < 0$. Therefore, it follows from (19) that

$$0 \le v^2(\sigma) + 2\gamma_1\sqrt{k(\sigma)}\|z(\sigma)\| \left(\int_\sigma^t k\phi^\diamond(u)v^2\right)^{1/2} - \mathbf{G}(0)\int_\sigma^t k\phi^\diamond(u)v^2 \quad \forall\, t \ge \sigma,$$

and so

$$\int_\sigma^\infty k\phi^\diamond(u)v^2 < \infty. \tag{20}$$

Moreover, by (14) we deduce that

$$\int_\sigma^\infty |C_L z|\,|k\phi^\diamond(u)v| < \infty. \tag{21}$$

Combining (18), (20), and (21) shows that there exists a number $\nu \in \mathbb{R}_+$ such that

$$\lim_{t\to\infty} v^2(t) = v^2(\sigma) + 2\lim_{t\to\infty}\int_\sigma^t v\dot{v} = \nu,$$

whence assertion (a) of the theorem. Assertion (b) now follows by Lemma 2.2 (part (c)). ∎

Let $\mathcal{M}$ denote the space of finite signed Borel measures on $\mathbb{R}_+$.

LEMMA 3.7. *Let $(A, B, C, D) \in \mathcal{R}$, $\phi \in \mathcal{N}$, and $(x_0, u_0) \in X \times \mathbb{R}$. Assume that $\mathcal{L}^{-1}(\mathbf{G}) \in \mathcal{M}$. Let $k : \mathbb{R}_+ \to (0,\infty)$ be bounded and such that $\int_0^t k =: K(t) \to \infty$ as $t \to \infty$. Let $r \in \mathbb{R}$ be feasible, that is, $\phi_r := [\mathbf{G}(0)]^{-1}r \in \operatorname{clos}(\operatorname{im}\phi)$. Let $(x(\cdot), u(\cdot)) : \mathbb{R}_+ \to X \times \mathbb{R}$ be the unique solution of* (10).

If $\lim_{t\to\infty}\phi(u(t))$ exists and is finite, then the following statements hold:

(a) $\lim_{t\to\infty}\phi(u(t))=\phi_r$,
(b) $\lim_{t\to\infty}\|x(t)+A^{-1}B\phi_r\|=0$,
(c) $\lim_{t\to\infty}[r-y(t)+(\mathbf{\Psi}_\infty x_0)(t)]=0$, *where* $y(t)=C_Lx(t)+D\phi(u(t))$,
(d) *if* $\phi_r\in\operatorname{im}\phi$, *then* $\lim_{t\to\infty}\operatorname{dist}(u(t),\phi^{-1}(\phi_r))=0$, *and*
(e) *if* $\phi_r\in\operatorname{int}(\operatorname{im}\phi)$, *then* $u(\cdot)$ *is bounded.*

Proof. By hypothesis, there exists $\phi^*\in\mathbb{R}$ such that $\lim_{t\to\infty}\phi(u(t))=\phi^*$. The essence of the proof is to show that $\phi^*=\phi_r$. Setting

$$y_0(t)=(\mathbf{\Psi}_\infty)(x_0)(t),\quad y_1(t)=(\mathcal{L}^{-1}(\mathbf{G})\star\phi(u))(t),$$

we have

$$\dot u(t)=k(t)[\mathbf{G}(0)(\phi_r-\phi^*)-y_0(t)-(y_1(t)-\mathbf{G}(0)\phi^*)].$$

Seeking a contradiction, suppose that $|\phi_r-\phi^*|\neq 0$. Since $\lim_{t\to\infty}\phi(u(t))=\phi^*$ and $\mathcal{L}^{-1}(\mathbf{G})\in\mathcal{M}$, it follows that $\lim_{t\to\infty}y_1(t)=\mathbf{G}(0)\phi^*$ (see [7, Theorem 6.1, part (ii), p. 96]). Let $s>0$ be such that

$$|y_1(t)-\mathbf{G}(0)\phi^*|\le\tfrac12\mathbf{G}(0)|\phi_r-\phi^*|\quad\forall\ t\ge s.$$

As a consequence we obtain

$$\begin{aligned}-\tfrac12k(t)\mathbf{G}(0)|\phi_r-\phi^*|-k(t)|y_0(t)|&\le\dot u(t)-k(t)\mathbf{G}(0)(\phi_r-\phi^*)\\&\le\tfrac12k(t)\mathbf{G}(0)|\phi_r-\phi^*|+k(t)|y_0(t)|\quad\forall\ t\ge s.\end{aligned}\tag{22}$$

Since $\phi_r\neq\phi^*$, either $\phi_r>\phi^*$ or $\phi_r<\phi^*$. If $\phi_r>\phi^*$, then

$$\tfrac12k(t)\mathbf{G}(0)(\phi_r-\phi^*)-k(t)|y_0(t)|\le\dot u(t)\quad\forall\ t\ge s,$$

which, on integration, yields

$$\tfrac12\mathbf{G}(0)(K(t)-K(s))(\phi_r-\phi^*)-\int_s^t k(\tau)|y_0(\tau)|d\tau\le u(t)-u(s)\quad\forall\ t\ge s.$$

By exponential stability, $y_0\in L^2_\alpha(\mathbb{R}_+,\mathbb{R})$ for some $\alpha<0$, and thus $y_0\in L^1(\mathbb{R}_+,\mathbb{R})$, which in turn implies that $ky_0\in L^1(\mathbb{R}_+,\mathbb{R})$. Since $K(t)\to\infty$ as $t\to\infty$, we conclude that $u(t)\to\infty$ as $t\to\infty$, whence the contradiction

$$\phi_r\le\sup\phi=\lim_{t\to\infty}\phi(u(t))=\phi^*<\phi_r.$$

If $\phi_r<\phi^*$, then

$$\dot u(t)\le k(t)|y_0(t)|-\tfrac12k(t)\mathbf{G}(0)|\phi_r-\phi^*|\quad\forall\ t\ge s,$$

which, on integration, yields $u(t)\to-\infty$ as $t\to\infty$ and the contradiction

$$\phi_r<\phi^*=\lim_{t\to\infty}\phi(u(t))=\inf\phi\le\phi_r.$$

Therefore, we may conclude that $\lim_{t\to\infty}\phi(u(t))=\phi_r$, which is assertion (a). Assertion (b) follows from Lemma 2.2, part (c); assertion (c) is a consequence of assertion (a), together with the identity

$$r-y(t)+(\mathbf{\Psi}_\infty x_0)(t)=\mathbf{G}(0)\phi_r-(\mathcal{L}^{-1}(\mathbf{G})\star\phi(u))(t),$$

and the fact that $\lim_{t\to\infty}(\mathcal{L}^{-1}(\mathbf{G}) \star \phi(u))(t) = \mathbf{G}(0)\phi_r$.

To prove assertion (d), let $\phi_r \in \operatorname{im}\phi$ and suppose that the claim is false. Then there exists a sequence $(t_n) \subset \mathbb{R}_+$ with $\lim_{n\to\infty} t_n = \infty$ and $\varepsilon > 0$ such that

$$\operatorname{dist}(u(t_n), \phi^{-1}(\phi_r)) \geq \varepsilon \quad \forall\ n. \tag{23}$$

If the sequence $(u(t_n))$ is bounded, we may assume without loss of generality that it converges to a finite limit u_∞. By continuity of ϕ and assertion (a), we have that $\phi(u_\infty) = \phi_r$, and thus $u_\infty \in \phi^{-1}(\phi_r)$. This contradicts (23). So, suppose that $(u(t_n))$ is unbounded. Without loss of generality, we may then assume that $\lim_{n\to\infty} u(t_n) = \infty$. By monotonicity and assertion (a), it follows that $\phi_r = \sup\phi$. Since $\phi_r \in \operatorname{im}\phi$, there exists ξ^* such that $\phi(\xi^*) = \phi_r = \sup\phi = \max\phi$. By monotonicity of ϕ, $\phi(\xi) = \phi_r = \max\phi \ \ \forall\ \xi \geq \xi^*$. In particular, we see that $u(t_n) \in \phi^{-1}(\phi_r)$ for all sufficiently large n, which contradicts (23).

Now, assume that $\phi_r \in \operatorname{int}(\operatorname{im}\phi)$ and, for contradiction, suppose that assertion (e) is false. Then there exists a sequence $(t_n) \subset (0,\infty)$ with $t_n \to \infty$ and $|u(t_n)| \to \infty$ as $n \to \infty$. Without loss of generality, we may assume that $\lim_{n\to\infty} u(t_n) = \infty$. By monotonicity, it then follows that $\phi_r = \lim_{n\to\infty}\phi(u(t_n)) = \sup\phi$, contradicting the assumption that $\phi_r \in \operatorname{int}(\operatorname{im}\phi)$. □

For $\alpha \in \mathbb{R}$, we define the exponentially weighted space $\mathcal{M}_\alpha$ as the set of all locally finite signed Borel measures μ on $\mathbb{R}_+$ (see, e.g., [6]) with the property that the weighted measure $E \mapsto \int_E e^{-\alpha t}\mu(dt)$ belongs to $\mathcal{M}$.

THEOREM 3.8. *Let $(A,B,C,D) \in \mathcal{R}$, $\phi \in \mathcal{N}$, and $(x_0,u_0) \in X \times \mathbb{R}$. Assume that $\mathcal{L}^{-1}(\mathbf{G}) \in \mathcal{M}$. Let $k : \mathbb{R}_+ \to (0,\infty)$ be bounded, monotone, and such that $k(t) \downarrow 0$ and $\int_0^t k =: K(t) \to \infty$ as $t \to \infty$. Let $r \in \mathbb{R}$ be feasible, that is, $\phi_r := [\mathbf{G}(0)]^{-1}r \in \operatorname{clos}(\operatorname{im}\phi)$.*

The unique solution $(x(\cdot),u(\cdot)) : \mathbb{R}_+ \to X \times \mathbb{R}$ of (10) satisfies

(a) $\lim_{t\to\infty}\phi(u(t)) = \phi_r$,

(b) $\lim_{t\to\infty}\|x(t) + A^{-1}B\phi_r\| = 0$,

(c) $\lim_{t\to\infty}[r - y(t) + (\mathbf{\Psi}_\infty x_0)(t)] = 0$, *where* $y(t) = C_L x(t) + D\phi(u(t))$,

(d) *if* $\phi_r \in \operatorname{im}\phi$, *then* $\lim_{t\to\infty}\operatorname{dist}(u(t),\phi^{-1}(\phi_r)) = 0$,

(e) *if* $\phi_r \in \operatorname{int}(\operatorname{im}\phi)$, *then $u(\cdot)$ is bounded, and*

(f) *if* $\phi_r \in \operatorname{im}\phi$ *is not a critical value of ϕ, then the convergence in* (a) *and* (b) *is of order* $\exp(-\rho K(t))$ *for some $\rho > 0$; moreover, the convergence in* (c) *is of the same order, provided that* $\mathcal{L}^{-1}(\mathbf{G}) \in \mathcal{M}_\alpha$ *for some $\alpha < 0$.*

REMARK 3.9. (i) *The assumption that* $\mathcal{L}^{-1}(\mathbf{G}) \in \mathcal{M}$ *(or that* $\mathcal{L}^{-1}(\mathbf{G}) \in \mathcal{M}_\alpha$ *for some $\alpha < 0$) is not very restrictive and seems to be satisfied in all practical examples of exponentially stable systems. In particular, this assumption is satisfied if B or C is bounded (see Lemma* 2.3*).*

(ii) *Since* $(\mathbf{\Psi}_\infty x_0)(t)$ *converges exponentially to* 0 *as* $t \to \infty\ \forall\ x_0 \in X_1 = \operatorname{dom}(A)$, *it follows from* (c) *that the error* $e(t) = r - y(t)$ *converges to* 0 $\forall\ x_0 \in \operatorname{dom}(A)$. *(Moreover, the convergence is of order* $\exp(-\rho K(t))$ *if the extra assumptions in* (f) *are satisfied.) If C is bounded, then this statement is true $\forall\ x_0 \in X$. If C is unbounded and* $x_0 \notin \operatorname{dom}(A)$, *then $e(t)$ does not necessarily converge to* 0 *as* $t \to \infty$. *However, $e(t)$ is small for large t in the sense that $e(t) = e_1(t) + e_2(t)$, where the function e_1 is bounded with* $\lim_{t\to\infty} e_1(t) = 0$ *and* $e_2 \in L^2_\alpha(\mathbb{R}_+,\mathbb{R})$ *for some $\alpha < 0$.*

(iii) *In particular,* (d) *asserts that $u(t)$ converges as $t \to \infty$ if the set* $\phi^{-1}(\phi_r)$ *is a singleton, which, in turn, will be true if* $\phi_r \in \operatorname{im}\phi$ *is not a critical value of ϕ.*

Proof. Assertions (a)–(e) follow immediately from Theorem 3.6 combined with Lemma 3.7. It remains only to establish assertion (f). By hypothesis, $\phi_r \in \operatorname{im}\phi$ is

not a critical value of ϕ, and so, by monotonicity, its preimage $\phi^{-1}(\phi_r)$ is a singleton $\{u_r\}$ and

$$\phi^{\diamond}(u_r) \geq \phi^{-}(u_r) =: \phi_r^{-} > 0.$$

Therefore, by assertion (d), $u(t) \to u_r$ as $t \to \infty$. By lower semicontinuity of ϕ^{-}, there exists $\sigma_1 > 0$ such that

$$\phi^{\diamond}(u(t)) \geq \phi^{-}(u(t)) \geq \tfrac{1}{2}\phi_r^{-} > 0 \quad \forall\, t \geq \sigma_1. \tag{24}$$

Define $\rho := \frac{1}{4}\mathbf{G}(0)\phi_r^{-} > 0$, and introduce exponentially weighted variables given by

$$z_e(t) := \exp(\rho K(t))[x(t) + A^{-1}B\phi(u(t))], \tag{25a}$$

$$v_e(t) := \exp(\rho K(t))[\phi(u(t)) - \phi_r] \tag{25b}$$

$\forall\, t \in \mathbb{R}_+$. Since (z_e, v_e) is absolutely continuous as an $(X_{-1} \times \mathbb{R})$-valued function, and using Proposition 3.4, we obtain by direct calculation

$$\begin{aligned} &\dot{z_e} = (A + \rho k I)z_e - k\phi^{\diamond}(u)A^{-1}B(C_L z_e + \mathbf{G}(0)v_e), \\ &z_e(0) = x_0 + A^{-1}B\phi(u_0), \end{aligned} \tag{26a}$$

$$\begin{aligned} &\dot{v}_e = -k\phi^{\diamond}(u)(C_L z_e + \mathbf{G}(0)v_e) + \rho k v_e, \\ &v_e(0) = \phi(u_0) - \phi_r. \end{aligned} \tag{26b}$$

For each (t, s) with $0 \leq s \leq t$, define

$$\mathbf{U}(t, s) := \exp(\rho[K(t) - K(s)])\mathbf{T}_{t-s}. \tag{27}$$

We briefly digress to prove a technicality.

LEMMA 3.10. *Let $s \in \mathbb{R}_+$, $u \in L^2_{\mathrm{loc}}(\mathbb{R}_+)$, and, on $[s, \infty)$, define a function p by*

$$p(t) := \int_s^t \mathbf{U}(t, \xi)Bu(\xi)d\xi.$$

Then, $\forall\, t$, $p(t) \in X$ and, as an X_{-1}-valued function, p is absolutely continuous with

$$\dot{p}(t) = (A + \rho k(t)I)p(t) + Bu(t) \quad \text{a.e.}$$

Proof. On $[s, \infty)$, define a function q by

$$q(t) := e^{-\rho K(t)}p(t) = \int_s^t \mathbf{T}_{t-\xi}Be^{-\rho K(\xi)}u(\xi)\, d\xi.$$

Clearly, $q(t) \in X\ \forall\, t$ and so $p(t) \in X\ \forall\, t$. Moreover, q is absolutely continuous as an X_{-1}-valued function and, by Pazy [18, Theorem 2.9, p. 109], for a.a. $t \geq s$,

$$\dot{q}(t) = Aq(t) + e^{-\rho K(t)}Bu(t).$$

Thus, as an X_{-1}-valued function, p is absolutely continuous with

$$\dot{p}(t) = \rho k(t)e^{\rho K(t)}q(t) + e^{\rho K(t)}\dot{q}(t) = (A + \rho k(t)I)p(t) + Bu(t)$$

for a.a. $t \geq s$. ∎

Returning to the proof of the theorem, for notational convenience write

$$w_e := \phi^{\diamond}(u)\left[C_L z_e + \mathbf{G}(0)v_e\right].$$

Let $s \in \mathbb{R}_+$ and, on $[s,\infty)$, define $f := f_1 - A^{-1}f_2$ with

$$f_1(t) := \mathbf{U}(t,s)z_e(s), \quad f_2(t) := \int_s^t \mathbf{U}(t,\xi)Bk(\xi)w_e(\xi)d\xi.$$

Clearly, $f_1(t) \in X \ \forall\ t$ and, as an X_{-1}-valued function, f_1 is absolutely continuous with

$$\dot{f}_1(t) = (A + \rho k(t)I)f_1(t) \quad \text{a.e.}$$

By Lemma 3.10, it now follows that $f(t) \in X \ \forall\ t$ and, as an X_{-1}-valued function, f is absolutely continuous with

$$\begin{aligned}\dot{f}(t) &= (A + \rho k(t)I)f_1(t) - A^{-1}\left((A + \rho k(t)I)f_2(t) + Bw_e(t)\right)\\ &= (A + \rho k(t)I)f(t) - A^{-1}Bw_e(t) \quad \text{a.e.}\end{aligned}$$

In view of (26a) (together with uniqueness of solutions), we may now conclude that

$$z_e(t) = \mathbf{U}(t,s)z_e(s) - A^{-1}\int_s^t \mathbf{U}(t,\xi)Bk(\xi)w_e(\xi)\,d\xi \quad \forall\ t,s \text{ with } 0 \le s \le t. \tag{28}$$

By exponential stability of the semigroup $\mathbf{T}$, there exist constants $N,\ \nu > 0$ such that $\|\mathbf{T}_t\| \le N\exp(-\nu t)\ \forall\ t \in \mathbb{R}_+$. Let $\varepsilon \in (0,\nu)$ be sufficiently small such that $(A + \varepsilon I, B, C, D) \in \mathcal{R}$ (recall Proposition 3.2). Fix $\sigma_2 > \sigma_1$ such that

$$k(\sigma_2) < \min\{\varepsilon/\rho, \nu/(\rho N)\}. \tag{29}$$

Again, we digress to prove a technicality.

LEMMA 3.11. *There exists constant $\gamma > 0$ such that, $\forall\ u \in L^2_{\mathrm{loc}}(\mathbb{R}_+)$,*

$$\left(\int_s^t \left\|\int_s^\tau \mathbf{U}(\tau,\xi)Bu(\xi)d\xi\right\|^2 d\tau\right)^{\frac{1}{2}} \le \gamma\left(\int_s^t u^2(\xi)d\xi\right)^{\frac{1}{2}} \quad \forall\ s,t \ge \sigma_2 \text{ with } s \le t.$$

Proof. Let $s \ge \sigma_2$ and let $u \in L^2_{\mathrm{loc}}(\mathbb{R}_+)$ be arbitrary. On $[s,\infty)$ define (as before) the function $p : t \mapsto \int_s^t \mathbf{U}(t,\xi)Bu(\xi)d\xi$. By Lemma 3.10, for a.a. $t \ge s$, we have $\dot{p}(t) = Ap(t) + \rho k(t)p(t) + Bu(t)$. Therefore,

$$p(t) = \int_s^t \mathbf{T}_{t-\xi}\rho k(\xi)p(\xi)\,d\xi + \int_s^t \mathbf{T}_{t-\xi}Bu(\xi)\,d\xi \quad \forall\ t \ge s.$$

Using exponential stability of the semigroup $\mathbf{T}$, monotonicity of k, and Lemma 2.2 (part (a)), we may conclude

$$\left(\int_s^t \|p(\xi)\|^2 d\xi\right)^{\frac{1}{2}} \le \rho N\nu^{-1}k(\sigma_2)\left(\int_s^t \|p(\xi)\|^2 d\xi\right)^{\frac{1}{2}} + \alpha_1\left(\int_s^t u^2(\xi)\,d\xi\right)^{\frac{1}{2}} \quad \forall\ t \ge s.$$

By (29), $1 - \rho N\nu^{-1}k(\sigma_2) > 0$, and the result follows. ∎

Once more, we return to the proof of the theorem. By monotonicity of k, $K(t) - K(s) \le k(s)(t-s)\,\forall\, t, s$ with $0 \le s \le t$. Since $k(\sigma_2) \le \varepsilon/\rho$, it follows that

$$\exp(\rho[K(t) - K(s)]) \le \exp(\varepsilon[t-s]) \quad \forall\ t, s \text{ with } \sigma_2 \le s \le t. \tag{30}$$

Observe that, $\forall\, t, s$ with $\sigma_2 \le s \le t$,

$$\begin{aligned} |C_L \mathbf{U}(t,s) z_e(s)| &= |C_L \mathbf{T}_{t-s} z_e(s)| \exp(\rho[K(t) - K(s)]) \\ &\le |C_L \mathbf{T}_{t-s} \exp(\varepsilon[t-s]) z_e(s)|. \end{aligned}$$

Invoking (4), (5) (in the context of the regular system $(A + \varepsilon I, B, C, D)$), (28), and Lemma 3.11, and recalling that $C_L A^{-1}$ maps X boundedly into $\mathbb{R}$, there exist constants $\alpha_2, \alpha_3 > 0$ such that

$$\begin{aligned} \left(\int_s^t |C_L z_e|^2 \right)^{1/2} \le \alpha_2 \|z_e(s)\| + k(s)\alpha_3 \left(\int_s^t |\phi^\diamond(u)|^2 |C_L z_e|^2 \right)^{1/2} \\ + \alpha_3 \mathbf{G}(0) \left(\int_s^t |k\phi^\diamond(u) v_e|^2 \right)^{1/2} \end{aligned} \tag{31}$$

$\forall\, t, s$ with $\sigma_2 \le s \le t$.

Inequality (31) is the exponentially weighted version of (16). Following the argument in the proof of Theorem 3.6, (31) may be used to derive an exponentially weighted version of (14), i.e., there exist positive constants $\gamma_1, \gamma_2 > 0$ and $\sigma_3 \ge \sigma_2$ such that

$$\int_s^t |C_L z_e|\, |k\phi^\diamond(u) v_e| \le \gamma_1 \|z_e(s)\| \left(\int_s^t k^2 \phi^\diamond(u) v_e^2 \right)^{1/2} + \gamma_2 \int_s^t k^2 \phi^\diamond(u) v_e^2 \tag{32}$$

$\forall\ t, s$ with $\sigma_3 \le s \le t$.

By (26b), for a.a. $t \ge 0$,

$$v_e(t)\dot{v}_e(t) = -k(t)\mathbf{G}(0)\phi^\diamond(u(t))v_e^2 + \rho k(t) v_e^2(t) - k(t)\phi^\diamond(u(t)) v_e(t) C_L z_e(t). \tag{33}$$

By (24), $\mathbf{G}(0)\phi^\diamond(u(t)) - \rho \ge \frac{1}{2}\mathbf{G}(0)\phi_r^- - \rho = \rho > 0\ \forall\ t \ge \sigma_3$. Hence, we have

$$v_e(t)\dot{v}_e(t) \le -\tfrac{1}{2}\rho k(t) v_e^2(t) + |C_L z_e(t)|\, |k(t)\phi^\diamond(u(t)) v_e(t)| \quad \text{for a.a. } t \ge \sigma_3.$$

Integrating this inequality and using (32) and monotonicity of k yields, $\forall\ t, s$ with $t \ge s \ge \sigma_3$,

$$v_e^2(t) \le v_e^2(s) + 2\gamma_1 \sqrt{\lambda k(s)} \|z_e(s)\| \left(\int_s^t k v_e^2 \right)^{1/2} - \int_s^t (\rho - 2k\gamma_2\lambda) k v_e^2. \tag{34}$$

Fix $\sigma \ge \sigma_3$ such that $\rho - 2k(t)\gamma_2\lambda > \frac{1}{2}\rho\ \forall\ t \ge \sigma$. From (34) and (32), we deduce

$$\int_\sigma^\infty k v_e^2 < \infty.$$

Hence, by (34), $v_e(\cdot) = \exp(\rho K(\cdot))[\phi(u(\cdot)) - \phi_r]$ is bounded and so the convergence in (a) is of order $\exp(-\rho K(t))$. We proceed to prove that the convergence in (b) is of the same order. Define $x_r := -A^{-1}B\phi_r$, and introduce a new variable given by

$$x_e(t) = \exp(\rho K(t))[x(t) - x_r] \quad \forall\ t \ge 0.$$

It suffices to show that $x_e(\cdot)$ is bounded. By (8) and (25), we have

$$\dot{x}_e = (A + \rho k I)x_e + Bv_e, \quad x_e(0) = x_0 - x_r,$$

and so, $\forall\ t \geq \sigma$,

$$x_e(t) = \mathbf{T}_{t-\sigma}x_e(\sigma) + \int_\sigma^t \mathbf{T}_{t-\xi}Bv_e(\xi)\,d\xi + \int_\sigma^t \mathbf{T}_{t-\xi}\rho k(\xi)x_e(\xi)\,d\xi.$$

Therefore, by boundedness of v_e together with Lemma 2.2 (part (b)) and exponential stability of $\mathbf{T}$, there exists a constant $\beta > 0$ such that

$$\sup_{s\in[\sigma,t]} \|x_e(s)\| \leq \beta + \rho N \nu^{-1} k(\sigma) \sup_{s\in[\sigma,t]} \|x_e(s)\| \quad \forall\ t \geq \sigma.$$

Since $\sigma \geq \sigma_2$, we have, by (29), $\rho N \nu^{-1} k(\sigma) < 1$, and hence we may conclude boundedness of x_e. Therefore, the convergence in part (b) is of order $\exp(-\rho K(t))$.

It remains only to prove that the convergence in (c) is also of order $\exp(-\rho K(t))$, provided that $\mu := \mathcal{L}^{-1}(\mathbf{G}) \in \mathcal{M}_\alpha$ for some $\alpha < 0$. Denoting the unit-step function by θ, we have $\forall\ t \geq 0$

$$|r - y(t) + (\mathbf{\Psi}_\infty x_0)(t)| \leq |[\mu \star (\phi(u) - \phi_r\theta)(t)]| + |\phi_r[(\mu \star \theta)(t) - \mathbf{G}(0)]|. \tag{35}$$

For convenience we set $w(t) = \exp(\rho K(t))\ \forall\ t \geq 0$. We have already shown that the function $t \mapsto w(t)|\phi(u(t)) - \phi_r|$ remains bounded as $t \to \infty$. If we extend w to a function defined on $\mathbb{R}$ by setting $w(t) = 1\ \forall\ t < 0$, then it is easy to show that w is a submultiplicative weight function in the sense of [7, p. 118]. Moreover, since $\mu \in \mathcal{M}_\alpha$, the measure $\mu_w : E \mapsto \int_E w(t)\mu(dt)$ belongs to $\mathcal{M}$. Hence, by [7, Theorem 3.5, part (i), p. 119], we may conclude that the function $t \mapsto w(t)[\mu\star(\phi(u)-\phi_r\theta)](t)$ is bounded on $\mathbb{R}_+$.

Since $\mu_w \in \mathcal{M}$ (a space of finite measures), $\int_0^\infty w(t)|\mu|(dt)$, where $|\mu|$ denotes the total variation of μ. Hence

$$|w(t)[(\mu \star \theta)(t) - \mathbf{G}(0)]| = w(t)\left|\int_t^\infty \mu(d\tau)\right| \leq \int_0^\infty w(\tau)|\mu|(d\tau) < \infty,$$

showing that the function $t \mapsto w(t)[(\mu\star\theta)(t)-\mathbf{G}(0)]$ is bounded on $\mathbb{R}_+$. Consequently, appealing to (35), we deduce that the function

$$\mathbb{R}_+ \to \mathbb{R}, \quad t \mapsto \exp(\rho K(t))|r - y(t) + (\mathbf{\Psi}_\infty x_0)(t)|$$

is bounded. □

3.2. Adaptive gain. Whilst Theorem 3.8 identifies conditions under which the tracking objective is achieved through the use of a monotone gain function, the resulting control strategy is somewhat unsatisfactory insofar as the gain function is selected a priori: no use is made of the output information from the plant to update the gain. We now consider the possibility of exploiting this output information to generate, by feedback, an appropriate gain function. In particular, let the gain $k(\cdot)$ be generated by the law:

$$k(t) = \frac{1}{l(t)}, \quad \dot{l}(t) = |r - y(t)|, \quad l(0) = l_0 > 0, \tag{36}$$

which yields the feedback system

$$\dot{x}(t) = Ax(t) + B\phi(u(t)), \quad x(0) = x_0 \in X, \tag{37a}$$

$$\dot{u}(t) = (1/l(t))[r - C_L x(t) - D\phi(u(t))], \quad u(0) = u_0 \in \mathbb{R}, \tag{37b}$$

$$\dot{l}(t) = |r - C_L x(t) - D\phi(u(t))|, \quad l(0) = l_0 \in (0, \infty). \tag{37c}$$

The concept of a solution of this feedback system is the obvious modification of the solution concept defined in subsection 3.1. The proof of the following existence and uniqueness result can be found in the appendix.

LEMMA 3.12. *Let* $(A, B, C, D) \in \mathcal{R}$, $\phi \in \mathcal{N}$, *and* $r \in \mathbb{R}$. *For each* $(x_0, u_0, l_0) \in X \times \mathbb{R} \times (0, \infty)$, *the initial-value problem given by* (37) *has a unique solution defined on* $\mathbb{R}_+$.

We now arrive at the main adaptive control result.

THEOREM 3.13. *Let* $(A, B, C, D) \in \mathcal{R}$, $\phi \in \mathcal{N}$, *and let* $r \in \mathbb{R}$ *be such that* $\phi_r := [\mathbf{G}(0)]^{-1} r \in \operatorname{clos}(\operatorname{im} \phi)$. *Assume that* $\mathcal{L}^{-1}(\mathbf{G}) \in \mathcal{M}$.

$\forall\ (x_0, u_0, l_0) \in X \times \mathbb{R} \times (0, \infty)$, *the unique solution of the initial-value problem given by* (37) *is such that assertions* (a) *to* (e) *of Theorem* 3.8 *hold. Moreover, if* $\phi_r \in \operatorname{im} \phi$ *is not a critical value and* $\mathcal{L}^{-1}(\mathbf{G}) \in \mathcal{M}_\alpha$ *for some* $\alpha < 0$, *then the monotone gain* k *converges to a positive value.*

Proof. Set $k(t) = 1/l(t)$. Since $l(\cdot)$ is monotone increasing, either $l(t) \to \infty$ as $t \to \infty$ (Case 1), or $l(t) \to l^* \in (0, \infty)$ as $t \to \infty$ (Case 2). We consider these two cases separately.

Case 1. In this case, $k(t) \downarrow 0$ as $t \to \infty$ and the hypotheses of Theorem 3.6 are satisfied. Therefore, $(\phi(u))(\cdot)$ converges. It follows that $\lim_{t\to\infty}(\mathcal{L}^{-1}(\mathbf{G}) \star \phi(u))(t)$ converges (and so, in particular, is a bounded function). Moreover, by exponential stability, $\mathbf{\Psi}_\infty x_0 \in L^1(\mathbb{R}_+, \mathbb{R})$, and it follows from

$$\dot{l}(t) = |r - y(t)| \le |r - (\mathcal{L}^{-1}(\mathbf{G}) \star \phi(u))(t)| + |(\mathbf{\Psi}_\infty x_0)(t)|,$$

via integration that

$$k(t) = \frac{1}{l(t)} \ge \frac{1}{\alpha + \beta t} \quad \forall\ t \ge 0, \tag{38}$$

where

$$\alpha := l_0 + \int_0^\infty |\mathbf{\Psi}_\infty x_0(\tau)|\, d\tau, \quad \beta \ge \sup_{t \ge 0} |r - (\mathcal{L}^{-1}(\mathbf{G}) \star \phi(u))(t)|.$$

Therefore, assertions (a) to (e) of Theorem 3.8 hold.

Case 2. In this case, $k(t) \to k^* := 1/l^* > 0$ as $t \to \infty$. By boundedness of $l(\cdot)$ and (36), we may conclude that $e(\cdot) := r - C_L x(\cdot) - D\phi(u(\cdot)) \in L^1(\mathbb{R}_+)$ and so (by (10b)) $u(t)$ converges to a finite limit as $t \to \infty$. Consequently, $\phi(u(t))$ converges to a finite limit as $t \to \infty$, and hence, by Lemma 3.7, assertions (a) to (e) of Theorem 3.8 hold.

Finally, assume that $\phi_r \in \operatorname{im} \phi$ is not a critical value and that $\mathcal{L}^{-1}(\mathbf{G}) \in \mathcal{M}_\alpha$ for some $\alpha < 0$. We will show that the monotone gain k converges to a positive value. Seeking a contradiction, suppose that the monotone function l is unbounded (equivalently, $k(t) \downarrow 0$ as $t \to \infty$). Then the hypotheses of Theorem 3.6 are satisfied and so (38) holds. By Theorem 3.8, $\phi(u(\cdot))$ converges to ϕ_r, and $y(\cdot) - (\mathbf{\Psi}_\infty x_0)(\cdot)$

converges to r; moreover, the convergence is of order $\exp(-\rho K(t))$ for some $\rho > 0$; that is, there exists constant $L > 0$ such that

$$|r - y(t) + (\mathbf{\Psi}_\infty x_0)(t)| \leq L \exp(-\rho K(t)) \quad \forall\, t \in \mathbb{R}_+. \tag{39}$$

Choose $\gamma \geq \beta$ such that $\rho/\gamma < 1$. By (38), $k(t) = 1/l(t) \geq (\alpha + \gamma t)^{-1}\ \forall\ t \in \mathbb{R}_+$. Therefore,

$$K(t) = \int_0^t k \ \geq\ \ln[((\alpha + \gamma t)/\alpha)^{1/\gamma}] \quad \forall\, t \geq 0.$$

Consequently for a.a. $t \geq 0$,

$$\dot{l}(t) = |r - y(t)| \leq L \exp(-\rho K(t)) + |(\mathbf{\Psi}_\infty x_0)(t)| \leq M(\alpha + \gamma t)^{-\eta} + |(\mathbf{\Psi}_\infty x_0)(t)|,$$

where $\eta = \rho/\gamma \in (0, 1)$ and $M = L\alpha^\eta$. Since, by exponential stability, $\mathbf{\Psi}_\infty x_0 \in L^1(\mathbb{R}_+, \mathbb{R})$, integration gives

$$l(t) \leq N(\alpha + \gamma t)^{1-\eta} \quad \forall\, t \geq 0,$$

for some suitable constant $N > 0$. It follows that

$$-K(t) = -\int_0^t k \leq -(N\gamma\eta)^{-1} \left[(\alpha + \gamma t)^\eta - \alpha^\eta\right] \quad \forall\, t \geq 0.$$

Therefore, $\exp(-\rho K(\cdot))$ is of class $L^1(\mathbb{R}_+, \mathbb{R})$, and, by (39), it follows that $|r - y(\cdot) + (\mathbf{\Psi}_\infty x_0)(\cdot)|$ is also of class $L^1(\mathbb{R}_+, \mathbb{R})$. Since $\mathbf{\Psi}_\infty x_0 \in L^1(\mathbb{R}_+, \mathbb{R})$, we have $|r - y(\cdot)| \in L^1(\mathbb{R}_+, \mathbb{R})$. This contradicts the supposition of unboundedness of $l(\cdot)$. Therefore, $l(\cdot)$ is bounded. □

4. Example: Controlled diffusion process with output delay. Consider a diffusion process (with diffusion coefficient $a > 0$ and with Dirichlet boundary conditions), on the one-dimensional spatial domain $I = [0, 1]$, with scalar nonlinear pointwise control action (applied at point $x_b \in I$, via a nonlinearity ϕ with Lipschitz constant $\lambda > 0$) and delayed (delay $h \geq 0$) pointwise scalar observation (output at point $x_c \in I$). We formally write this single-input, single-output system (previously considered, in a nonadaptive control context, in the precursor [9] to the present paper) as

$$z_t(t, x) = a z_{xx}(t, x) + \delta(x - x_b)\phi(u(t)), \quad y(t) = z(t - h, x_c),$$

$$z(t, 0) = 0 = z(t, 1) \quad \forall\, t > 0.$$

For simplicity, we assume zero initial conditions:

$$z(t, x) = 0 \quad \forall\ (t, x) \in [-h, 0] \times [0, 1].$$

With input $\phi(u(\cdot))$ and output $y(\cdot)$, this example qualifies as a regular linear system with transfer function given by

$$\mathbf{G}(s) = \frac{e^{-sh/a} \sinh(x_b\sqrt{s}\,) \sinh((1 - x_c)\sqrt{s}\,)}{a\sqrt{s} \sinh\sqrt{s}}.$$

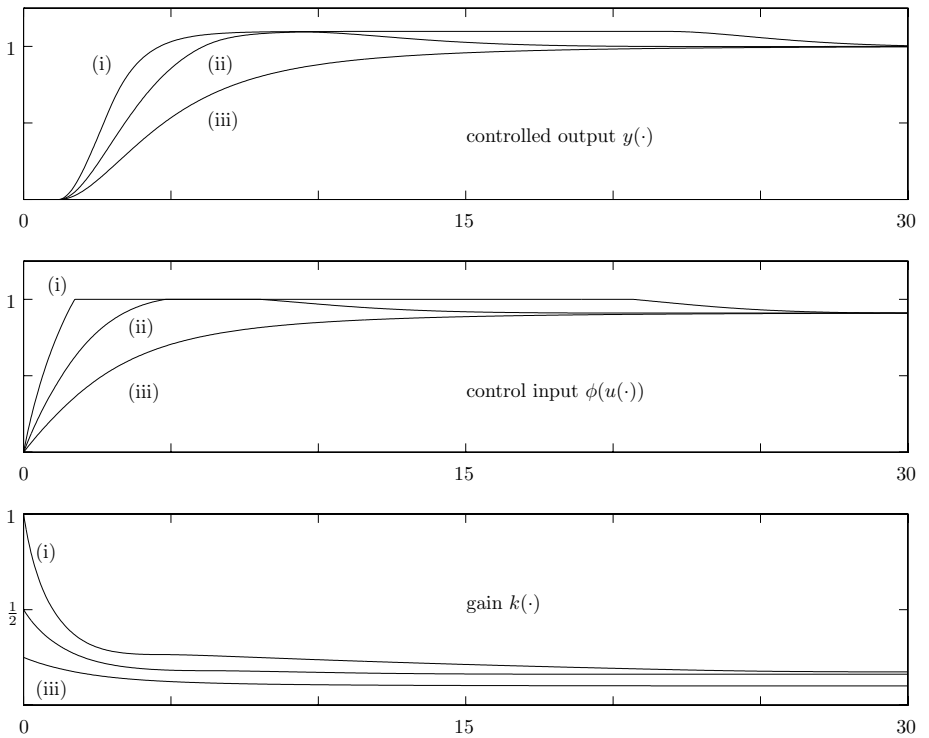


FIG. 3. *Controlled output, control input, and adapting gain.*

Therefore, by Theorem 3.8, the adaptive integral control

$$u(t) = \int_0^t k(t)[r - y(t)]\,dt, \qquad k(t) = \frac{1}{l(t)},$$

where the evolution of $l(t)$ is given by the adaptation law

$$\dot{l}(t) = |r - y(t)|, \quad l(0) = l_0 > 0,$$

guarantees asymptotic tracking of every constant reference signal r satisfying

$$\frac{r}{\mathbf{G}(0)} = \frac{ar}{x_b(1 - x_c)} \in \operatorname{clos}(\operatorname{im}\phi).$$

For purposes of illustration, we adopt the following values:

$$a = 0.1, \quad x_b = \frac{1}{3}, \quad x_c = \frac{2}{3}, \quad h = 0.1.$$

We consider a nonlinearity ϕ of saturation type, defined as follows

$$u \mapsto \phi(u) := \begin{cases} 1, & u \geq 1, \\ u, & u \in (0, 1), \\ 0, & u \leq 0, \end{cases}$$

in which case $\lambda = 1$. For unit reference signal $r = 1$, we have

$$\frac{r}{\mathbf{G}(0)} = \frac{a}{x_b(1 - x_c)} = 0.9 \in \operatorname{int}(\operatorname{im}\phi).$$

Figure 3 depicts the behavior of the system (with reference $r = 1$) under adaptive integral control in each of the following three cases:

$$\text{(i) } l_0 = 1, \quad \text{(ii) } l_0 = 2, \quad \text{(iii) } l_0 = 4.$$

This figure was generated using SIMULINK simulation software within MATLAB wherein a truncated eigenfunction expansion, of order 10, was adopted to model the diffusion process.

5. Appendix: Proof of Lemmas 3.5 and 3.12. In proving Lemmas 3.5 and 3.12, we will first study an abstract Volterra integrodifferential equation. Let $\alpha \geq 0$, and let $w_\alpha \in C([0,\alpha],\mathbb{R}^n)$. Consider the initial-value problem

$$\dot{w}(t) = (Vw)(t), \quad t \geq \alpha, \tag{40a}$$

$$w(t) = w_\alpha(t), \quad t \in [0,\alpha], \tag{40b}$$

where the operator $V : C(\mathbb{R}_+,\mathbb{R}^n) \to L^1_{\text{loc}}(\mathbb{R}_+,\mathbb{R}^n)$ is causal and *weakly Lipschitz* in the following sense.

$\forall\ \alpha \geq 0$, $\delta > 0$, $\rho > 0$, and $\theta \in C([0,\alpha],\mathbb{R}^n)$, there exists a continuous function $f : [0,\delta] \to \mathbb{R}_+$, with $f(0) = 0$, such that

$$\int_\alpha^{\alpha+\varepsilon} \|(Vv)(t) - (Vw)(t)\|\, dt \leq f(\varepsilon) \sup_{\alpha \leq t \leq \alpha+\varepsilon} \|v(t) - w(t)\|$$

$\forall\ \varepsilon \in [0,\delta]$ and $\forall\ v, w \in \mathcal{C}(\alpha,\delta,\rho,\theta)$, where

$$\mathcal{C}(\alpha,\delta,\rho,\theta) := \{w \in C([0,\alpha+\delta],\mathbb{R}^n) |\ w(t) = \theta(t)\ \forall\ t \in [0,\alpha], \\ \|w(t) - \theta(\alpha)\| \leq \rho\ \forall\ t \in [\alpha,\alpha+\delta]\}.$$

A *solution* of the initial-value problem (40) on an interval $[0,\beta)$, where $\beta > \alpha$, is a function $w \in C([0,\beta),\mathbb{R}^n)$, with $w(t) = w_\alpha(t)\ \forall\ t \in [0,\alpha]$, such that w is absolutely continuous on $[\alpha,\beta)$ and (40a) is satisfied for a.a. $t \in [\alpha,\beta)$.

Strictly speaking, to make sense of (40), we have to give a meaning to $(Vw)(t)$, $t \in [0,\beta)$, when w is a continuous function defined on a *finite* interval $[0,\beta)$ (recall that V operates on the space of continuous functions defined on the *infinite* interval $\mathbb{R}_+$). This can be done easily using causality of V: $\forall\ t \in [0,\beta)$, $(Vw)(t) := (Vw^*)(t)$, where $w^* : \mathbb{R}_+ \to \mathbb{R}^n$ is any continuous function with $w^*(s) = w(s)\ \forall\ s \in [0,t]$.

PROPOSITION 5.1. *For every $\alpha \geq 0$ and every $w_\alpha \in C([0,\alpha],\mathbb{R}^n)$, there exists a unique solution $w(\cdot)$ of (40) defined on a maximal interval $[0,t_{\max})$, with $t_{\max} > \alpha$. Moreover, if $t_{\max} < \infty$, then*

$$\limsup_{t \to t_{\max}} |w(t)| = \infty. \tag{41}$$

Proof. Fix $\alpha \geq 0$, $w_\alpha \in C([0,\alpha],\mathbb{R}^n)$ arbitrarily. Define a continuous extension $w^*_\alpha : \mathbb{R}_+ \to \mathbb{R}^n$ of w_α by setting $w^*_\alpha(t) = w_\alpha(\alpha)\ \forall\ t > \alpha$. For later convenience, we introduce the continuous function

$$\varepsilon \mapsto g(\varepsilon) := \int_\alpha^{\alpha+\varepsilon} \|(Vw^*_\alpha)(t)\|\, dt.$$

We proceed in three steps.

Step 1. Existence and uniqueness on a small interval.

For each $\varepsilon \in (0,1)$, define

$$\mathcal{C}_\varepsilon := \mathcal{C}(\alpha, \varepsilon, 1, w_\alpha),$$

which, endowed with the metric

$$(v,w) \mapsto \sup_{\alpha \le t \le \alpha+\varepsilon} \|v(t) - w(t)\|,$$

is a complete metric space.

Existence and uniqueness of a solution on a small interval is proved by showing that

$$(\Gamma w)(t) := \begin{cases} w_\alpha(t), & 0 \le t \le \alpha, \\ w_\alpha(\alpha) + \displaystyle\int_\alpha^t (Vw)(\tau)\, d\tau, & \alpha \le t \le \alpha + \varepsilon \end{cases}$$

defines a contraction on $\mathcal{C}_\varepsilon$ for sufficiently small $\varepsilon > 0$.

By the weak Lipschitz property of V, there exists a continuous function $f : [0,1] \to \mathbb{R}_+$ with $f(0) = 0$, such that, $\forall\ \varepsilon \in (0,1)$, $v, w \in \mathcal{C}_\varepsilon$, and $t \in [\alpha, \alpha+\varepsilon]$,

$$\begin{aligned} \|(\Gamma w)(t) - w_\alpha(\alpha)\| &\le \int_\alpha^{\alpha+\varepsilon} \|(Vw)(\tau)\|\, d\tau \\ &\le g(\varepsilon) + \int_\alpha^{\alpha+\varepsilon} \|(Vw)(\tau) - (Vw^*_\alpha)(\tau)\|\, d\tau \\ &\le g(\varepsilon) + f(\varepsilon) \\ &\le 1 \qquad \text{for all sufficiently small } \varepsilon > 0 \end{aligned} \tag{42}$$

and

$$\begin{aligned} \|(\Gamma v)(t) - (\Gamma w)(t)\| &\le \int_\alpha^{\alpha+\varepsilon} \|(Vv)(\tau) - (Vw)(\tau)\|\, d\tau \\ &\le f(\varepsilon) \sup_{\alpha \le \tau \le \alpha+\varepsilon} \|v(\tau) - w(\tau)\| \\ &\le \tfrac{1}{2} \sup_{\alpha \le \tau \le \alpha+\varepsilon} \|v(\tau) - w(\tau)\| \quad \text{for all sufficiently small } \varepsilon > 0. \end{aligned} \tag{43}$$

By (42), $\Gamma(\mathcal{C}_\varepsilon) \subset \mathcal{C}_\varepsilon$ for all sufficiently small $\varepsilon > 0$. Consequently, we obtain from (43) that Γ is a contraction on $\mathcal{C}_\varepsilon$ for all sufficiently small $\varepsilon > 0$.

Step 2. Extended uniqueness.

Let $v : [0, \beta_1) \to \mathbb{R}^n$ and $w : [0, \beta_2) \to \mathbb{R}^n$, $\beta_1, \beta_2 > \alpha$, be solutions of (40) (existence of v and w is ensured by Step 1).

We claim that $v(t) = w(t)\ \forall\ t \in [0, \beta)$, where $\beta = \min\{\beta_1, \beta_2\}$. Seeking a contradiction, suppose that there exists $t \in (0, \beta)$ such that $v(t) \ne w(t)$. Defining

$$t^* = \inf\{t \in (0,\beta) \,|\, v(t) \ne w(t)\},$$

it follows that $t^* > \alpha$ (by Step 1), $t^* < \beta$ (by supposition), and $v(t^*) = w(t^*)$ (by continuity of v and w). Clearly, the initial-value problem

$$\dot{z}(t) = (Vz)(t),\ \ t \ge t^*\,; \quad z(t) = v(t),\ \ t \in [0, t^*]$$

is solved by v and w. This implies (by the argument in Step 1) that there exists an $\varepsilon > 0$ such that $v(t) = w(t)\ \forall\ t \in [0, t^* + \varepsilon)$, which contradicts the definition of t^*.

Step 3. Continuation of solutions.

Let $\alpha \geq 0$ and $w_\alpha \in C([0,\alpha], \mathbb{R}^n)$ be arbitrary and, as before, let w_α^* be the continuous extension of w_α with $w_\alpha^*(t) = w_\alpha(\alpha)\ \forall\ t > \alpha$.

Let w be a solution of (40) on the interval $[0,\beta)$, $\alpha < \beta < \infty$. In order to prove that w can be extended to a maximal solution (which satisfies (41) if $t_{\max} < \infty$), it is sufficient to show that w can be continued to the right (beyond β) if w is bounded on $[0,\beta)$. Suppose that w is bounded. Set $\delta := \beta - \alpha$ and $\rho =: \sup\{\|w(\tau) - w_\alpha(\alpha)\| \mid \alpha < \tau < \beta\}$. By the weak Lipschitz property of V, there exists continuous $f : [0,\delta] \to \mathbb{R}_+$, with $f(0) = 0$, such that $\forall\ \varepsilon \in (0,\delta)$

$$\int_\alpha^{\alpha+\varepsilon} \|(Vw)(\tau)\|\, d\tau \leq g(\varepsilon) + \rho f(\varepsilon),$$

implying, by boundedness of g and f on $[0,\delta]$, that $Vw \in L^1([0,\beta], \mathbb{R}^n)$ and so the following limit exists:

$$\lim_{t\uparrow\beta} \int_\alpha^t (Vw)(\tau)\, d\tau =: L \in \mathbb{R}^n,$$

whence

$$L + w_\alpha(\alpha) = \lim_{t\uparrow\beta} (\Gamma w)(t).$$

Now $w(t) = (\Gamma w)(t)\ \forall\ t \in [0,\beta)$. Therefore, defining $w(\beta) = L + w_\alpha(\alpha)$ we can extend w into a continuous function on $[0,\beta]$. Finally, by the argument in Step 1, the initial-value problem

$$\dot z(t) = (Vz)(t),\ \ t \geq \beta\,; \quad z(t) = w(t),\ \ t \in [0,\beta]$$

has a unique solution w^* on $[0,\beta+\varepsilon)$ for some $\varepsilon > 0$. By causality of V, the function w^* is a solution of (40) on $[0,\beta+\varepsilon)$, and so w^* is a proper right continuation of w. □

REMARK 5.2. *In what follows, we shall invoke Proposition* 5.1 *only in the special case* $\alpha = 0$. *Note, however, that Steps* 2 *and* 3 *in the above proof of the proposition required the local existence and uniqueness result in the more general context of* $\alpha \geq 0$.

In the following, Proposition 5.1 will be used to prove Lemmas 3.5 and 3.12. First note that, by setting $k(t) = 1/l(t)$, the adaptive feedback system (37) (with $(A,B,C,D) \in \mathcal{R}$) can be written in the following form which will be more convenient for our purposes:

$$\dot x(t) = Ax(t) + B\phi(u(t)), \quad x(0) = x_0 \in X, \tag{44a}$$

$$\dot u(t) = k(t)[r - C_L x(t) - D\phi(u(t))], \quad u(0) = u_0 \in \mathbb{R}, \tag{44b}$$

$$\dot k(t) = -k^2(t)|r - C_L x(t) - D\phi(u(t))|, \quad k(0) = k_0 \in (0,\infty). \tag{44c}$$

The feedback systems (10) and (44) are both of the form

$$\dot x(t) = Ax(t) + B\phi(u(t)), \quad x(0) = x_0 \in X \tag{45a}$$

$$\dot u(t) = \kappa(t)\theta(t)[r - C_L x(t) - D\phi(u(t))], \quad u(0) = u_0 \in \mathbb{R}, \tag{45b}$$

$$\dot\theta(t) = h(\theta(t))|r - C_L x(t) - D\phi(u(t))|, \quad \theta(0) = \theta_0 \in \mathbb{R}, \tag{45c}$$

where $\kappa \in L^\infty(\mathbb{R}_+, \mathbb{R})$ and $h : \mathbb{R} \to \mathbb{R}$ is locally Lipschitz. To recover (10) from (45), set $h(\theta) \equiv 0$ and $\theta_0 = 1$ (in this case $\kappa(\cdot)$ plays the role of the gain function $k(\cdot)$). Considering the special case $\kappa(t) \equiv 1$ and $h(\theta) = -\theta^2$ gives the adaptive feedback equations (44) (with $k(\cdot) = \theta(\cdot)$).

For $a \in (0, \infty]$, a continuous function

$$[0, a) \to X \times \mathbb{R} \times \mathbb{R}, \quad t \mapsto (x(t), u(t), \theta(t))$$

is a *solution* of (45) if $(x(\cdot), u(\cdot), \theta(\cdot))$ is absolutely continuous as a $(X_{-1} \times \mathbb{R} \times \mathbb{R})$-valued function, $x(t) \in \mathrm{dom}(\mathrm{C_L})$ for a.a. $t \in [0, a)$, $(x(0), u(0), \theta(0)) = (x_0, u_0, \theta_0)$, and the differential equations in (45) are satisfied almost everywhere on $[0, a)$, where the derivative in (45a) should be interpreted in the space X_{-1}.

On noting that $C_L x(t) + D\phi(u(t)) = (\mathbf{\Psi}_\infty x_0)(t) + (\mathbf{F}_\infty \phi(u))(t)$ (with $\mathbf{\Psi}_\infty$ and $\mathbf{F}_\infty$ defined by (2)), the variable $x(t)$ can be eliminated from (45b) and (45c) to obtain

$$\dot{u}(t) = \kappa(t)\theta(t)[r - (\mathbf{\Psi}_\infty x_0)(t) - (\mathbf{F}_\infty \phi(u))(t)], \quad u(0) = u_0, \tag{46a}$$

$$\dot{\theta}(t) = h(\theta(t))|r - (\mathbf{\Psi}_\infty x_0)(t) - (\mathbf{F}_\infty \phi(u))(t)|, \quad \theta(0) = \theta_0. \tag{46b}$$

In order to proceed we need the following lemma.

LEMMA 5.3. *$\forall\ \alpha \geq 0$, $v \in C([0, \alpha], \mathbb{R})$, $\delta > 0$, and $\rho > 0$, there exist $\gamma_1, \gamma_2 > 0$ such that $\forall\ \varepsilon \in [0, \delta]$ and $u, w \in \mathcal{C}(\alpha, \delta, \rho, v)$*

$$\int_\alpha^{\alpha+\varepsilon} |(\mathbf{F}_\infty \phi(u))(\tau) - (\mathbf{F}_\infty \phi(w))(\tau)|\, d\tau \leq \varepsilon\gamma_1 \sup_{\alpha \leq \tau \leq \alpha+\varepsilon} |u(\tau) - w(\tau)|, \tag{47}$$

$$\int_\alpha^{\alpha+\varepsilon} |(\mathbf{F}_\infty \phi(u))(\tau)|\, d\tau \leq \varepsilon\gamma_1\rho + \sqrt{\varepsilon}\gamma_2. \tag{48}$$

Proof. Let $\alpha \geq 0$, $v \in C([0, \alpha], \mathbb{R})$, $\delta > 0$, $\rho > 0$, and $u, w \in \mathcal{C}(\alpha, \delta, \rho, v)$. Let λ be a Lipschitz constant for $\phi \in \mathcal{N}$. Then, using the Cauchy–Schwarz inequality and the boundedness of $\mathbf{F}_\infty$ as an operator from $L^2(\mathbb{R}_+, \mathbb{R})$ into $L^2(\mathbb{R}_+, \mathbb{R})$, we obtain $\forall\ \varepsilon \in [0, \delta]$,

$$\begin{aligned}
\int_\alpha^{\alpha+\varepsilon} |\mathbf{F}_\infty \phi(u) - \mathbf{F}_\infty \phi(w)| &\leq \sqrt{\varepsilon} \left(\int_\alpha^{\alpha+\varepsilon} |\mathbf{F}_\infty \phi(u) - \mathbf{F}_\infty \phi(w)|^2 \right)^{1/2} \\
&\leq \sqrt{\varepsilon}\lambda \|\mathbf{F}_\infty\| \left(\int_\alpha^{\alpha+\varepsilon} |u - w|^2 \right)^{1/2} \\
&\leq \varepsilon\lambda \|\mathbf{F}_\infty\| \sup_{\alpha \leq \tau \leq \alpha+\varepsilon} |u(\tau) - w(\tau)|,
\end{aligned}$$

which is (47) with $\gamma_1 = \lambda\|\mathbf{F}_\infty\|$.

To establish (48), define a continuous extension $v^* : \mathbb{R}_+ \to \mathbb{R}$ of v by setting $v^*(t) = v(\alpha)\ \forall\ t > \alpha$. Applying (47), it follows $\forall\ \varepsilon \in [0, \delta]$ that

$$\begin{aligned}
\int_\alpha^{\alpha+\varepsilon} |(\mathbf{F}_\infty \phi(u))(\tau)|\, d\tau &\leq \int_\alpha^{\alpha+\varepsilon} |(\mathbf{F}_\infty \phi(v^*))(\tau)|\, d\tau + \varepsilon\gamma_1 \sup_{\alpha \leq \tau \leq \alpha+\varepsilon} |u(\tau) - v^*(\tau)| \\
&\leq \sqrt{\varepsilon} \left(\int_\alpha^{\alpha+\delta} |(\mathbf{F}_\infty \phi(v^*))(\tau)|^2 \right)^{1/2} d\tau + \varepsilon\gamma_1\rho,
\end{aligned}$$

which yields (48) with $\gamma_2 := (\int_\alpha^{\alpha+\delta} |(\mathbf{F}_\infty\phi(v^*))(\tau)|^2\,d\tau)^{1/2}$. □

Lemmas 3.5 and 3.12 are special cases of the following corollary.

COROLLARY 5.4. *Let $(A,B,C,D) \in \mathcal{R}$, $\phi \in \mathcal{N}$, $r \in \mathbb{R}$, $\kappa \in L^\infty(\mathbb{R}_+,\mathbb{R})$, and let $h : \mathbb{R} \to \mathbb{R}$ be locally Lipschitz. If $h(\theta) \le 0\ \forall\ \theta \in \mathbb{R}$ and $h(0) = 0$, then $\forall$ $(x_0,u_0,\theta_0) \in X \times \mathbb{R} \times (0,\infty)$, the initial-value problem given by* (45) *has a unique solution defined on $\mathbb{R}_+$.*

Proof. Let $(x_0,u_0,\theta_0) \in X\times\mathbb{R}\times(0,\infty)$. It is clear that the map $V : C(\mathbb{R}_+,\mathbb{R}^2) \to L^1_{\mathrm{loc}}(\mathbb{R}_+,\mathbb{R}^2)$ given by

$$V\begin{pmatrix}u\\ \theta\end{pmatrix}(t) = \begin{pmatrix}\kappa(t)\theta(t)[r - (\mathbf{\Psi}_\infty x_0)(t) - (\mathbf{F}_\infty\phi(u))(t)]\\ h(\theta(t))|r - (\mathbf{\Psi}_\infty x_0)(t) - (\mathbf{F}_\infty\phi(u))(t)|\end{pmatrix}$$

is causal, and it follows from Lemma 5.3 via a routine argument that it is also weakly Lipschitz. Hence it follows from Proposition 5.1 that the initial-value problem (46) has a unique solution (u,θ) on a maximal interval of existence $[0,t_{\max})$. To prove that $t_{\max} = \infty$, we first show that θ is bounded on $[0,t_{\max})$. Note that since $h \le 0$, $\theta(\cdot)$ is nonincreasing, and combining this with the assumption that $\theta_0 > 0$, we see that boundeness of $\theta(\cdot)$ follows if we can show that $\theta(t) > 0\ \forall\ t \in [0,t_{\max})$. Seeking a contradiction, suppose that there exists a $t^* \in (0,t_{\max})$ such that $\theta(t^*) = 0$. Consider the following initial-value problem on $[0,t_{\max})$:

$$\dot\zeta(t) = h(\zeta(t))|e(t)|, \quad \zeta(t^*) = 0, \tag{49}$$

where $e(t) = r - (\mathbf{\Psi}_\infty x_0)(t) - (\mathbf{F}_\infty\phi(u))(t)$. Then $\theta(\cdot)$ is a solution of (49). Since $h(0) = 0$, the function $\zeta \equiv 0$ is also a solution of (49). By uniqueness it follows that $\theta \equiv 0$, which is in contradiction to $\theta_0 > 0$. Therefore, the function $\theta(\cdot)$ is bounded on $[0,t_{\max})$ and hence there exists a constant $\gamma > 0$ such that

$$|\kappa(t)\theta(t)| \le \gamma \quad \forall\, t \in [0,t_{\max}).$$

Let $[0,T)$ be an arbitrary interval with $[0,T) \subset [0,t_{\max})$ and $T < \infty$. Multiplying (46a) by u and estimating we obtain that, $\forall\ \tau \in [0,T)$,

$$u(\tau)\dot u(\tau) \le \gamma[r^2 + (\mathbf{\Psi}_\infty x_0)^2(\tau) + u^2(\tau) + |(\mathbf{F}_\infty\phi(u))(\tau)u(\tau)|\,]. \tag{50}$$

Integrating (50) from 0 to t and combining the estimate

$$\int_0^t |(\mathbf{F}_\infty\phi(u))u| \le \int_0^t |\mathbf{F}_\infty(\phi(u)-\phi(0))|\,|u| + \frac{1}{2}\left(\int_0^t (\mathbf{F}_\infty\phi(0))^2 + \int_0^t u^2\right),$$

the Cauchy–Schwarz inequality, and the global Lipschitz property of ϕ, we can show readily that there exist positive constants α and β such that, $\forall\ t \in [0,T)$,

$$u^2(t) \le \alpha + \beta\int_0^t u^2(\tau)\,d\tau.$$

An application of Gronwall's lemma then shows that $u^2(t) \le \alpha e^{\beta t}\ \forall\ t \in [0,T)$. Hence u is bounded on $[0,T)$. Since this holds $\forall\ T$ with $T \le t_{\max}$ and $T < \infty$, it follows by Proposition 5.1 that $t_{\max} = \infty$.

Finally, to obtain a solution of (45), *define*

$$x(t) = \mathbf{T}_t x_0 + \int_0^t \mathbf{T}_{t-\tau}B\phi(u(\tau))\,d\tau. \tag{51}$$

By well-posedness, x is a continuous X-valued function, and moreover, since A, considered as a generator on X_{-1}, is in $\mathcal{B}(X, X_{-1})$, the function $t \mapsto Ax(t)$ is a continuous X_{-1}-valued function. Consequently, by Pazy [18, Theorem 2.4, p. 107], we have that in X_{-1}

$$\dot{x}(t) = Ax(t) + B\phi(u(t)) \quad \forall\, t \in \mathbb{R}_+.$$

It follows that (x, u, θ) is the unique solution of (45) defined on $\mathbb{R}_+$. □

REFERENCES

[1] V. Barbu, *Analysis and Control of Nonlinear Infinite-Dimensional Systems*, Academic Press, Boston, 1993.

[2] F. H. Clarke, *Optimization and Nonsmooth Analysis*, Wiley, New York, 1983.

[3] P. A. Cook, *Controllers with universal tracking properties*, in Proceedings of the International IMA Conference on Control: Modelling, Computation, Information, Manchester, UK, 1992.

[4] G. W. M. Coppus, S. L. Sha, and R. K. Wood, *Robust multivariable control of a binary distillation column*, IEE Proceedings D, 130 (1983), pp. 201–208.

[5] E. J. Davison, *Multivariable tuning regulators: The feedforward and robust control of a general servomechanism problem*, IEEE Trans. Automat. Control, 21 (1976), pp. 35–47.

[6] N. Dunford and J. T. Schwartz, *Linear Operators, Part* I: *General Theory*, Wiley, New York, 1958.

[7] G. Gripenberg, S.-O. Londen, and O. Staffans, *Volterra Integral and Functional Equations*, Cambridge University Press, Cambridge, UK, 1990.

[8] H. Logemann, E. P. Ryan, and S. Townley, *Integral control of linear systems with actuator nonlinearities: Lower bounds for the maximal regulating gain*, IEEE Trans. Automat. Control, 44 (1999), pp. 1315–1319.

[9] H. Logemann, E. P. Ryan, and S. Townley, *Integral control of infinite-dimensional linear systems subject to input saturation*, SIAM J. Control Optim., 36 (1998), pp. 1940–1961.

[10] H. Logemann and S. Townley, *Discrete-time low-gain control of uncertain infinite-dimensional systems*, IEEE Trans. Automat. Control, 42 (1997), pp. 22–37.

[11] H. Logemann and S. Townley, *Low-gain control of uncertain regular linear systems*, SIAM J. Control Optim., 35 (1997), pp. 78–116.

[12] H. Logemann and S. Townley, *Adaptive integral control of time-delay systems*, IEE Proc. Control Theory Appl., 144 (1997), pp. 531–536.

[13] J. Lunze, *Robust Multivariable Feedback Control*, Prentice–Hall, London, 1988.

[14] J. Lunze, *Experimentelle Erprobung einer Einstellregel für PI-Mehrgrößenregler bei der Herstellung von Ammoniumnitrat-Harnstoff-Lösung*, Messen Steuern Regeln, 30 (1987), pp. 2–6.

[15] D. E. Miller and E. J. Davison, *An adaptive tracking problem with a control input constraint*, Automatica, 29 (1993), pp. 877–887.

[16] D. E. Miller and E. J. Davison, *The self-tuning robust servomechanism problem*, IEEE Trans. Automat. Control, 34 (1989), pp. 511–523.

[17] M. Morari, *Robust stability of systems with integral control*, IEEE Trans. Automat. Control, 30 (1985), pp. 574–577.

[18] A. Pazy, *Semigroups of Linear Operators and Applications to Partial Differential Equations*, Springer-Verlag, New York, 1983.

[19] D. Salamon, *Realization theory in Hilbert space*, Math. Systems Theory, 21 (1989), pp. 147–164.

[20] G. Weiss, *Transfer functions of regular linear systems, part* I: *Characterization of regularity*, Trans. Amer. Math. Soc., 342 (1994), pp. 827–854.

[21] G. Weiss, *Two conjectures on the admissibility of control operators*, in Control and Estimation of Distributed Parameter Systems, F. Kappel, K. Kunisch, and W. Schappacher, eds., Birkhäuser-Verlag, Basel, 1991, pp. 367–378.

[22] G. Weiss, *Admissibility of unbounded control operators*, SIAM J. Control Optim., 27 (1989), pp. 527–545.

[23] G. Weiss, *Admissible observation operators for linear semigroups*, Israel J. Math., 65 (1989), pp. 17–43.

[24] G. Weiss, *The representation of regular linear systems on Hilbert spaces*, in Distributed Parameter System, F. Kappel, K. Kunisch, and W. Schappacher, eds., Birkhäuser-Verlag, Basel, 1989, pp. 401–416.

SIAM J. CONTROL OPTIM.
Vol. 38, No. 4, pp. 1145–1170

EXACT BOUNDARY CONTROLLABILITY OF MAXWELL'S EQUATIONS IN HETEROGENEOUS MEDIA AND AN APPLICATION TO AN INVERSE SOURCE PROBLEM*

SERGE NICAISE†

Abstract. We examine the question of control of Maxwell's equations in a heterogeneous medium with a nonsmooth boundary by means of control currents on the boundary of that medium. This requires the introduction and analysis of some functions spaces. Some energy estimates are established which allow us to get the control results owing to the Hilbert uniqueness method. We finally give an application to an inverse source problem.

Key words. Maxwell's equations, interface problems, singularities, control, inverse problem

AMS subject classifications. 93C20, 35B37, 35A20

PII. S0363012998344373

1. Introduction. Recently the controllability and stabilization of Maxwell's equations gave rise to many works. To obtain the controllability results, four approaches are commonly used, namely the method of moments, the energy decay method of Russell, the skew symmetric operators method of Bensoussan and the Hilbert uniqueness method (HUM) of Lions [23]. The method of moments [31, 19] consists in reducing the control problem to the question of finding solutions to a countable set of moment problems related to some eigenfunctions and is only convenient for particular regions. The energy decay method [18, 1] uses the so-called Russell's "controllability via stability" principle. The skew symmetric operators method takes advantage of the purely imaginary point spectrum to obtain (sufficient) algebraic characterization of the exact controllability, algebraic conditions which are checked for the Maxwell's equations by the multiplier method [4]. The HUM is based on observation estimates (or energy estimates) of the adjoint problem. Such estimates are obtained either by the multiplier method [22, 20] or by microlocal analysis [24, 29, 30]. The above-mentioned papers require regularity properties of the solutions which are satisfied if the electric permittivity ε and the magnetic permeability μ are smooth (mainly for the sake of simplicity they assume that $\varepsilon = \mu = 1$) and the involved domain Ω is smooth. Our first goal is then to extend some of their results to the nonsmooth case (i.e., ε and μ piecewise constant and nonsmooth boundaries) using HUM.

The stabilization of Maxwell's equations with Ohm's law or with Silver–Müller absorbing boundary conditions is treated in [18, 28, 29, 30].

The relationship between exact controllability results and some inverse problems was underlined by Yamamoto [33] for the wave equation. For Maxwell's equations with constant permittivity and permeability, the determination of antennas from boundary measurements was obtained in [34]. Our second goal is to extend his result to the inhomogeneous case and to show reconstruction of antennas from boundary measurements.

Let us now define our framework and shortly describe our results. Let Ω be a

*Received by the editors September 9, 1998; accepted for publication (in revised form) May 18, 1999; published electronically April 4, 2000.
http://www.siam.org/journals/sicon/38-4/34437.html
†Université de Valenciennes et du Hainaut Cambrésis, LIMAV, Institut des Sciences et Techniques de Valenciennes, F-59313 - Valenciennes Cedex 9, France (snicaise@univ-valenciennes.fr).

bounded, simply connected domain with a Lipschitz boundary Γ. We suppose that Ω is occupied by an electromagnetic medium of electric permittivity ε and magnetic permeability μ, which are real-valued bounded functions and uniformly positive definite (i.e., there exist $\varepsilon_0, \mu_0 > 0$ such that $\varepsilon(x) \geq \varepsilon_0, \mu(x) \geq \mu_0$ for all $x \in \Omega$). Two particular cases will retain our attention: First is the case when ε and μ are smooth and constant near Γ and a smooth boundary Γ. Second is the case when ε and μ are piecewise constant with Lipschitz polyhedral subdomains, in the sense that we assume that there exists a partition $\mathcal{P}$ of Ω in a finite set of Lipschitz polyhedra $\Omega_1, \ldots, \Omega_J$ such that on each Ω_j, $\varepsilon = \varepsilon_j$ and $\mu = \mu_j$, where ε_j and μ_j are positive constants (throughout this paper a Lipschitz polyhedron is a bounded, simply connected Lipschitz domain with piecewise plane boundary).

For any $T > 0$, we denote by Q_T the cylinder $\Omega\times]0,T[$ and by $\Sigma_T = \Gamma\times]0,T[$ its lateral boundary. If Γ_0 is a fixed open part of Γ, we will write $\Sigma_{0T} = \Gamma_0\times]0,T[$. When no confusion is possible we drop the index T.

We now consider (nonstationary) Maxwell's equations:

$$
\left\{
\begin{array}{l}
\varepsilon \frac{\partial E}{\partial t} - \mathbf{curl}\, H = 0 \text{ in } Q_T, \\
\mu \frac{\partial H}{\partial t} + \mathbf{curl}\, E = 0 \text{ in } Q_T, \\
\operatorname{div}(\varepsilon E) = \operatorname{div}(\mu H) = 0 \text{ in } Q_T, \\
H \times \nu = J \text{ on } \Sigma_{0T}, \\
H \times \nu = 0 \text{ on } \Sigma_T \setminus \Sigma_{0T},\ E \cdot \nu = 0 \text{ on } \Sigma_T, \\
E(0) = E_0,\ H(0) = H_0 \text{ in } \Omega,
\end{array}
\right. \tag{1.1}
$$

where ν denotes the unit outer normal vector on Γ. This means that we suppose that the time evolution of the electric field E and magnetic field H is driven by an externally applied density of current J flowing tangentially on Γ_0.

In that paper we want to find sufficient conditions that guarantee that the next exact controllability problem is solvable: Given a time $T > 0$ and initial data $\{E_0, H_0\}$, find a surface density of current J in appropriate function spaces such that the solution of (1.1) satisfies

$$E(T) = H(T) = 0 \text{ in } \Omega.$$

As already mentioned, this exact controllability problem was already investigated in [1, 19, 20, 22, 24, 29, 30, 31] but in the smooth case. Our goal is then to extend some of their results to the inhomegeneous case and for nonsmooth boundaries, which requires some nontrivial adaptations. The first step is to introduce and analyze adapted function spaces. Existence results for equations like (1.1) with homogeneous boundary conditions follow from semigroup theory. Afterwards we shall attack the exact controllability problem by means of the HUM of Lions [23] as in [22]. In that case, the exact controllability problem is equivalent to the unique solvability of the adjoint problem of (1.1), which is obtained with the help of observation estimates. In the general situation, we introduce the notion of weak and strong observation estimates satisfied by Σ_0. In both cases, we conclude the exact controllability with different controls on Σ_0. When ε and μ are smooth and constant near Γ and Γ is smooth, the strong observation estimate can be deduced from [24, 30], when Σ_0 satisfies the geometrical control property. These authors actually use microlocal analysis extending the method which has been developed by Bardos, Lebeau, and Rauch in [3] for the wave equation. When ε and μ are piecewise constant with Lipschitz polyhedra Ω_j, using multiplier techniques, we prove that Σ satisfies the weak and strong observation

estimates under some assumptions ensuring regularity properties of some function spaces [9, 10] as well as a geometrical condition (namely the condition (6.17)) in order to avoid internal control (in the homogeneous case, the first assumptions reduce to the convexity of Ω, while the geometrical condition (6.17) disappears). In this setting, a nonglobal strong observation estimate (i.e., for $\Sigma_0 \neq \Sigma$) is probably available using microlocal analysis, but this requires more investigations (see [6, 7] in the case of the wave equation with Dirichlet boundary conditions in nonsmooth domains). Note that the multiplier technique has the advantage of allowing us to get an estimate of the minimal time T_0 for which the observation estimates hold. Consequently, we obtain an estimate of the minimal time of control. Remark that when ε and μ are in $C^1(\bar{\Omega})$ and Γ is smooth, the multiplier method allows us to show that Σ satisfies the weak and strong observation estimates under some monotonicity assumptions on ε and μ. Carleman's estimate for Maxwell's equations with ε and μ in $C^2(\bar{\Omega})$ has been recently proved in [35, 14]. Such an estimate yields controllability results but with another kind of boundary controls.

As an application of the above results, we finally solve an inverse source problem in the spirit of [33, 34], namely the determination and reconstruction of antennas from boundary measurements.

Note that we have not considered Maxwell's equations with Ohm's law (which consists in adding to the left-hand side of the first equation of (1.1) the term σE, when $\sigma \geq 0$ is the conductivity of the medium) because our method only yields a controllability result for σ small enough. Probably another choice of the boundary observation would be more judicious.

The schedule of the paper is the following one: In section 2, we introduce some function spaces and prove some useful properties. Section 3 is devoted to the well-posedness of the adjoint problem of (1.1), established using the theory of maximal dissipative operators. In section 4, we introduce the notion of observation estimates and deduce from them the exact controllability results using the HUM. Section 5 is devoted to the inverse source problem. Finally, in section 6, for ε and μ piecewise constant with Lipschitz polyhedral subdomains, we establish different observation estimates for Σ adapting to our setting the results of [22, section 3], and we give an estimate of the minimal time of control.

2. Functions spaces. For any $s \geq 0$, $H^s(\Omega)$ denotes the usual Sobolev space on Ω [17]. In what follows, $\mathcal{D}(\Omega)$ is the space of all C^∞ functions with compact support in Ω while $C^\infty(\bar{\Omega})$ is the space of restrictions to Ω of functions from $\mathcal{D}(\mathbf{R}^3)$.

Let us now introduce the following spaces (compare with [21, 22]):

$$J(\Omega, \varepsilon) = \{\chi \in L^2(\Omega)^3 | \operatorname{div}(\varepsilon\chi) = 0\}, \tag{2.1}$$

$$\hat{J}(\Omega, \varepsilon) = \{\chi \in J(\Omega, \varepsilon) | \chi \cdot \nu = 0 \text{ on } \Gamma\}. \tag{2.2}$$

According to the definition of the spaces $J(\Omega)$ and $\hat{J}(\Omega)$ in [21, 22] corresponding to ε constant, we first prove the following lemma.

LEMMA 2.1. *The space $J(\Omega, \varepsilon)$ is equal to the closure in $L^2(\Omega)^3$ of*

$$X = \{\varphi \in L^2(\Omega)^3 | \varepsilon\varphi \in C^\infty(\bar{\Omega}) \text{ and } \operatorname{div}(\varepsilon\varphi) = 0\}.$$

Similarly, the space $\hat{J}(\Omega, \varepsilon)$ is equal to the closure in $L^2(\Omega)^3$ of

$$\hat{X} = \{\varphi \in L^2(\Omega)^3 | \varepsilon\varphi \in \mathcal{D}(\Omega) \text{ and } \operatorname{div}(\varepsilon\varphi) = 0\}.$$

Proof. Let us first assume that $\varepsilon = 1$. In that case, the second density result is proved in Theorem I.2.8 of [16]. For the first one, let us fix $u \in J(\Omega, 1)$. Then by Theorem I.3.4 of [16], there exists $\psi_0 \in H^1(\Omega)^3$ such that

$$u = \mathbf{curl}\, \psi_0.$$

Since $C^\infty(\bar{\Omega})$ is dense in $H^1(\Omega)$ (see, for instance, Theorem 1.4.2.1 of [17]), there exists a sequence of $\psi_n \in C^\infty(\bar{\Omega})^3$ such that

$$\mathbf{curl}\, \psi_n \to \mathbf{curl}\, \psi_0 \text{ in } L^2(\Omega)^3, \text{ as } n \to \infty.$$

Consequently, $\mathbf{curl}\, \psi_n$ belongs to $C^\infty(\bar{\Omega})$, is divergence-free, and converges to u in $L^2(\Omega)^3$.

For an arbitrary ε, we simply use the equivalence $\varphi \in J(\Omega, \varepsilon)$ (resp., $\hat{J}(\Omega, \varepsilon)$) if and only if $\varepsilon\varphi \in J(\Omega, 1)$ (resp., $\hat{J}(\Omega, 1)$). ∎

For the different formulations of our Maxwell equation (1.1), we further need the following spaces:

$$J^1_\nu(\Omega, \mu) = \{\chi \in \hat{J}(\Omega, \mu) |\, \mathbf{curl}\, \chi \in L^2(\Omega)^3\}, \tag{2.3}$$

$$J^1_\tau(\Omega, \varepsilon) = \{\chi \in J(\Omega, \varepsilon) |\, \mathbf{curl}\, \chi \in L^2(\Omega)^3 \text{ and } \chi \times \nu = 0 \text{ on } \Gamma\}, \tag{2.4}$$

$$J^\star_\nu(\Omega, \varepsilon, \mu) = \{\chi \in J^1_\nu(\Omega, \mu) |\, \mathbf{curl}(\varepsilon^{-1}\, \mathbf{curl}\, \chi) \in L^2(\Omega)^3 \text{ and } \mathbf{curl}\, \chi \times \nu = 0 \text{ on } \Gamma\}, \tag{2.5}$$

$$J^\star_\tau(\Omega, \varepsilon, \mu) = \{\chi \in J^1_\tau(\Omega, \varepsilon) |\, \mathbf{curl}(\mu^{-1}\, \mathbf{curl}\, \chi) \in L^2(\Omega)^3 \text{ and } \mathbf{curl}\, \chi \cdot \nu = 0 \text{ on } \Gamma\}. \tag{2.6}$$

When ε and μ are constant in Ω and Ω has a $C^{1,1}$ boundary, the above spaces coincide with those introduced in [21, 22], owing to Theorems 2.9 and 2.12 of [2] and the results from [8, 9]. When $\varepsilon = \mu = 1$ and Ω is a polyhedral domain, these spaces no longer coincide, in general, with those of [21, 22] according to the results of [8, 9] (see also [5, 15]; note that if Ω is convex, then the spaces $J^1_\nu(\Omega, 1)$ and $J^1_\tau(\Omega, 1)$ are embedded into $H^1(\Omega)$). Indeed in these papers, it is shown that any function in the above spaces admits a decomposition into a regular part with the optimal regularity (H^1 for the first two spaces and H^2 for the last two ones) and a singular part induces by some singularities of the Laplace operator with Dirichlet and Neumann conditions in Ω. The same kind of results were recently extended to the case when ε and μ are piecewise constant with polyhedral subdomains in [10].

For further purposes, we need the following results.

THEOREM 2.2. *There exist two positive constants c_1, c_2 such that*

$$\|\chi\|_{J^1_\nu(\Omega,\mu)} \le c_1 \|\, \mathbf{curl}\, \chi\|_{L^2(\Omega)^3} \quad \forall \chi \in J^1_\nu(\Omega, \mu), \tag{2.7}$$

$$\|\chi\|_{J^1_\tau(\Omega,\varepsilon)} \le c_2 \|\, \mathbf{curl}\, \chi\|_{L^2(\Omega)^3} \quad \forall \chi \in J^1_\tau(\Omega, \varepsilon). \tag{2.8}$$

Proof. The proof is based on the compact embeddings of $J^1_\nu(\Omega, \mu)$ and $J^1_\tau(\Omega, \varepsilon)$ into $L^2(\Omega)^3$ [32]. ∎

LEMMA 2.3. *The space $J^1_\tau(\Omega, \varepsilon)$ is dense in $J(\Omega, \varepsilon)$, while $J^1_\nu(\Omega, \mu)$ is dense in $\hat{J}(\Omega, \mu)$.*

Proof. Endow $L^2(\Omega)^3$ with the inner product

$$(\chi, \varphi) = \int_\Omega \varepsilon\chi \cdot \varphi \, dx,$$

and let P be the orthogonal projection on $J(\Omega,\varepsilon)$ in $L^2(\Omega)^3$. As $\mathcal{D}(\Omega)$ is dense in $L^2(\Omega)$, the subspace $P\mathcal{D}(\Omega)^3$ is clearly dense in $J(\Omega,\varepsilon)$. Consequently the first density result will be proved if the inclusion

$$P\mathcal{D}(\Omega)^3 \subset J^1_\tau(\Omega,\varepsilon) \tag{2.9}$$

holds. Fix $\chi \in \mathcal{D}(\Omega)^3$; then for any $\varphi \in \mathcal{D}(\Omega)^3$, we have

$$\int_\Omega \mathbf{curl}(P\chi)\cdot\varphi\,dx = \int_\Omega \varepsilon P\chi\cdot(\varepsilon^{-1}\,\mathbf{curl}\,\varphi)\,dx.$$

As the function $\varepsilon^{-1}\,\mathbf{curl}\,\varphi$ belongs to $J(\Omega,\varepsilon)$, the usual property of the projection yields

$$\int_\Omega \mathbf{curl}(P\chi)\cdot\varphi\,dx = \int_\Omega \varepsilon\chi\cdot(\varepsilon^{-1}\,\mathbf{curl}\,\varphi)\,dx.$$

This last identity shows that

$$\mathbf{curl}(P\chi) = \mathbf{curl}\,\chi. \tag{2.10}$$

This implies that $\mathbf{curl}(P\chi)$ belongs to $L^2(\Omega)^3$, and by Green's formula [16, section I.2], we have

$$\begin{aligned}\int_\Omega \mathbf{curl}(P\chi)\cdot\varphi\,dx = &\int_\Omega \varepsilon(P\chi)\cdot(\varepsilon^{-1}\,\mathbf{curl}\,\varphi)\,dx\\ &+\langle(P\chi)\times\nu,\varphi\rangle,\forall\varphi\in H^1(\Omega)^3.\end{aligned}$$

As the function $\varepsilon^{-1}\,\mathbf{curl}\,\varphi$ still belongs to $J(\Omega,\varepsilon)$, we get

$$\begin{aligned}\int_\Omega \mathbf{curl}(P\chi)\cdot\varphi\,dx = &\int_\Omega \varepsilon\chi\cdot(\varepsilon^{-1}\,\mathbf{curl}\,\varphi)\,dx\\ &+\langle(P\chi)\times\nu,\varphi\rangle,\forall\varphi\in H^1(\Omega)^3.\end{aligned}$$

Owing to (2.10), we arrive at

$$\langle(P\chi)\times\nu,\varphi\rangle = 0 \quad \forall\varphi\in H^1(\Omega)^3,$$

which shows that

$$(P\chi)\times\nu = 0 \text{ on } \Gamma.$$

This property and (2.10) prove that the inclusion (2.9) holds.

The second density result is similarly proved by endowing $L^2(\Omega)^3$ with the inner product $\int_\Omega \mu\chi\cdot\varphi\,dx$ and considering the orthogonal projection on $\hat{J}(\Omega,\mu)$ in $L^2(\Omega)^3$. □

3. Weak and strong solutions of the adjoint Maxwell equation. The homogeneous adjoint problem to (1.1) is (see section 5 for a justification)

$$\left\{\begin{array}{l}\mu\frac{\partial\varphi}{\partial t} - \mathbf{curl}\,\psi = 0 \text{ in } Q,\\ \varepsilon\frac{\partial\psi}{\partial t} + \mathbf{curl}\,\varphi = 0 \text{ in } Q,\\ \mathrm{div}(\mu\varphi) = \mathrm{div}(\varepsilon\psi) = 0 \text{ in } Q,\\ \varphi\times\nu = 0, \psi\cdot\nu = 0 \text{ on } \Sigma,\\ \varphi(0) = \varphi_0,\ \psi(0) = \psi_0 \text{ in } \Omega.\end{array}\right. \tag{3.1}$$

Contrary to [22], where a vector potential is used leading to a second order evolution equation, we shall prove existence of (3.1) by keeping the first order system (compare with [13, 18]). For that reason, let us introduce the Hilbert space

$$H = J(\Omega,\mu) \times \hat{J}(\Omega,\varepsilon),$$

equipped with the inner product

$$\left(\begin{pmatrix} \varphi \\ \psi \end{pmatrix}, \begin{pmatrix} \varphi_1 \\ \psi_1 \end{pmatrix} \right)_H = \int_\Omega \{\mu\varphi\bar{\varphi}_1 + \varepsilon\psi\bar{\psi}_1\}\, dx.$$

Now define the operator A as

$$\begin{aligned} D(A) &= J^1_\tau(\Omega,\mu) \times J^1_\nu(\Omega,\varepsilon), \\ A \begin{pmatrix} \varphi \\ \psi \end{pmatrix} &= \begin{pmatrix} \mu^{-1}\,\mathbf{curl}\,\psi \\ -\varepsilon^{-1}\,\mathbf{curl}\,\varphi \end{pmatrix}. \end{aligned}$$

We then see that formally problem (3.1) is equivalent to

$$\left\{ \begin{array}{l} \frac{\partial \Phi}{\partial t} = A\Phi, \\ \Phi(0) = \Phi_0, \end{array} \right. \tag{3.2}$$

when $\Phi = \begin{pmatrix} \varphi \\ \psi \end{pmatrix}$ and $\Phi_0 = \begin{pmatrix} \varphi_0 \\ \psi_0 \end{pmatrix}$.

We shall prove that this problem (3.2) has a unique solution using Lumer–Phillips's theorem [27] by showing the following lemma.

LEMMA 3.1. *A and $-A$ are maximal dissipative operators.*

Proof. We start with the dissipativeness of $\pm A$; in other words we need to show that

$$\Re(A\Phi,\Phi)_H = 0 \quad \forall \Phi \in D(A). \tag{3.3}$$

With the above notation we have

$$(A\Phi,\Phi)_H = \int_\Omega \{\mathbf{curl}\,\psi\bar{\varphi} - \mathbf{curl}\,\varphi\bar{\psi}\}\, dx.$$

But the definition of $D(A)$ implies that φ belongs to $H_0(\mathbf{curl},\Omega)$, where (see [16, 2])

$$H_0(\mathbf{curl},\Omega) = \{\chi \in L^2(\Omega)^3 |\, \mathbf{curl}\,\chi \in L^2(\Omega)^3 \text{ and } \chi \times \nu = 0 \text{ on } \Gamma\}.$$

Moreover, by section I.2.3 of [16], the space $\mathcal{D}(\Omega)^3$ is dense in $H_0(\mathbf{curl},\Omega)$. Consequently, there exists a sequence φ_n of functions in $\mathcal{D}(\Omega)^3$ such that

$$\varphi_n \to \varphi \text{ in } H_0(\mathbf{curl},\Omega), \text{ as } n \to \infty.$$

Now applying Green's formula (I.2.22) of [16] to the couple (φ_n,ψ), we get

$$\int_\Omega \{\mathbf{curl}\,\psi\bar{\varphi}_n - \mathbf{curl}\,\bar{\varphi}_n\psi\}\, dx = 0.$$

Taking the limit on n, we arrive at the identity

$$\int_\Omega \mathbf{curl}\,\psi\bar{\varphi}\, dx = \overline{\int_\Omega \mathbf{curl}\,\varphi\bar{\psi}\, dx}.$$

The real part of this identity yields (3.3).

Let us now pass to the maximality. This means that for all $\begin{pmatrix} f \\ g \end{pmatrix}$ in H, we are looking for $\begin{pmatrix} \varphi \\ \psi \end{pmatrix}$ in $D(A)$ such that

$$(I \pm A) \begin{pmatrix} \varphi \\ \psi \end{pmatrix} = \begin{pmatrix} f \\ g \end{pmatrix}.$$

Equivalently, we have

$$\psi = g \pm \varepsilon^{-1} \mathbf{curl}\, \varphi \tag{3.4}$$

and

$$\varphi + \mu^{-1} \mathbf{curl}(\varepsilon^{-1} \mathbf{curl}\, \varphi) = f \mp \mu^{-1} \mathbf{curl}\, g.$$

This last problem has a unique solution φ in $J^1_\tau(\Omega, \mu)$ because its variational formulation is

$$\int_\Omega \{\varepsilon^{-1} \mathbf{curl}\, \varphi\, \mathbf{curl}\, \bar{\theta} + \mu \varphi \bar{\theta}\}\, dx = \int_\Omega \{\mu f \bar{\theta} \mp g\, \mathbf{curl}\, \bar{\theta}\}\, dx \quad \forall \theta \in J^1_\tau(\Omega, \mu). \tag{3.5}$$

This problem has a unique solution by the Lax–Milgram lemma because the bilinear form defined as the left-hand side is directly coercive on $J^1_\tau(\Omega, \mu)$.

It then remains to show that ψ given by (3.4) belongs to $J^1_\nu(\Omega, \varepsilon)$. By (3.5), we see that

$$\mathbf{curl}\, \psi = \pm \mu (f - \varphi),$$

which shows that $\mathbf{curl}\, \psi \in L^2(\Omega)^3$. The other properties of ψ follow from (3.4) and from the fact that $g \in \hat{J}(\Omega, \varepsilon)$ and $\varphi \in J^1_\tau(\Omega, \mu)$ (recall that the boundary condition $\varphi \times \nu = 0$ implies that $\mathbf{curl}\, \varphi \cdot \nu = 0$). □

Since Lemma 2.3 guarantees the density of $D(A)$ into H, by Lumer–Phillips's theorem (see, for instance, Theorem I.4.3 of [27]), we conclude that A generates a C_0-group of contraction $T(t)$. Therefore, we have the following existence result.

THEOREM 3.2. *For all $\Phi_0 \in H$, the problem* (3.2) *has a weak solution $\Phi \in C([0, \infty), H)$ given by $\Phi = T(t)\Phi_0$. If, moreover, $\Phi_0 \in D(A^k)$, with $k \in \mathbf{N}^*$, the problem* (3.2) *has a strong solution $\Phi \in C([0, \infty), D(A^k)) \cap C^1([0, \infty), D(A^{k-1}))$.*

In the particular case $k = 1$ and 2, our result is in accordance with those from [22] because we easily see that

$$D(A^2) = J^\star_\tau(\Omega, \mu, \varepsilon) \times J^\star_\nu(\Omega, \mu, \varepsilon).$$

To finish this section, we establish the conservation of energy for weak and strong solutions.

LEMMA 3.3. *If $\Phi = \begin{pmatrix} \varphi \\ \psi \end{pmatrix}$ is a weak solution of problem* (3.2) *(or equivalently* (3.1)*), define the energy at time t by*

$$E(t) = \frac{1}{2} \int_\Omega \{\mu |\varphi(x,t)|^2 + \varepsilon |\psi(x,t)|^2\}\, dx.$$

Then we have

$$E(t) = E(0) \quad \forall t \geq 0. \tag{3.6}$$

Proof. Since $D(A)$ is dense in H, it suffices to prove (3.6) for $\begin{pmatrix} \varphi_0 \\ \psi_0 \end{pmatrix}$ in $D(A)$. For such an initial datum, φ and ψ are differentiable and therefore

$$\begin{aligned} \frac{d}{dt}E(t) &= \Re \int_\Omega \left\{ \mu \frac{\partial \varphi}{\partial t}\bar{\varphi} + \varepsilon \frac{\partial \psi}{\partial t}\bar{\psi} \right\} dx \\ &= \Re \int_\Omega \{ \mathbf{curl}\, \psi \bar{\varphi} - \mathbf{curl}\, \varphi \bar{\psi} \}\, dx \\ &= \Re \left(A \begin{pmatrix} \varphi \\ \psi \end{pmatrix}, \begin{pmatrix} \varphi \\ \psi \end{pmatrix} \right)_H = 0, \end{aligned}$$

owing to (3.3). ∎

LEMMA 3.4. *If* $\Phi = \begin{pmatrix} \varphi \\ \psi \end{pmatrix}$ *is a strong solution of problem* (3.2) *(or equivalently* (3.1)*) with initial datum in* $D(A)$*, define the modified energy at time* t *by*

$$\tilde{E}(t) = \frac{1}{2}\int_\Omega \{\mu^{-1} |\, \mathbf{curl}\, \psi(x,t)|^2 + \varepsilon^{-1} |\, \mathbf{curl}\, \varphi(x,t)|^2\}\, dx.$$

Then we have

$$\tilde{E}(t) = \tilde{E}(0) \quad \forall t \geq 0. \tag{3.7}$$

Proof. The proof is a direct consequence of the fact that

$$\frac{d}{dt}\tilde{E}(t) = \Re \left(A \begin{pmatrix} \frac{\partial \varphi}{\partial t} \\ \frac{\partial \psi}{\partial t} \end{pmatrix}, \begin{pmatrix} \frac{\partial \varphi}{\partial t} \\ \frac{\partial \psi}{\partial t} \end{pmatrix} \right)_H. \qquad \blacksquare$$

Let us finally notice that the compact embeddings of $J^1_\tau(\Omega,\mu)$ and $J^1_\nu(\Omega,\varepsilon)$ into $L^2(\Omega)^3$ imply that $D(A)$ is also compactly embedded into H. This fact and the monotonicity of A guarantee that A has a discrete spectrum included in the imaginary axis and that the set of associated eigenvectors forms an orthonormal basis of H.

4. Exact controllability. We start with the following definition.

DEFINITION 4.1. *Let* Γ_0 *be an open part of* Γ *and* $T > 0$*. We say that* $\Sigma_0 = \Gamma_0 \times (0,T)$ *satisfies the* (ε,μ)*-strong observation estimate (in short* (ε,μ)*-SOE) if there exists* $C > 0$ *such that*

$$\tilde{E}(0) \leq C \int_{\Sigma_0} \varepsilon^{-1} |\, \mathbf{curl}\, \varphi|^2 d\sigma dt \tag{4.1}$$

for all solutions (φ,ψ) *of* (3.1)*.*

We say that Σ_0 *satisfies the* (ε,μ)*-weak observation estimate (in short* (ε,μ)*-WOE) if there exists* $C > 0$ *such that*

$$\tilde{E}(0) \leq C \int_{\Sigma_0} (\varepsilon^{-1} |\, \mathbf{curl}\, \varphi|^2 + \mu^{-1} |\, \mathbf{curl}\, \psi|^2) d\sigma dt \tag{4.2}$$

for all solutions (φ,ψ) *of* (3.1)*.*

We will see that the weak and strong observation estimates yield control results at time T. Here are two examples where they hold. If ε and μ are smooth and constant near Γ and if Γ is smooth, then combining the results from [3] and the techniques from [24, 30], we can define the rays of the principal symbol, the diffractive

points of Σ and finally we can say that Σ_0 satisfies the geometrical control property if any ray encounters a nondiffractive point of Σ_0 (see [3, 24, 30] for the details). The method developed in [24, 30] then allows us to show that if Σ_0 satisfies the geometrical control property, then Σ_0 satisfies the (ε, μ)-SOE. If ε and μ are piecewise constant with Lipschitz polyhedral subdomains, we shall give in section 6 sufficient conditions ensuring that Σ satisfies the weak and strong observation estimates; we further obtain an explicit upper bound on the constants C appearing in (4.1) and (4.2).

If Σ_0 satisfies the (ε, μ)-SOE, the expression

$$\|\{\varphi_0, \psi_0\}\|_{F_1} = \left(\int_{\Sigma_0} |\operatorname{\mathbf{curl}} \varphi|^2 d\sigma dt \right)^{1/2}, \tag{4.3}$$

defines a norm on $J^\star_\tau(\Omega, \mu, \varepsilon) \times J^\star_\nu(\Omega, \mu, \varepsilon)$. We then define F_1 as the closure of this space for the norm (4.3). From (4.1), the following algebraic and topological inclusion holds:

$$F_1 \subset J^1_\tau(\Omega, \mu) \times J^1_\nu(\Omega, \varepsilon).$$

Unfortunately, this space does not furnish boundary controls in $L^2(\Sigma_0)^3$. We then use the next argument inspired from [22].

LEMMA 4.2. *If Σ_0 satisfies the (ε, μ)-SOE, then*

$$E(0) \leq C \sup_{x \in \Gamma_0} \varepsilon(x) \int_{\Sigma_0} |\psi|^2 d\sigma dt \tag{4.4}$$

for all solutions (φ, ψ) of (3.1) with initial data satisfying

$$\varphi_0 \in J^1_\tau(\Omega, \mu), \psi_0 \in J^1_\nu(\Omega, \varepsilon),$$

where C is the constant appearing in (4.1).

Proof. Introduce the auxiliary functions

$$\theta = \int_0^t \psi(s) ds + \theta_0, \chi = \int_0^t \varphi(s) ds + \chi_0,$$

where θ_0, χ_0 are chosen such that

$$\theta_0 \in J^\star_\nu(\Omega, \mu, \varepsilon), \operatorname{\mathbf{curl}} \theta_0 = \mu \varphi_0, \tag{4.5}$$

$$\chi_0 \in J^\star_\tau(\Omega, \mu, \varepsilon), \operatorname{\mathbf{curl}} \chi_0 = -\varepsilon \psi_0, \tag{4.6}$$

whose existence follows from the next lemma. Since θ, χ are solutions of (3.1) with initial data θ_0, χ_0, applying the estimate (4.1), we obtain

$$\begin{aligned} E(0) &\leq C \int_{\Sigma_0} \varepsilon^{-1} |\operatorname{\mathbf{curl}} \chi|^2 d\sigma dt \\ &\leq C \sup_{x \in \Gamma_0} \varepsilon(x) \int_{\Sigma_0} |\psi|^2 d\sigma dt, \end{aligned}$$

because $\operatorname{\mathbf{curl}} \chi = -\varepsilon \psi$. ∎

LEMMA 4.3. *The mappings*

$$\begin{aligned} J^\star_\nu(\Omega, \mu, \varepsilon) &\to J^1_\tau(\Omega, \mu) : \theta_0 \to \mu^{-1} \operatorname{\mathbf{curl}} \theta_0, \\ J^\star_\tau(\Omega, \mu, \varepsilon) &\to J^1_\nu(\Omega, \varepsilon) : \chi_0 \to \varepsilon^{-1} \operatorname{\mathbf{curl}} \chi_0 \end{aligned}$$

are onto.

Proof. We first prove that the first mapping is onto. Let us fix $\varphi \in J^1_\tau(\Omega,\mu)$. By Lemma 3.1 of [10], there exists $\psi \in H^1(\Omega)$ such that

$$\begin{aligned} &\mathbf{curl}\,\psi = \mu\varphi \text{ in } \Omega,\\ &\psi\cdot\nu = 0 \text{ on } \Gamma. \end{aligned}$$

We now look for θ_0 in the form

$$\theta_0 = \psi + \nabla u,$$

where u has to be determined so that θ_0 belongs to $J^\star_\nu(\Omega,\mu,\varepsilon)$. We see that it holds if

$$\begin{aligned} &\operatorname{div}(\varepsilon\theta_0) = 0 \text{ in } \Omega,\\ &\theta_0\cdot\nu = 0 \text{ on } \Gamma. \end{aligned}$$

From the expression of θ_0, this is equivalent to

$$\begin{aligned} &\operatorname{div}(\varepsilon\nabla u) = -\operatorname{div}(\varepsilon\psi) \text{ in } \Omega,\\ &\frac{\partial u}{\partial\nu} = 0 \text{ on } \Gamma. \end{aligned}$$

Therefore, it remains to look for the solution $u \in H^1(\Omega)$ of the above problem, whose variational formulation is

$$\int_\Omega \varepsilon\nabla u\nabla v\,dx = -\int_\Omega \varepsilon\psi\cdot\nabla v\,dx \quad \forall v \in H^1(\Omega).$$

This last problem has a unique (up to an additive constant) solution since the right-hand side is a continuous linear form on $H^1(\Omega)/\mathbf{R}$.

The proof of the surjectivity of the second mapping is similar, replacing Lemma 3.1 of [10] by Lemma 3.2 of [10]. □

Thanks to Lemma 4.2, if Σ_0 satisfies the (ε,μ)-SOE, the expression

$$\|\{\varphi_0,\psi_0\}\|_{F_2} = \left(\int_{\Sigma_0} |\psi|^2 d\sigma dt\right)^{1/2}, \tag{4.7}$$

defines a norm on $J^1_\tau(\Omega,\mu)\times J^1_\nu(\Omega,\varepsilon)$. If we define F_2 as the closure of this space for the norm (4.7), from (4.4), the following algebraic and topological inclusion holds:

$$F_2 \subset J(\Omega,\mu)\times\hat{J}(\Omega,\varepsilon).$$

If Σ_0 satisfies the (ε,μ)-WOE, we deduce from (4.2) (see [22] for the details) that

$$\|\{\varphi_0,\psi_0\}\|_{F_3} = \left(\int_{\Sigma_0} (|\,\mathbf{curl}\,\varphi|^2 + |\,\mathbf{curl}\,\psi|^2) d\sigma dt\right)^{1/2} \tag{4.8}$$

is a norm on $J^\star_\tau(\Omega,\mu,\varepsilon)\times J^\star_\nu(\Omega,\mu,\varepsilon)$. We then define F_3 as the closure of this space for the norm (4.8), and we have the following algebraic and topological inclusion:

$$F_3 \subset J^1_\tau(\Omega,\mu)\times J^1_\nu(\Omega,\varepsilon). \tag{4.9}$$

Following section 4 of [22], we deduce that the above observation estimates and the use of HUM allow us to prove exact controllability results for our Maxwell's equations. Here the main difficulty is to make precise the transposition method (the remainder is made exactly as in [22] and is therefore omitted).

Let us now assume that a solution $\{E, H\}$ of (1.1) exists such that

$$E \in C([0,T], J^1_\nu(\Omega,\varepsilon)) \cap C^1([0,T], \hat{J}(\Omega,\varepsilon)),$$
$$H \in C([0,T], J^1_\tau(\Omega,\mu)) \cap C^1([0,T], J(\Omega,\mu)),$$

which is the case if $E_0 \in J^1_\nu(\Omega,\varepsilon)$, $H_0 \in J^1_\tau(\Omega,\mu)$, and J is sufficiently regular.

Fix also a solution $\{\varphi, \psi\}$ of (3.1) with initial data $\{\varphi_0, \psi_0\}$ in $J^1_\tau(\Omega,\mu) \times J^1_\nu(\Omega,\varepsilon)$ with the regularity (thanks to Theorem 3.2)

$$\varphi \in C([0,T], J^1_\tau(\Omega,\mu)) \cap C^1([0,T], J(\Omega,\mu)),$$
$$\psi \in C([0,T], J^1_\nu(\Omega,\varepsilon)) \cap C^1([0,T], \hat{J}(\Omega,\varepsilon)).$$

Then we may write

$$0 = \int_0^t \int_\Omega [\varepsilon(E' - \varepsilon^{-1}\,\mathbf{curl}\, H)\cdot\psi - \mu(H' - \mu^{-1}\,\mathbf{curl}\, E)\cdot\varphi]\, dxds.$$

Applying integration by parts in s and Green's formula in x (allowed thanks to the assumptions (6.1)–(6.2)), we get

$$\int_\Omega (\mu H(t)\varphi(t) - \varepsilon E(t)\psi(t))dx = \int_\Omega (\mu H_0\varphi_0 - \varepsilon E_0\psi_0)dx$$
$$- \int_0^t \int_{\Gamma_0} J\cdot\psi d\sigma ds.$$

We rewrite this identity as

$$\langle\{H(t), -E(t)\}, \{\varphi(t), \psi(t)\}\rangle = \langle\{H_0, -E_0\}, \{\varphi_0, \psi_0\}\rangle - \int_0^t \int_{\Gamma_0} J\cdot\psi d\sigma ds. \tag{4.10}$$

In our two applications J is fixed such that the mapping

$$j : \{\varphi_0, \psi_0\} \to \int_0^t \int_{\Gamma_0} J\cdot\psi d\sigma ds$$

is a continuous linear form on F_k, $k = 2$ or 3. Consequently, the arguments of section 4 of [22] guarantee the existence of a solution $\{H, -E\} \in C([0,T], F_k')$ of (4.10) for all $\{\varphi, \psi\}$ solutions of (3.1) with initial data $\{\varphi_0, \psi_0\}$ in F_k (this solution $\{H, -E\}$ is then called the weak solution of (1.1)) with the property

$$E(T) = H(T) = 0,$$

once $\{H_0, -E_0\} \in F_k'$ is fixed such that (which is always possible)

$$\langle\{H_0, -E_0\}, \{\varphi_0, \psi_0\}\rangle = \|\{\varphi_0, \psi_0\}\|^2_{F_k}.$$

In the case $k = 2$, $J = -\psi \in L^2(\Sigma)^3$, and we directly arrive at the following theorem.

THEOREM 4.4. *Assume that Σ_0 satisfies the (ε,μ)-SOE. Then for all $\{H_0,-E_0\}\in H$, the control $J=-\psi\in L^2(\Sigma_0)^3$ drives our system* (1.1) *to rest at time T.*

In the case $k=3$, as in [22], we impose that

$$-\int_0^t\int_{\Gamma_0} J\cdot\psi d\sigma ds=\int_0^t\int_{\Gamma_0}(|\psi'|^2+|\,\mathbf{curl}\,\psi|^2)d\sigma ds,$$

since the continuity of j is then guaranteed. This yields

$$\begin{aligned} J=\psi''-\mu\sigma\times\varphi'\in(H^1(0,T;L^2(\Gamma_0)^3)'\\ \oplus L^2(\Sigma_0)\times L^2(0,T;H^{-1/2}(\mathrm{curl},\Gamma_0)'). \end{aligned}\tag{4.11}$$

Indeed, from the inclusion (4.9), we deduce that

$$\varphi,\psi\in H^1(0,T;H^{-1/2}(\Gamma_0)^3).\tag{4.12}$$

Moreover, writing

$$\mathbf{curl}\,\psi=\nu\times\frac{\partial\psi}{\partial\nu}+\sigma\times\psi,$$

where σ is a tangential differential operator of order 1, we see that

$$|\,\mathbf{curl}\,\psi|^2=\mu\varphi'\sigma\times\psi=\mu(\varphi'\cdot\nu)(\sigma\times\psi)\cdot\nu,$$

because $\varphi\times\nu=0$ on Σ. Moreover, as $\mathbf{curl}\,\psi\in L^2(\Sigma)^3$, we get

$$(\sigma\times\psi)\cdot\nu=\mathbf{curl}\,\psi\cdot\nu\in L^2(\Sigma).\tag{4.13}$$

The easily checked identity

$$(\sigma\times\psi)\cdot\nu=\mathrm{curl}\,\psi_T\text{ on }\Gamma$$

and (4.12)–(4.13) imply that

$$\psi_T\in L^2(0,T;H^{-1/2}(\mathrm{curl},\Gamma_0)),$$

where

$$H^{-1/2}(\mathrm{curl},\Gamma_0))=\{\chi\in H^{-1/2}(\Gamma_0)^2|\,\mathrm{curl}\,\chi\in L^2(\Gamma_0)\}.$$

Here $\mathrm{curl}\,\chi=\frac{\partial\chi_2}{\partial\tau_1}-\frac{\partial\chi_1}{\partial\tau_2}$ is the scalar curl of $\chi=(\chi_1,\chi_2)$ when $\{\tau_1,\tau_2\}$ is a direct orthonormal basis on Γ.

Therefore $\mu\sigma\times\varphi'$ is defined as

$$\langle\mu\sigma\times\varphi',\psi\rangle=\int_0^t\int_{\Gamma_0}\mu(\varphi'\cdot\nu)(\sigma\times\psi)\cdot\nu\,d\sigma dt.$$

Accordingly, we have proved the following theorem.

THEOREM 4.5. *Assume that Σ_0 satisfies the (ε,μ)-WOE. Then for all $\{H_0,-E_0\}\in D(A)'$, the control J given by* (4.11) *drives our system* (1.1) *to rest at time T.*

5. Determination of antennas. As in [33, 34], we shall give an application of the above results to an inverse problem, namely the determination and reconstruction of antennas from boundary measurements. From [33], we remark that the determination of antennas is simply based on the observation estimate (4.4) and on the isomorphic property of the integral operator K, defined by (5.5) hereafter, proved in [33] (see Theorem 5.1 below). On the contrary, from [33], we see that the reconstruction formula requires the observation estimate (4.4) (which yields the exact controllability result with control in $L^2(\Sigma_0)^3$) and appropriate properties of the operators K and Φ (defined by (5.8)) obtained in [33] (see Theorem 5.2 below). For that last problem, we do not use a relationship between the observation and the eigenvectors orthogonal basis but duality arguments.

To be more precise we consider Maxwell's equations where the volume current is injected from the exterior:

$$\left\{\begin{array}{l} \varepsilon\frac{\partial E}{\partial t} - \mathbf{curl}\, H = j \text{ in } Q, \\ \mu\frac{\partial H}{\partial t} + \mathbf{curl}\, E = 0 \text{ in } Q, \\ \operatorname{div}(\varepsilon E) = \operatorname{div}(\mu H) = 0 \text{ in } Q, \\ E \times \nu = 0, H \cdot \nu = 0 \text{ on } \Sigma, \\ E(0) = 0,\ H(0) = 0 \text{ in } \Omega. \end{array}\right. \tag{5.1}$$

Here $j(x,t)$ corresponds to an antenna and causes the electric and magnetic fields by (5.1). As in [33, 34], we assume that

$$j(x,t) = \lambda(t) f(x) \text{ for } x \in \Omega, t > 0.$$

The case $\lambda(t) = \cos(\omega t)$ for some fixed $\omega \in \mathbf{R}$ corresponds to an exterior current varying harmonically in time.

The problem that we will solve is the following one: Assume that λ is given. Determine f from the observation of H on a part Σ_0 of the boundary Σ. Here we will be concerned with the uniqueness, stability and reconstruction problems. These results are based on the results from Theorem 4.4 exchanging the role of ε and μ; therefore from now on we assume that Σ_0 satisfies the (μ,ε)-SOE.

The uniqueness and stability are mainly direct and were treated in [34] in the homogeneous case (ε and μ constant in Ω). To our knowledge, the reconstruction of f from the observation of H on Σ_0 is new even in the homogeneous case.

Hereafter we assume that

$$\lambda \in C^1([0,\infty)) \text{ and } \lambda(0) \neq 0. \tag{5.2}$$

The regularity assumption on λ guarantees that for any $f \in J(\Omega,\varepsilon)$, there exists a unique solution to (5.1) in $C^1([0,T]; J(\Omega,\varepsilon) \times \hat{J}(\Omega,\mu))$, denoted by $\{E(f), H(f)\}$.

We start with the uniqueness and stability results (compare with Theorems 4.1 and 4.2 of [34]).

THEOREM 5.1. *Under the above assumptions, there exists a positive constant C such that*

$$\int_\Omega \varepsilon |f(x)|^2\, dx \le C \int_{\Sigma_0} \left(|H(f)(x,t)|^2 + \left| \frac{\partial H(f)}{\partial t}(x,t) \right|^2 \right) d\sigma dt. \tag{5.3}$$

Therefore, if the solution $\{E(f), H(f)\}$ of (5.1) satisfies $H(f) = 0$ on Σ_0, then $f = 0$.

Proof. By Duhamel's principle, we have

$$E(f) = K(\varphi(f)), H(f) = K(\psi(f)), \tag{5.4}$$

where K is the integral operator defined by

$$(K\psi)(x,s) = \int_0^t \lambda(t-s)\psi(x,s)\, ds \quad \forall x \in \Gamma_0, t \in [0,T], \tag{5.5}$$

and $\{\varphi(f), \psi(f)\}$ is the unique solution (in $C([0,T]; J(\Omega,\varepsilon) \times \hat{J}(\Omega,\mu))$) of

$$\left\{\begin{array}{l} \varepsilon \frac{\partial \varphi(f)}{\partial t} - \mathbf{curl}\, \psi(f) = 0 \text{ in } Q, \\ \mu \frac{\partial \psi(f)}{\partial t} + \mathbf{curl}\, \varphi(f) = 0 \text{ in } Q, \\ \mathrm{div}(\varepsilon\varphi(f)) = \mathrm{div}(\mu\psi(f)) = 0 \text{ in } Q, \\ \varphi(f) \times \nu = 0, \psi(f) \cdot \nu = 0 \text{ on } \Sigma, \\ \varphi(f)(0) = f,\ \psi(f)(0) = 0 \text{ in } \Omega. \end{array}\right. \tag{5.6}$$

As Lemma 3 of [33] shows that the assumption (5.2) guarantees that K is an isomorphism from $L^2(\Gamma_0 \times (0,T))^3$ into $H^1(0,T; L^2(\Gamma_0)^3)$, there exists a positive constant $C(\lambda, T)$ (which means that the constant depends only on λ and T) such that

$$\int_{\Sigma_0} |\psi(f)(x,t)|^2\, d\sigma dt \le C(\lambda,T) \int_{\Sigma_0} \left(|H(f)(x,t)|^2 + \left| \frac{\partial H(f)}{\partial t}(x,t) \right|^2 \right) d\sigma dt. \tag{5.7}$$

The conclusion now follows from the estimate (4.4). ∎

We now pass to the reconstruction of f from the knowledge of $H(f)$ on Σ_0. First we need to introduce the following mappings. Thanks to Theorem 4.4, the mapping Π hereafter is well defined.

$$\Pi : J(\Omega,\varepsilon) \times \hat{J}(\Omega,\mu) \to L^2(\Sigma_0)^3 : (\varphi_0, \psi_0) \to J(\varphi_0, \psi_0),$$

where $J(\varphi_0, \psi_0)$ is the control (given by Theorem 4.4) which drives the problem

$$\left\{\begin{array}{l} \varepsilon \frac{\partial \varphi}{\partial t} - \mathbf{curl}\, \psi = 0 \text{ in } Q, \\ \mu \frac{\partial \psi}{\partial t} + \mathbf{curl}\, \varphi = 0 \text{ in } Q, \\ \mathrm{div}(\varepsilon\varphi) = \mathrm{div}(\mu\psi) = 0 \text{ in } Q, \\ \varphi \times \nu = -J(\varphi_0, \psi_0), \psi \cdot \nu = 0 \text{ on } \Sigma_0, \\ \varphi \times \nu = 0 \text{ on } \Sigma \setminus \Sigma_0, \psi \cdot \nu = 0 \text{ on } \Sigma, \\ \varphi(0) = \varphi_0,\ \psi(0) = \psi_0 \text{ in } \Omega \end{array}\right.$$

to rest at time T. Moreover, let us consider the bounded operator (see [33])

$$\Phi : L^2(\Sigma_0)^3 \to H^1(0,T; L^2(\Gamma_0)^3) : \eta \to \theta, \tag{5.8}$$

where θ is the unique solution of the Volterra integral equation (of the second kind):

$$\lambda(0)\theta'(x,t) + \int_t^T (\lambda'(\xi - t)\theta'(x,\xi) + \lambda(\xi - t)\theta(x,\xi))\, d\xi = \eta(x,t) \quad \forall (x,t) \in \Sigma_0,$$
$$\theta(x,0) = 0 \quad \forall x \in \Gamma_0.$$

According to section 3, we can finally fix the orthonormal basis $\{\Phi_k\}_{k\in\mathbf{Z}}$ of $J(\Omega,\varepsilon)\times\hat{J}(\Omega,\mu)$ of eigenvectors of the operator A (exchanging the role of ε and μ). For all $k\in\mathbf{Z}$, we shall write $\Phi_k=\{\varphi_k,\psi_k\}$.

We are now ready to state the reconstruction result.

THEOREM 5.2. *For all $k\in\mathbf{Z}$, set*

$$\theta_k=\Phi\Pi\Phi_k. \tag{5.9}$$

Then we have

$$\int_\Omega \varepsilon f\varphi_k\,dx=(H(f),\theta_k)_{H^1(0,T;L^2(\Gamma_0)^3)}\quad\forall k\in\mathbf{Z}. \tag{5.10}$$

Consequently, we have

$$f=\sum_{k\in\mathbf{Z}}(H(f),\theta_k)_{H^1(0,T;L^2(\Gamma_0)^3)}\varphi_k.$$

Proof. Since the set $\{\Phi_k\}_{k\in\mathbf{Z}}$ is a basis of $J(\Omega,\varepsilon)\times\hat{J}(\Omega,\mu)$, we may write

$$(f,0)=\sum_{k\in\mathbf{Z}}c_k\Phi_k,$$

where

$$c_k=\int_\Omega \varepsilon f\varphi_k\,dx.$$

Therefore it remains to prove the identity (5.10).

By (5.4) and the definition of θ_k, we have

$$\begin{aligned}(H(f),\theta_k)_{H^1(0,T;L^2(\Gamma_0)^3)}&=(K(\psi(f)),\Phi\Pi\Phi_k)_{H^1(0,T;L^2(\Gamma_0)^3)}\\&=(\psi(f),K^*\Phi\Pi\Phi_k)_{H^1(0,T;L^2(\Gamma_0)^3)}.\end{aligned}$$

Since it was shown in [33] (identity (5.9)) that $K^*\Phi=I$, the above identity becomes

$$(H(f),\theta_k)_{H^1(0,T;L^2(\Gamma_0)^3)}=\int_{\Sigma_0}\psi(f)\cdot\Pi\Phi_k\,d\sigma dt. \tag{5.11}$$

If we denote by $\{\tilde{\varphi}_k,\tilde{\psi}_k\}$ the (weak) solution of

$$\left\{\begin{array}{l}\varepsilon\frac{\partial\tilde{\varphi}_k}{\partial t}-\mathbf{curl}\,\tilde{\psi}_k=0\text{ in }Q,\\ \mu\frac{\partial\tilde{\psi}_k}{\partial t}+\mathbf{curl}\,\tilde{\varphi}_k=0\text{ in }Q,\\ \operatorname{div}(\varepsilon\tilde{\varphi}_k)=\operatorname{div}(\mu\tilde{\psi}_k)=0\text{ in }Q,\\ \tilde{\varphi}_k\times\nu=-\Pi\Phi_k\text{ on }\Sigma_0,\\ \tilde{\varphi}_k\times\nu=0\text{ on }\Sigma\setminus\Sigma_0,\tilde{\psi}_k\cdot\nu=0\text{ on }\Sigma,\\ \tilde{\varphi}_k(0)=\varphi_k,\ \tilde{\psi}_k(0)=\psi_k\text{ in }\Omega,\\ \tilde{\varphi}_k(T)=\tilde{\psi}_k(T)=0\text{ in }\Omega,\end{array}\right. \tag{5.12}$$

we have (see section 4)

$$\langle\{\tilde{\varphi}_k(0),\tilde{\psi}_k(0)\},\{\tilde{\varphi}_0,\tilde{\psi}_0\}\rangle=\int_{\Sigma_0}\Pi\Phi_k\cdot\tilde{\psi}\,d\sigma dt$$

for all $\{\tilde{\varphi}, \tilde{\psi}\}$ solutions of the adjoint problem to (5.12), i.e.,

$$\begin{cases} \varepsilon \frac{\partial \tilde{\varphi}}{\partial t} - \mathbf{curl}\, \tilde{\psi} = 0 \text{ in } Q, \\ \mu \frac{\partial \tilde{\psi}}{\partial t} + \mathbf{curl}\, \tilde{\varphi} = 0 \text{ in } Q, \\ \operatorname{div}(\varepsilon\tilde{\varphi}) = \operatorname{div}(\mu\psi) = 0 \text{ in } Q, \\ \tilde{\varphi} \times \nu = 0, \tilde{\psi} \cdot \nu = 0 \text{ on } \Sigma, \\ \tilde{\varphi}(0) = \tilde{\varphi}_0,\ \tilde{\psi}(0) = \tilde{\psi}_0 \text{ in } \Omega. \end{cases}$$

Since the pair $\{\varphi(f), \psi(f)\}$ solution of (5.6) is also solution of the above problem, we get

$$\langle \{\tilde{\varphi}_k(0), \tilde{\psi}_k(0)\}, \{f, 0\}\rangle = \int_{\Sigma_0} \Pi\Phi_k \cdot \psi(f)\, d\sigma dt. \tag{5.13}$$

The conclusion follows from the two identities (5.11) and (5.13). □

Remark 5.3. When ε and μ are piecewise constant with Lipschitz polyhedral subdomains, the above observation and reconstruction results hold for $\Sigma_0 = \Sigma$ under the assumptions of Corollary 6.11 (exchanging the rule of ε and μ). In that case the constant C appearing in (5.3) may be estimated as follows (see Lemma 4.2 and Corollary 6.11)

$$C \leq C(\lambda, T) \frac{\sup_{x\in\Gamma} \mu(x) \sup_{x\in\Gamma} m(x)\cdot\nu(x)}{2(T - T_0)},$$

where $C(\lambda, T)$ is the constant appearing in (5.7) and T_0 may be estimated by (6.26).

Similar results from partial observation (i.e., $\Sigma_0 \neq \Sigma$) hold when Σ_0 satisfies the geometrical control property under the assumptions that ε and μ are smooth and constant near Γ and that Γ is smooth (without estimation of the constant in (5.3)); for nonsmooth ε, μ and Γ, some investigations are still necessary (see sections 1 and 4).

6. Checking the observation estimates. In this section, we assume that Ω is a Lipschitz polyhedron. We further suppose that ε and μ are piecewise constant with Lipschitz polyhedra Ω_j. As already mentioned, we will give sufficient conditions on the partition $\mathcal{P}$ and the parameters ε and μ such that the strong and weak observation estimates hold. First of all, we need to introduce the space $PH^1(\Omega, \mathcal{P})$ of piecewise H^1 functions relatively to the partition $\mathcal{P}$; more precisely

$$PH^1(\Omega, \mathcal{P}) = \{\varphi \in L^2(\Omega) | \varphi_j \in H^1(\Omega_j) \quad \forall j = 1, \ldots, J\},$$

where, of course, φ_j means the restriction of φ to Ω_j.

For all $j = 1, \ldots, J$, we denote by $F_{jk}, k = 1, \ldots, k_j$, the open faces of the boundary of Ω_j. Let $\mathcal{F}_{\text{int}} = \{F_{jk} | F_{jk} \subset \Omega\}$ be the set of interior faces (contained in Ω) and let $\mathcal{F}_{\text{ext}}$ be the set of exterior faces (contained in Γ).

6.1. Some identities. As in [22], the observation estimates will be obtained with the help of an identity with multiplier, which, in our case, will be permitted under the following regularity assumptions:

$$J^1_\tau(\Omega, \mu) \hookrightarrow PH^1(\Omega, \mathcal{P})^3, \tag{6.1}$$

$$J^1_\nu(\Omega, \varepsilon) \hookrightarrow PH^1(\Omega, \mathcal{P})^3. \tag{6.2}$$

Owing to Theorem 3.5 of [10], the first inclusion (resp., second) holds if the operator $-\Delta_\mu^{\mathrm{Dir}}$ (resp. $-\Delta_\varepsilon^{\mathrm{Neu}}$) has no edge singular exponents in $(0,1]$ and no corner singular exponents in $(0,1/2]$ (see [10] for the right definitions). Before going on, let us remark that the above assumptions do not guarantee (see Theorem 7.1 of [10]) that the spaces $J_\tau^\star(\Omega,\mu,\varepsilon)$ and $J_\nu^\star(\Omega,\mu,\varepsilon)$ are embedded into the space of piecewise H^2 functions (as it is the case when ε and μ are constant and Γ is smooth [22]). Consequently, the integrations by parts used hereafter in order to establish the identity with multiplier require careful attention.

As already mentioned, when ε and μ are constant, the above inclusions hold if Ω is convex or if Ω has a $C^{1,1}$ boundary. Here are three examples when the inclusions (6.1) and (6.2) hold. In the first one, we leave the general setting of this section, but all the results below still hold for that example.

Example 6.1. Assume that ε and μ are constant. If Ω has a smooth boundary except at some points $x_k, k = 1,\dots,K$, where it coincides with a revolution cone centered at x_k with opening $\theta_k \in]0,\pi[$, then thanks to [8, section 2] and [11, section 18.D], (6.1) holds for all $\theta_k \le \theta_0 \cong 134^\circ$ (which is nonconvex if one θ_k is larger than 90°). On the other hand, by [8, section 2] and Proposition 10 of [12], (6.2) holds $\forall \theta_k \in]0,\pi[$.

Example 6.2. If Ω is a parallelepiped divided into two subdomains separated by a plane parallel to two faces, then by the results from section 7.a of [10], the assumptions (6.1) and (6.2) hold.

Example 6.3. Assume that Ω is convex and that any edge of any Ω_j is an edge of Ω (see Figures 1 and 2). Denote by $\varepsilon_{\min} = \min_{j=1,\dots,J} \varepsilon_j$ and $\varepsilon_{\max} = \max_{j=1,\dots,J} \varepsilon_j$. If the ratio $\frac{\mu_{\min}}{\mu_{\max}}$ (resp., $\frac{\varepsilon_{\min}}{\varepsilon_{\max}}$) is sufficiently close to 1, then (6.1) (resp., (6.2)) holds. Let us sketch the proof of (6.1). With the notation from [10], if we denote by $\nu_{\mu,c}$ (resp., $\nu_{\mu,e}$) the smallest eigenvalue of the Laplace–Beltrami operator $L_{\mu,c}^{\mathrm{Dir}}$ (resp., $L_{\mu,e}^{\mathrm{Dir}}$) associated with the corner c (resp., edge e) (a corner (resp., edge) is any corner (resp., edge) of any Ω_j), by the min-max principle we have

$$\nu_{\mu,c} \ge \frac{\mu_{\min}}{\mu_{\max}} \nu_{1,c},$$
$$\nu_{\mu,e} \ge \frac{\mu_{\min}}{\mu_{\max}} \nu_{1,e}.$$

We further recall that (see [10]) the smallest singular exponent $\lambda_{\mu,c}$ (resp., $\lambda_{\mu,e}$) of $-\Delta_\mu^{\mathrm{Dir}}$ associated with the corner e (resp., edge e) is given by

$$\lambda_{\mu,c} = -\frac{1}{2} + \sqrt{\nu_{\mu,c} + \frac{1}{4}},$$
$$\lambda_{\mu,e} = \sqrt{\nu_{\mu,e}}.$$

Consequently, $\lambda_{\mu,c} > \frac{1}{2}$ (resp., $\lambda_{\mu,e} > 1$) if and only if $\nu_{\mu,c} > \frac{3}{4}$ (resp., $\nu_{\mu,e} > 1$).

As the convexity of Ω and the assumption on $\mathcal{P}$ guarantee that $\lambda_{1,c} > \frac{1}{2}$ and $\lambda_{1,e} > 1$, we conclude that $\lambda_{\mu,c} > \frac{1}{2}$ and $\lambda_{\mu,e} > 1$ if

$$\frac{\mu_{\min}}{\mu_{\max}} > \max\left\{\frac{3}{4\nu_{1,c}}, \frac{1}{\nu_{1,2}}\right\}$$

for all corners c and all edges e.

Note that the assumption on the edges is made for the sake of simplicity; indeed if one edge e of some Ω_j is included into one face of Ω or included into Ω, then $\lambda_{1,e} = 1$

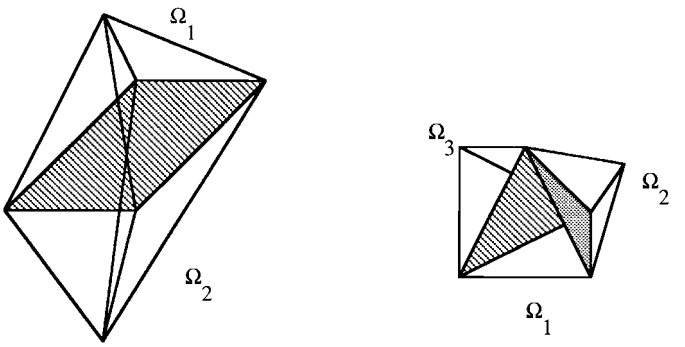


FIG. 1.

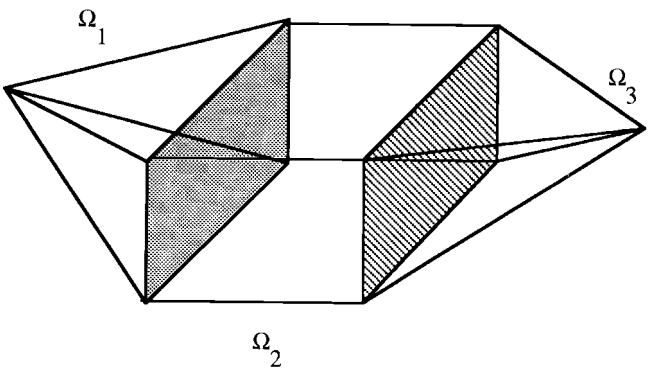


FIG. 2.

and the above argument does not allow us to conclude that $\lambda_{\mu,e} > 1$ (see [26] for some numerical examples). Nevertheless in some particular cases (as in Figure 3), a direct calculation shows that $\lambda_{\mu,e} > 1$ if the ratio $\frac{\mu_{\min}}{\mu_{\max}}$ is sufficiently close to 1.

We further introduce a vector field m defined by

$$m(x) = x - x_0,$$

where x_0 is a fixed point.

Let us now fix $\varphi_0 \in J_\tau^\star(\Omega, \mu, \varepsilon)$ and $\psi_0 \in J_\nu^\star(\Omega, \mu, \varepsilon)$, and let φ, ψ be the solution of (3.1) which has the regularity (due to Theorem 3.2)

$$\varphi \in C^2([0,\infty), J(\Omega,\mu)) \cap C^1([0,\infty), J_\tau^1(\Omega,\mu)) \cap C([0,\infty), J_\tau^\star(\Omega,\mu,\varepsilon)), \tag{6.3}$$

$$\psi \in C^2([0,\infty), \hat{J}(\Omega,\varepsilon)) \cap C^1([0,\infty), J_\nu^1(\Omega,\varepsilon)) \cap C([0,\infty), J_\nu^\star(\Omega,\mu,\varepsilon)). \tag{6.4}$$

We start with a technical property.

LEMMA 6.4. *Under the above assumptions, ψ satisfies*

$$m \cdot \nabla\psi_j \in H(\mathbf{curl}, \Omega_j) \quad \forall j = 1, \dots, J,$$

where we recall that (see [16, 2]*)* $H(\mathbf{curl}, \Omega) = \{\psi \in L^2(\Omega)^3 | \, \mathbf{curl}\, \psi \in L^2(\Omega)^3\}$.

Proof. The fact that $m \cdot \nabla\psi_j$ belongs to $L^2(\Omega_j)^3$ follows from the inclusion

$$J_\nu^\star(\Omega,\mu,\varepsilon) \subset J_\nu^1(\Omega,\varepsilon)$$

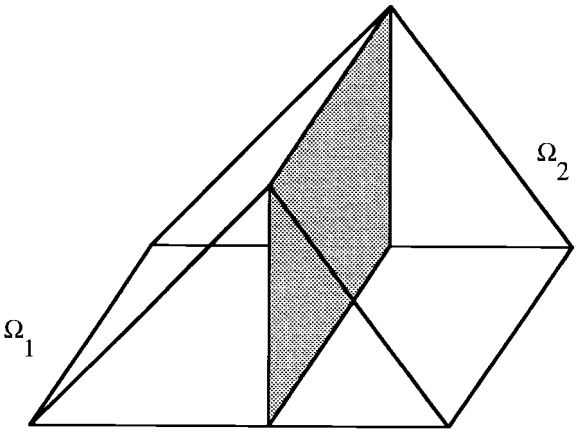


FIG. 3.

and the assumption (6.2).

For the curl, we make use of the identity

$$\mathbf{curl}(m \cdot \nabla\psi) = \mathbf{curl}\,\psi + m \cdot \nabla(\mathbf{curl}\,\psi). \tag{6.5}$$

But the regularity (6.4) implies that

$$\mu^{-1}\,\mathbf{curl}\,\psi \in J^1_\tau(\Omega,\mu),$$

and by the assumption (6.1) we conclude that $\mathbf{curl}\,\psi \in PH^1(\Omega,\mathcal{P})$. □

We now prove the following identity.

LEMMA 6.5. *Under the above assumptions, we have (hereafter ψ' means the time derivative of ψ)*

$$\begin{aligned}\frac{1}{2}\int_\Sigma m\cdot\nu(\varepsilon|\psi'|^2 - \mu^{-1}|\,\mathbf{curl}\,\psi|^2)d\sigma dt\\ +\frac{1}{2}\int_Q(\mu^{-1}|\,\mathbf{curl}\,\psi|^2 - 3\varepsilon|\psi'|^2)dxdt + X_1 + I_1 + I_2 + I_3 = 0,\end{aligned} \tag{6.6}$$

where we have set

$$\begin{aligned}X_1 &= -\int_\Omega \varepsilon\psi' m\cdot\nabla\psi dx|_0^T,\\ I_1 &= \sum_{F\in\mathcal{F}_{\rm int}}\int_{F\times(0,T)}\varphi'\times\nu_F[m\cdot\nabla\psi]_F d\sigma dt,\\ I_2 &= \frac{1}{2}\sum_{F\in\mathcal{F}_{\rm int}}\int_{F\times(0,T)} m\cdot\nu_F[\varepsilon|\psi'|^2]_F d\sigma dt,\\ I_3 &= -\frac{1}{2}\sum_{F\in\mathcal{F}_{\rm int}}\int_{F\times(0,T)}\mu^{-1}m\cdot\nu_F[|\,\mathbf{curl}\,\psi|^2]_F d\sigma dt,\end{aligned}$$

where the jump $[q]_F = q_j - q_{j'}$ if F belongs to $\partial\Omega_j$ and to $\partial\Omega_{j'}$, while ν_F is the normal vector directed from Ω_j to $\Omega_{j'}$.

Proof. Since φ' belongs to $PH^1(\Omega,\mathcal{P})$, by Lemma 6.4, Green's formula yields

$$\int_Q \varphi' \,\mathbf{curl}(m\cdot\nabla\psi)dxdt = \int_Q \mathbf{curl}\,\varphi' m\cdot\nabla\psi dxdt + I_1.$$

As $\mathbf{curl}\,\varphi' = -\varepsilon\psi''$, an integration by parts in t leads to

$$\int_Q \varphi' \,\mathbf{curl}(m\cdot\nabla\psi)dxdt = \int_Q \varepsilon\psi' m\cdot\nabla\psi' dxdt + X_1 + I_1. \tag{6.7}$$

To transform the first term of this right-hand side, we use the identity

$$\psi' m\cdot\nabla\psi' = \frac{1}{2}\mathrm{div}(m|\psi'|^2) - \frac{\mathrm{div}m}{2}|\psi'|^2. \tag{6.8}$$

Moreover, in order to integrate by parts in the term $\int_Q \varepsilon\mathrm{div}(m|\psi'|^2)$, we shall check that

$$m|\psi_j'|^2 \in W_p(\mathrm{div},\Omega_j) \quad \forall j=1,\dots,J \tag{6.9}$$

for some $p>1$. Here we use the space

$$W_p(\mathrm{div},\Omega) = \{u\in L^p(\Omega)^3 | \mathrm{div}u \in L^p(\Omega)\}.$$

In the event that $p=2$, $W_p(\mathrm{div},\Omega)$ is nothing else than $H(\mathrm{div},\Omega)$ [16]. Note that if $u\in W_p(\mathrm{div},\Omega)$, then its trace $u\cdot\nu$ belongs to $W^{-1/p,p}(\Gamma)$ and the next Green's formula holds (the proof follows the same line as the one given in [16] for $p=2$):

$$\int_\Omega (u\cdot\nabla v + \mathrm{div}uv)\,dx = \langle u\cdot\nu, v\rangle \quad \forall v\in W^{1,q}(\Omega), \tag{6.10}$$

when $1/p+1/q=1$.

Let us now verify (6.9). First, as ψ_j' belongs to $H^1(\Omega_j)^3$, the Sobolev imbedding theorem leads to

$$|\psi_j'|^2 \in L^p(\Omega_j) \quad \forall p\le 3.$$

Secondly, using the identity (6.8), it remains to check that

$$\psi_j' m\cdot\nabla\psi_j' \in L^p(\Omega_j) \tag{6.11}$$

for some $p>1$. Using again the above regularity of ψ_j', the Sobolev imbedding theorem, and Hölder's inequality, we arrive at (6.11) for all $p\le 3/2$. Therefore, (6.9) also holds for all $p\le 3/2$.

Using (6.9) and (6.10), we obtain

$$\int_Q \varepsilon\mathrm{div}(m|\psi'|^2) = I_2 + \frac{1}{2}\int_\Sigma \varepsilon m\cdot\nu|\psi'|^2 d\sigma dt.$$

Inserting this identity into (6.7), we arrive at

$$\begin{aligned}\int_Q \varphi' \,\mathbf{curl}(m\cdot\nabla\psi)dxdt &= \frac{1}{2}\int_\Sigma \varepsilon m\cdot\nu|\psi'|^2 \\ &\quad -\frac{3}{2}\int_Q \varepsilon|\psi'|^2 dxdt + X_1 + I_1 + I_2.\end{aligned} \tag{6.12}$$

In a second step using the identity (6.5), we have

$$\int_Q \mu^{-1}\,\mathbf{curl}\,\psi\,\mathbf{curl}(m\cdot\nabla\psi)dxdt = -\frac{1}{2}\int_Q \mu^{-1}|\,\mathbf{curl}\,\psi|^2 dxdt + \frac{1}{2}\int_Q \mu^{-1}\mathrm{div}(m|\,\mathbf{curl}\,\psi|^2)dxdt. \tag{6.13}$$

As before, in order to integrate by parts in the second term of this right-hand side, we readily check that

$$m|\,\mathbf{curl}\,\psi_j|^2 \in W_p(\mathrm{div},\Omega_j) \quad \forall j=1,\dots,J$$

for all $p \le 3/2$. This regularity and (6.10) allow us to transform (6.13) into

$$\int_Q \mu^{-1}\,\mathbf{curl}\,\psi\,\mathbf{curl}(m\cdot\nabla\psi)dxdt = -\frac{1}{2}\int_Q \mu^{-1}|\,\mathbf{curl}\,\psi|^2 dxdt - I_3 + \frac{1}{2}\int_\Sigma \mu^{-1} m\cdot\nu|\,\mathbf{curl}\,\psi|^2 d\sigma dt. \tag{6.14}$$

The difference between (6.12) and (6.14) directly gives (6.6). ∎

The second interesting identity is the goal of the next lemma.

LEMMA 6.6. *Under the above assumptions, we have*

$$\int_Q \varepsilon|\psi'|^2 dxdt = \int_Q \mu^{-1}|\,\mathbf{curl}\,\psi|^2 dxdt - X_2, \tag{6.15}$$

where we have set

$$X_2 = \int_\Omega \mu\varphi\cdot\varphi' dx|_0^T.$$

Proof. We multiply the identity $\psi' + \varepsilon^{-1}\,\mathbf{curl}\,\varphi = 0$ by $\varepsilon\psi'$ and integrate on Q. This gives

$$0 = \int_Q \varepsilon|\psi'|^2 dxdt + \int_Q \mathbf{curl}\,\varphi\psi' dxdt.$$

In this second term we can integrate by parts in each Ω_j, since ψ_j' is in $H^1(\Omega_j)$ and φ_j in $H(\mathbf{curl},\Omega_j)$. This leads to

$$0 = \int_Q \varepsilon|\psi'|^2 dxdt + \int_Q \varphi\,\mathbf{curl}\,\psi' dxdt,$$

because the jump of $\varphi\times\nu_F\cdot\psi'$ is zero through any interior face F. Using now the fact that $\mathbf{curl}\,\psi' = \mu\varphi''$ and an integration by part in t, we arrive at (6.15). ∎

At this stage, we are able to give the main identity with multiplier (compare with the identity (3.12) of [22]).

LEMMA 6.7. *With the above assumptions and notation, it holds that*

$$\frac{1}{2}\int_Q (\varepsilon^{-1}|\,\mathbf{curl}\,\varphi|^2 + \mu^{-1}|\,\mathbf{curl}\,\psi|^2)dxdt = \frac{1}{2}\int_\Sigma m\cdot\nu(\varepsilon^{-1}|\,\mathbf{curl}\,\varphi|^2 - \mu^{-1}|\,\mathbf{curl}\,\psi|^2)d\sigma dt + X_1 + X_2 + I_1 + I_2 + I_3. \tag{6.16}$$

Proof. The identity (6.6) may be equivalently written

$$\int_Q \varepsilon|\psi'|^2 dxdt + \frac{1}{2}\int_Q (\varepsilon|\psi'|^2 - \mu^{-1}|\operatorname{\mathbf{curl}}\psi|^2)dxdt$$
$$= \frac{1}{2}\int_\Sigma m\cdot\nu(\varepsilon|\psi'|^2 - \mu^{-1}|\operatorname{\mathbf{curl}}\psi|^2)d\sigma dt + X_1 + I_1 + I_2 + I_3.$$

Substituting (6.15) into this identity, we get (6.16) because $\psi' = -\varepsilon^{-1}\operatorname{\mathbf{curl}}\varphi$. ∎

At this stage, to obtain the observation estimates, we need a geometrical assumption in order to cancel the interface terms $I_i, i = 1, 2, 3$ (to avoid internal control!). Namely we assume that x_0 may be chosen such that

$$m\cdot\nu_F = 0 \text{ on } F \quad \forall F \in \mathcal{F}_{\text{int}}. \tag{6.17}$$

Note that this condition holds in the setting of Example 6.2. Actually it holds if and only if the planes containing the interior faces have (at least) one point in common (see Figures 1 and 3 for some examples and Figure 2 for a counterexample).

LEMMA 6.8. *Under the assumption* (6.17)*, we have*

$$I_i = 0 \quad \forall i = 1, 2, 3.$$

Proof. The nullity of I_2 and I_3 is direct. For I_1, we use the fact that $\psi\times\nu_F$ is continuous through any interior face F. Consequently, we can prove that

$$(\varphi'\times\nu_F)[m\cdot\nabla\psi]_F = (\varphi'\times\nu_F)m\cdot\nu_F\left[\frac{\partial\psi_T}{\partial\nu_F}\right]_F \text{ on } F,$$

where $\psi_T = \psi - (\psi\cdot\nu_F)\nu_F$ is the tangent part of ψ. ∎

6.2. Some estimates. We are now ready to formulate the desired estimates.

THEOREM 6.9. *Let φ, ψ be the solution of* (3.1) *with the regularity* (6.3)–(6.4)*. Assume that* (6.1), (6.2)*, and* (6.17) *hold. Then there exists a minimal time $T_0 > 0$ such that*

$$(T - T_0)\tilde{E}(0) \le \frac{1}{2}\int_\Sigma m\cdot\nu(\varepsilon^{-1}|\operatorname{\mathbf{curl}}\varphi|^2 - \mu^{-1}|\operatorname{\mathbf{curl}}\psi|^2)d\sigma dt. \tag{6.18}$$

Proof. The assumption (6.2) and Theorem 2.2 guarantee the existence of two positive constants C_1, C_2 such that

$$\left(\int_\Omega \varepsilon|\nabla\chi|^2\,dx\right)^{1/2} \le C_1\left(\int_\Omega \mu^{-1}|\operatorname{\mathbf{curl}}\chi|^2\,dx\right)^{1/2} \quad \forall\chi\in J^1_\nu(\Omega,\varepsilon), \tag{6.19}$$

$$\left(\int_\Omega \mu|\chi|^2\,dx\right)^{1/2} \le C_2\left(\int_\Omega \varepsilon^{-1}|\operatorname{\mathbf{curl}}\chi|^2\,dx\right)^{1/2} \quad \forall\chi\in J^1_\tau(\Omega,\mu). \tag{6.20}$$

These two estimates allow us to estimate X_1 and X_2; namely, by the Cauchy–Schwarz inequality, we have

$$|X_1| \le 2R(x_0)C_1\tilde{E}(0),$$
$$|X_2| \le 2C_2\tilde{E}(0),$$

where we have set

$$R(x_0) = \max_{x\in\bar{\Omega}}|x - x_0|.$$

Starting from the identity (6.16), recalling that the I_i are zero, and using the above estimates, we obtain (6.18) with

$$T_0 = 2\max(C_1 R(x_0), C_2). \qquad \square \tag{6.21}$$

COROLLARY 6.10. *Under the assumptions of Theorem* 6.9, Σ_T *satisfies the* (ε,μ)*-WOE for all* $T > T_0$ *with*

$$C \le \frac{\sup_{x\in\Gamma} m(x)\cdot\nu(x)}{2(T-T_0)}. \tag{6.22}$$

COROLLARY 6.11. *In addition to the above hypotheses* (6.1), (6.2), *and* (6.17), *if we assume that* Γ *is star-shaped with respect to some point* x_0 *(for which* (6.17) *holds), i.e.,*

$$m\cdot\nu \ge 0 \text{ on } \Gamma, \tag{6.23}$$

then Σ_T *satisfies the* (ε,μ)*-SOE for all* $T > T_0$, *with* C *estimated by* (6.22).

Proof. Under the assumption (6.23), the estimate (6.18) simplifies to

$$(T-T_0)\tilde{E}(0) \le \frac{1}{2}\int_\Sigma m\cdot\nu\varepsilon^{-1}|\,\mathbf{curl}\,\varphi|^2 d\sigma dt, \tag{6.24}$$

which proves the corollary. $\square$

Example 6.12. If ε and μ are constant, then Theorems 4.4 and 4.5 hold if Ω is convex, if it has a $C^{1,1}$ boundary or if it satisfies the hypothesis of Example 6.1. The same results hold in the setting of Example 6.2. Finally, in the case of Example 6.3, they hold if, in addition to the hypotheses prescribed in Example 6.3, the assumption (6.17) holds (see Figures 1 and 3).

6.3. Estimate of the minimal time of control. We finish the paper by giving an upper bound on the constants C_1 and C_2 appearing in (6.19) and (6.20), which yields an estimate of the minimal time T_0 for which the observation estimates hold and consequently for which control results hold.

LEMMA 6.13. *If* (6.2) *holds, then* (6.19) *holds with* $C_1 \le C_J^{1/2}$, *where we have set* $C_J = \max_{j=1,\dots,J}\{\varepsilon_j\mu_j\}$.

Proof. By Lemma 2.2 and Theorem 2.1 of [10], it holds that

$$\int_\Omega \varepsilon|\nabla\chi|^2\,dx = \int_\Omega \varepsilon(|\,\mathbf{curl}\,\chi|^2 + |\mathrm{div}\chi|^2)\,dx \quad \forall\chi\in\mathbf{H}_T(\Omega,\varepsilon),$$

where the space $\mathbf{H}_T(\Omega,\varepsilon)$ is defined by

$$\mathbf{H}_T(\Omega,\varepsilon) = \{\chi\in PH^1(\Omega,\mathcal{P})^3 | \mathrm{div}(\varepsilon\chi)\in L^2(\Omega) \text{ and } \chi\cdot\nu = 0 \text{ on } \Gamma\}.$$

Since the assumption (6.2) guarantees that $J^1_\nu(\Omega,\varepsilon)\subset\mathbf{H}_T(\Omega,\varepsilon)$, the above identity implies that

$$\int_\Omega \varepsilon|\nabla\chi|^2\,dx = \int_\Omega \varepsilon|\,\mathbf{curl}\,\chi|^2\,dx \quad \forall\chi\in J^1_\nu(\Omega,\varepsilon).$$

This identity allows us to conclude because $\varepsilon \le C_J\mu^{-1}$. $\square$

LEMMA 6.14. *If* (6.1) *and* (6.17) *hold, then* (6.20) *holds with* $C_2 \le R(x_0)C_J^{1/2}$.

Proof. Fix $\chi \in J^1_\tau(\Omega,\mu)$. By Green's formula, we have

$$\int_\Omega \mu\, \mathbf{curl}\, \chi \cdot (m \times \chi)\, dx = \int_\Omega \mu\chi \cdot \mathbf{curl}(m \times \chi)\, dx + \sum_{F \in \mathcal{F}_{\rm int}} \int_F \chi \times \nu_F \cdot [\mu m \times \chi]_F d\sigma.$$

By the assumption (6.17) and the fact that $[\mu\chi \cdot \nu_F]_F = 0$ on F, we conclude that $[\mu m \times \chi]_F$ is orthogonal to $F\ \forall\ F \in \mathcal{F}_{\rm int}$. As $\chi \times \nu_F$ is tangent to F, for all $F \in \mathcal{F}_{\rm int}$, the boundary terms of the above identity are equal to zero. Since

$$\mathbf{curl}(m \times \chi) = -2\chi + (m \cdot \nabla)\chi,$$

the above identity becomes

$$2\int_\Omega \mu|\chi|^2\, dx = \int_\Omega \mu(\chi \cdot (m \cdot \nabla)\chi - \mathbf{curl}\, \chi \cdot (m \times \chi))\, dx.$$

By the Cauchy–Schwarz inequality, we get

(6.25)
$$2\left(\int_\Omega \mu|\chi|^2\, dx\right)^{1/2} \le R(x_0)\left\{\left(\int_\Omega \mu|\nabla\chi|^2\, dx\right)^{1/2} + \left(\int_\Omega \mu|\,\mathbf{curl}\,\chi|^2\, dx\right)^{1/2}\right\}.$$

As before, the assumption (6.1) and Lemma 2.2 and Theorem 2.1 of [10] imply that

$$\int_\Omega \mu|\nabla\chi|^2\, dx \le C_J \int_\Omega \varepsilon^{-1}|\,\mathbf{curl}\,\chi|^2\, dx.$$

Inserting this estimate in (6.25) and using the fact that $\mu \le C_J\varepsilon^{-1}$, we arrive at the conclusion. ∎

Corollary 6.15. *Under the assumptions of Theorem* 6.9*, we have*

$$T_0 \le 2R(x_0)\left(\max_{j=1,\dots,J}\{\varepsilon_j\mu_j\}\right)^{1/2}. \tag{6.26}$$

Remark 6.16. In the case where ε and μ are constant in the whole of Ω, the above estimate reduces to $T_0 \le 2R(x_0)\sqrt{\varepsilon\mu}$, which is in accordance with the estimate obtained for the wave equation [23] and with Theorem 1.3 of [20]. In particular, if Ω is convex, then for $x_0 \in \Omega$ such that $2R(x_0) = \operatorname{diam}\Omega$, we hit the optimal time of control since the speed of propagation of the electromagnetic wave is $\frac{1}{\sqrt{\varepsilon\mu}}$. In the case where ε and μ are piecewise constant, the estimate (6.26) is rather good since its right-hand side only retains the slower wave.

Acknowledgments. We thank Dr. F. Jochmann and Dr. K. D. Phung for some fruitful discussions. We are also grateful to the referees for useful remarks and comments on the first version of this paper, allowing its improvement.

REFERENCES

[1] N. U. Ahmed and T. Wan, *Exact boundary controllability of electromagnetic fields in general regions*, Dynam. Systems Appl., 5 (1996), pp. 229–243.

[2] C. Amrouche, C. Bernardi, M. Dauge, and V. Girault, *Vector potentials in three-dimensional nonsmooth domains*, Math. Methods Appl. Sci., 21 (1998), pp. 823–864.
[3] C. Bardos, G. Lebeau, and J. Rauch, *Contrôle et stabilisation dans les problèmes hyperboliques*, in Contrôlabilité exacte, perturbations et stabilisation de systèmes distribués, J.-L. Lions, ed., tome 1, Rech. Math. Appl. 8, Masson, Paris, 1988.
[4] A. Bensoussan, *Some remarks on the exact controllability of Maxwell's equations*, in Differential Equation and Control Theory, V. Barbu, ed., Pitman Res. Notes Math. Ser. 250, Longman, Harlow, UK, 1991, pp. 17–29.
[5] M. Birman and M. Solomyak, *L^2-theory of the Maxwell operator in arbitrary domains*, Russian Math. Surveys, 42 (1987), pp. 75–96.
[6] N. Burq, *Contrôlabilité exacte des ondes dans des ouverts peu réguliers,* Asymptot. Anal., 14 (1997), pp. 157–191.
[7] N. Burq and J.-M. Schlenker, *Contrôle de l'équation des ondes dans des ouverts comportant des coins,* Bull. Soc. Math. France, 126 (1998), pp. 601–637.
[8] M. Costabel and M. Dauge, *Singularité des équations de Maxwell dans un polyèdre*, C. R. Acad. Sci. Paris Sér. I Math., 324 (1997), pp. 1005–1010.
[9] M. Costabel and M. Dauge, *Singularities of Maxwell's Equations on Polyhedral Domains*, preprint IRMAR 97-19, Université de Rennes 1, France, 1997, http://www.maths.univ-rennes1.fr/~dauge/.
[10] M. Costabel, M. Dauge, and S. Nicaise, *Singularities of Maxwell interface problems,* RAIRO Modél. Math. Anal. Numér., 33 (1999), pp. 627–649.
[11] M. Dauge, *Elliptic Boundary Value Problems in Corner Domains. Smoothness and Asymptotics of Solutions,* Lecture Notes in Math. 1341, Springer-Verlag, Berlin, 1988.
[12] M. Dauge and M. Pogu, *Existence et régularité de la fonction potentiel pour des écoulements subcritiques s'établissant autour d'un corps à singularité conique,* Ann. Fac. Sci. Toulouse Math., 9 (1988), pp. 213–242.
[13] G. Duvaut and J.-L. Lions, *Les inéquations en mécanique et en physique,* Dunod, Paris, 1972.
[14] M. Eller, V. Isakov, G. Nakamura, and D. Tataru, *Uniqueness and Stability in the Cauchy Problem for Maxwell and Elasticity Systems*, preprint, Northwestern University, Evanston, IL, 1998, http://www.math.nwu.edu/~tataru.
[15] N. Filonov, *Système de Maxwell dans des domaines singuliers,* thesis, Université de Bordeaux 1, France, 1996.
[16] V. Girault and P.-A. Raviart, *Finite Element Methods for Navier-Stokes Equations,* Springer Ser. Comput. Math. 5, Springer-Verlag, New York, 1986.
[17] P. Grisvard, *Elliptic Problems in Nonsmooth Domains,* Monographs and Studies in Mathematics 21, Pitman, Boston, 1985.
[18] B. V. Kapitonov, *Stabilization and exact boundary controllability for Maxwell's equations*, SIAM J. Control Optim., 32 (1994), pp. 408–420.
[19] K. A. Kime, *Boundary controllability of Maxwell's equations in a spherical region*, SIAM J. Control Optim., 28 (1990), pp. 294–319.
[20] V. Komornik, *Boundary stabilization, observation and control of Maxwell's equations,* Panamer. Math. J., 4 (1994), pp. 47–61.
[21] O. A. Ladyzhenskaya and V. A. Solonikov, *The linearization principle and invariant manifolds for problems of magnetohydrodynamics*, J. Soviet Math., 8 (1977), pp. 384–422.
[22] J. E. Lagnese, *Exact boundary controllability of Maxwell's equations in a general region*, SIAM J. Control Optim., 27 (1989), pp. 374–388.
[23] J.-L. Lions, *Contrôlabilité exacte, perturbations et stabilisation de systèmes distribués,* tome 1, Rech. Math. Appl. 8, Masson, Paris, 1988.
[24] O. Nalin, *Contrôlabilité exacte sur une partie du bord des équations de Maxwell*, C. R. Acad. Sci. Paris Sér. I Math., 309 (1989), pp. 811–815.
[25] S. Nicaise, *About the Lamé system in a polygonal or a polyhedral domain and a coupled problem between the Lamé system and the plate equation* II: *Exact controllability*, Ann. Scuola Norm. Sup. Pisa Cl. Sci. (4), 20 (1993), pp. 163–191.
[26] S. Nicaise and A. M. Sändig, *General interface problems* I/II, Math. Methods Appl. Sci., 17 (1994), pp. 395–450.
[27] A. Pazy, *Semigroups of Linear Operators and Applications to Partial Differential Equations,* Appl. Math. Sci. 44, Springer-Verlag, New York, 1983.
[28] K. D. Phung, *Stabilisation frontière du système de Maxwell avec la condition aux limites absorbante de Silver-Müller*, C. R. Acad. Sci. Paris Sér. I Math., 320 (1995), pp. 187–192.
[29] K. D. Phung, *Contrôlabilité exacte et stabilisation interne des équations de Maxwell*, C. R. Acad. Sci. Paris Sér. I Math., 323 (1996), pp. 169–174.

[30] K. D. Phung, *Contrôle et stabilisation d'ondes électromagnétiques*, ESAIM Control Optim. Calc. Var., 5 (2000), pp. 87–137.
[31] D. L. Russell, *The Dirichlet–Neumann boundary control problem associated with Maxwell's equations in a cylindrical region*, SIAM J. Control Optim., 24 (1986), pp. 199–229.
[32] Ch. Weber, *A local compactness theorem for Maxwell's equations*, Math. Methods. Appl. Sci., 2 (1980), pp. 12–25.
[33] M. Yamamoto, *Stability, reconstruction and regularization for an inverse source hyperbolic problem by a control method*, Inverse Problems, 11 (1995), pp. 481–496.
[34] M. Yamamoto, *A mathematical aspect of inverse problems for non-stationary Maxwell's equations*, Internat. J. Appl. Electromagnetics and Mechanics, 8 (1997), pp. 77–98.
[35] M. Yamamoto, *An inverse problem of determining source terms in Maxwell's equations with a single measurement*, in Inverse Problems, Tomography and Image Processing, A. G. Ramm, ed., Plenum Press, New York, 1998, pp. 241–256.

SIAM J. CONTROL OPTIM.
Vol. 38, No. 4, pp. 1171–1203

A NEW SUBOPTIMAL APPROACH TO THE FILTERING PROBLEM FOR BILINEAR STOCHASTIC DIFFERENTIAL SYSTEMS*

FRANCESCO CARRAVETTA†, ALFREDO GERMANI†‡, AND MARAT K. SHUAKAYEV§

Abstract. The aim of this paper is to present a new approach to the filtering problem for the class of bilinear stochastic multivariable systems, consisting in searching for suboptimal state-estimates instead of the conditional statistics. As a first result, a finite-dimensional optimal linear filter for the considered class of systems is defined. Then, the more general problem of designing polynomial finite-dimensional filters is considered. The equations of a finite-dimensional filter are given, producing a state-estimate which is optimal in a class of polynomial transformations of the measurements with arbitrarily fixed degree. Numerical simulations show the effectiveness of the proposed filter.

Key words. square integrable martingales, wide-sense Wiener processes, stochastic bilinear systems, Kronecker algebra, Kalman–Bucy filtering, polynomial filtering, vector Ito formula

AMS subject classifications. 93E10, 93E11, 60H10

PII. S0363012997320912

1. Introduction.

Let us consider the class of nonlinear stochastic systems defined on some probability space, namely $(\Omega, \mathcal{F}, P)$, described by the Ito equations

$$dX(t) = A(t)X(t)dt + B^1(X(t), dW(t)), \tag{1.1}$$

$$dY(t) = C(t)\big(X(t)\big)dt + B^2(X(t), dW(t)), \tag{1.2}$$

where $X(t) \in \mathbf{R}^n$; $Y(t) \in \mathbf{R}^q$; $W(t) \in \mathbf{R}^p$ is a standard Wiener process with respect to some increasing family of σ-algebras, namely $\{\mathcal{F}_t\}$; $A(t), C(t)$ are matrices of proper dimensions; B^1 and B^2 are bilinear forms. System (1.1), (1.2) is commonly referred to in the literature as a *bilinear stochastic system* (BLSS) [4], [5], [6], [7], [8], [10].

The problem we are faced with consists in searching for finite-dimensional filters for the BLSS (1.1), (1.2). Indeed, for such a system even the *linear* optimal finite-dimensional filtering problem is still an interesting one.

With the name of finite-dimensional filter, we understand a stochastic differential equation in the form

$$dz(t) = f(z(t))dt + g(z(t))dY(t), \tag{1.3}$$

endowed with an output transformation

$$\hat{X}(t) = h(z(t)), \tag{1.4}$$

where $\{z(t),\ t > 0\}$ is some process taking values on a finite-dimensional linear space. We say that (1.3), (1.4) is a finite-dimensional optimal filter for system (1.1), (1.2) if

$$\hat{X}(t) = E(X(t)/\mathcal{F}_t^Y), \tag{1.5}$$

*Received by the editors May 8, 1997; accepted for publication (in revised form) September 21, 1999; published electronically April 18, 2000. This work was partially supported by the Progetto Finalizzato Trasporti 2 of the Italian National Research Council (CNR).

http://www.siam.org/journals/sicon/38-4/32091.html

†Istituto di Analisi dei Sistemi e Informatica del CNR, Viale Manzoni 30, 00185 Roma, Italy (carravetta@iasi.rm.cnr.it).

‡Dipartimento di Ingegneria Elettrica, Università dell'Aquila, 67100 Monteluco (L'Aquila), Italy (germani@ing.univaq.it).

§Department of Electrical Engineering, Kazakh National Technical University, 22 Satpaev Str., Almaty, Republic of Kazakhstan.

where we have denoted $\mathcal{F}_t^Y$ the σ-algebra generated by the observations $\{Y(s),\ 0 \leq s \leq t\}$.

As is well known, the optimal filter for system (1.1), (1.2) is an infinite-dimensional one. Nevertheless, from an application point of view, it becomes crucial to look for finite-dimensional approximations of the optimal filter.

In this paper we will derive, as an auxiliary result, the *optimal linear filtering* equations for a BLSS in the form of (1.1), (1.2) which will result in the finite-dimensional form (1.3), (1.4). We point out that in [3] the optimal linear filter is derived in the more general setting of linear stochastic equations driven by wide-sense Wiener (WSW) processes, resulting in a Kalman–Bucy scheme [1], [2]. Then, the optimal linear filter is defined for a scalar BLSS by representing the bilinear form as a WSW process. We will follow the same basic methodology in deriving the optimal linear filter for a *vector* BLSS.

Because of the infinite-dimensionality of the optimal filter for system (1.1), (1.2), it is of a great interest from an application point of view to search for finite-dimensional *suboptimal* filters showing a *better performance* with respect to the linear one.

This suboptimal approach has been recently developed for discrete-time systems in [9], [10], where a general *polynomial filter* of any arbitrarily fixed degree is defined for linear non-Gaussian systems [9] and bilinear systems [10]. The polynomial filter is able to produce, recursively, the optimal state-estimate in a class of polynomials of all the currently available measurements including the linear transformations. For this reason, in a non-Gaussian setting, it represents an improvement of the classical Kalman filtering. Indeed, many numerical simulations have shown that the improvement in performance may be very large especially when noise distributions are very far from Gaussianity.

In this paper we will propose this suboptimal approach for the filtering problem of continuous-time BLSSs. This will allow us to define a finite-dimensional filter in the form (1.3), (1.4), giving the optimal state-estimate in a suitably defined class of polynomial transformations of the measurements.

The program of the polynomial filtering methodology consists essentially in the following three steps.

(i) A class of polynomial estimators is defined.

(ii) The problem of finding the optimal filter for the BLSS in the above class of polynomial estimators is reduced to an optimal linear filtering problem for a suitable *augmented system*. The augmented system will result in a linear SDE with WSW diffusions. In particular, the state of the augmented system (augmented state) contains the original state, its Kronecker powers, and also Kronecker products with the observation process. The output of the augmented state (augmented observation) contains the original output process together with its Kronecker powers up to a fixed degree.

(iii) A Kalman–Bucy scheme is applied to the augmented system. This will give us the required polynomial filter.

The paper is organized as follows. Section 2 deals with point (i). In section 3, the overall setup of the problem is presented. Sections 4, 5, and 6 are concerned with some preliminary results. In particular, in section 4, a method for transforming a vector BLSS in a linear system with WSW diffusions is presented. In section 5 a vector Ito formula is defined by using the Kronecker formalism. In section 6, a general formula defining the stochastic differential of the Kronecker power of some process, solution of a bilinear SDE, is found. In section 7, point (ii) is treated. Finally, in

section 8, the complete solution of the problem is presented, resulting in a system of equations which defines a polynomial filter (of an arbitrarily fixed degree) for a BLSS. In section 9, numerical simulations are presented for a linear and third degree polynomial filter applied to a second order BLSS. A comparision is made with respect to the extended Kalman filter, which shows an unstable behavior for the presented case. Two appendices are included in order to make the paper more readable.

2. Suboptimal filtering. This section is devoted to the definition of the class of estimators considered in this paper. First of all, let us recall some results of linear filtering [3].

Let I be an interval (bounded or not) in the real line and consider a family $\{\xi_t,\ t \in I\}$ of L^2 random variables valued on some finite-dimensional euclidean space. For $t \in I$, let us define *the subspace* $\mathcal{L}_t(\xi) \subset L^2$ *linearly spanned by* $\{\xi_s,\ s \le t\}$ as the L^2-closure of the set $\mathcal{L}'_t(\xi)$:

$$\mathcal{L}'_t(\xi) \triangleq \Big\{ \lambda \in L^2 \ :\ \exists j \in \mathbf{N}, \exists t_1, \dots, t_j \in I,\ t_1 \le \cdots \le t_j \le t,$$
$$\exists \text{ matrices } M_{t_1}, \dots, M_{t_j}, \exists \text{ a vector } b, \text{ such that } \lambda = \sum_{i=1}^{j} M_{t_i} \xi_{t_i} + b \Big\}.$$

Let $\Pi(\cdot/\mathcal{L}_t(\xi))$ denote the orthogonal projection operator onto $\mathcal{L}_t(\xi)$. Then, for any given L^2 random variable η we can define the *optimal linear estimate of* η *given* $\{\xi_s,\ s \le t\}$ as $\Pi(\eta/\mathcal{L}_t(\xi))$. Now, suppose there exists an integer ν such that

$$E(\|\xi_t\|^{2\nu}) \le +\infty \quad \forall\ t \in I.$$

Let us denote by $X^{[i]}$ the ith Kronecker power of a vector X. We can give the following definition.

DEFINITION 2.1. *We call* ν*th degree polynomial estimate of* η *given* $\{\xi_s,\ s \le t\}$ *the random variable* $\Pi(\eta/\mathcal{P}_t^{(\nu)}(\xi))$*, where*

$$\mathcal{P}_t^{(\nu)}(\xi) \triangleq \mathcal{L}_t(\xi^{(\nu)})$$

and $\xi^{(\nu)}$ *is the process*

$$\xi^{(\nu)} \triangleq \begin{bmatrix} \xi^{[\nu]} \\ \xi^{[\nu-1]} \\ \vdots \\ \xi \\ 1 \end{bmatrix}.$$

From Definition 2.1 we see that $\Pi(\eta/\mathcal{P}_t^{(\nu)})$ is the mean square optimal estimate of η among all estimates, namely λ, that are either in the form

$$\lambda = \sum_{i,j=1}^{k} M_{i,j} \xi_{t_i}^{[j]} + b$$

for such a $k \in \mathbf{N}$, $t_1, \dots, t_k \in I$, $t_1 \le \cdots \le t_k$ for such a vector b and matrices $M_{i,j}$, $i, j = 1, \dots, k$, or are mean square limits of these. $\Pi(\eta/\mathcal{P}_t^{(\nu)})$ includes the linear estimates and, moreover,

$$\mathcal{P}_t^{(\nu)}(\xi) \subset \mathcal{P}_t^{(\nu+1)}(\xi) \quad \forall\ \nu \ge 1,$$

so that, for the polynomial estimates $\hat{\eta}^{(\nu)} = \Pi(\eta/\mathcal{P}_t^{(\nu)}(\xi))$, $\hat{\eta}^{(\nu+1)} = \Pi(\eta/\mathcal{P}_t^{(\nu+1)}(\xi))$ one has

$$E(\|\eta - \hat{\eta}^{(\nu+1)}\|^2) \leq E(\|\eta - \hat{\eta}^{(\nu)}\|^2) \quad \forall\ \nu \geq 1.$$

That is, the estimation quality is not decreasing for increasing ν.

Now, the aim of this paper can be expressed in a more precise manner as follows: for any given ν find a finite-dimensional filter in the form (1.3), (1.4) such that $\hat{X}(t)$ is the optimal νth degree polynomial estimate of the state of system (1.1), (1.2). Such a filter will be referred to in the following as a *νth degree polynomial filter*.

A crucial topic involved in the derivation of the polynomial filter is the linear estimation of stochastic processes generated by linear models driven by WSW processes, which we briefly describe below (see [3, Chap. 15], for a detailed discussion with proofs).

Let $\tilde{W}^{(i)}(t) \in \mathbf{R}^l$, $i = 1, \ldots, m$, be mutually uncorrelated WSW processes. Let us consider the linear stochastic system

$$\begin{aligned} dX(t) &= A(t)X(t)dt + \sum_{i=1}^{m} B_i(t)d\tilde{W}^{(i)}(t), \quad X(0) = \bar{X}, \\ dY(t) &= C(t)X(t)dt + \sum_{i=1}^{m} D_i(t)d\tilde{W}^{(i)}(t), \quad Y(0) = 0, \end{aligned} \tag{2.1}$$

where $t \in [0\ t_M]$, $X(t) \in \mathbf{R}^n$, $Y(t) \in \mathbf{R}^q$, $A(t), C(t), B_i(t), D_i(t)$, $i = 1, \ldots, m$, are suitably dimensioned matrices and $\bar{X}$ is a square integrable random vector. Model (2.1) can be interpreted as a continuous-time linear non-Gaussian system. We can consider the processes X, Y evolving in suitable L^2 spaces of square integrable random vectors. Let us denote with $\hat{X}(t)$ the optimal linear estimate of $X(t)$, that is $\hat{X}(t) = \Pi(X(t)/\mathcal{L}_t(Y))$. Then the following system of equations can be easily derived from [3, Thm. 15.3]:

(2.2)

$$\begin{aligned} d\hat{X}(t) &= A(t)\hat{X}(t)dts \\ &\quad + \left(\sum_{i=1}^{m} B_i(t)D_i(t)^T + P(t)C(t)^T\right) R(t)^{-1}(dY(t) - C(t)\hat{X}(t)dt), \\ \frac{dP(t)}{dt} &= A(t)P(t) + P(t)A(t)^T + Q(t) \\ &\quad - \left(\sum_{i=1}^{m} B_i(t)D_i(t)^T + P(t)C(t)^T\right) R(t)^{-1} \left(\sum_{i=1}^{m} B_i(t)D_i(t)^T + P_t(t)C(t)^T\right)^T, \\ \hat{X}(0) &= E(\bar{X}), \quad P(0) = E\Big((\bar{X} - E(\bar{X}))(\bar{X} - E(\bar{X}))^T\Big), \end{aligned}$$

where

$$R(t) \triangleq \sum_{i=1}^{m} D_i(t)D_i(t)^T; \quad Q(t) \triangleq \sum_{i=1}^{m} B_i(t)B_i(t)^T,$$

and $P(t)$ represents the filtering error covariance matrix. Note that in (2.2) the nonsingularity of the matrix function $R(t)$ over the time interval $[0\ t_M]$ is required.

As we will see in the next section, the BLSSs can be represented in the form (2.1). Then, (2.2) will allow us to obtain the optimal linear filter for a BLSS. This is a crucial point in the methodology here described. The way to derive the polynomial filter equations will consist indeed in reducing the original filtering problem to a linear one for a suitably defined BLSS.

3. The system to be filtered. Let $T = [0\ t_{_M}]$, let $(\Omega, \mathcal{F}, P)$ be a probability triple, and let $\{\mathcal{F}_t\}$, $t \in T$, be a family of nondecreasing sub-σ-algebras of $\mathcal{F}$. Moreover let $(W(t), \mathcal{F}_t)$ be an $\mathbf{R}^p$-valued standard Wiener process and $\bar{X} \in \mathbf{R}^n$ an $\mathcal{F}_0$-measurable random variable, independent of W, such that

$$E(\|\bar{X}\|^{2\nu}) < +\infty$$

for some integer $\nu \geq 1$. For the random variable $\bar{X}$ we suppose the moments, namely $m_{\bar{X}}^{(i)}$,

$$m_{\bar{X}}^{(i)} \triangleq E(\bar{X}^{[i]}), \quad i = 1, \ldots, 2\nu, \tag{3.1}$$

are known. Let us consider the stochastic system

$$\begin{aligned} dX(t) &= A(t)X(t)dt + H(t)u(t)dt \\ &\quad + \sum_{k=1}^{p} \big(B_k X(t) + F_k\big) dW_k(t), \quad X(0) = \bar{X}, \end{aligned} \tag{3.2}$$

$$\begin{aligned} dY(t) &= C(t)X(t)dt \\ &\quad + \sum_{k=1}^{p} \big(D_k X(t) + G_k\big) dW_k(t), \quad Y(0) = 0, \end{aligned} \tag{3.3}$$

where $A(t) \in \mathbf{R}^{n\times n}$, $C(t) \in \mathbf{R}^{q\times n}$, $H(t) \in \mathbf{R}^{n\times m}$, $B_k \in \mathbf{R}^{n\times n}$, $F_k \in \mathbf{R}^n$, $D_k \in \mathbf{R}^{q\times n}$, $G_k \in \mathbf{R}^q$, for $k = 1, \ldots, p$, $W_k(t)$ denotes the kth component of the standard Wiener process $W(t) \in \mathbf{R}^p$, and $u(t) \in \mathbf{R}^m$ is a deterministic input. Equation (3.2) is endowed with the initial condition $X(0) = \bar{X}$. In the following, we shall denote with I_α, $\alpha = 0, 1, \ldots$, the $\alpha \times \alpha$ identity matrix; we assume $I_0 = 1$. We make the following assumption on system (3.2), (3.3).

Assumption 3.1. There exists a $\bar{k}$, $1 \leq \bar{k} \leq p$, such that the matrix $D_{\bar{k}} D_{\bar{k}}^T$ is nonsingular.

Remark 3.2. Assumption 3.1 implies that we can assume, without loss of generality, that there exists a $\bar{k}$, $1 \leq \bar{k} \leq p$, such that

$$D_{\bar{k}} = [I_q\ 0]. \tag{3.4}$$

Indeed, let $\bar{k}$ be such that $D_{\bar{k}} D_{\bar{k}}^T$ is nonsingular, and define the matrix $T \in \mathbf{R}^{n\times n}$ as

$$T = \begin{bmatrix} D_{\bar{k}} \\ R \end{bmatrix},$$

where $R \in \mathbf{R}^{(n-q)\times n}$ is chosen such that the whole T results in a nonsingular matrix. It is easy to verify that $D_{\bar{k}} T^{-1} = [I_q\ 0]$. Hence we can always modify system (3.2), (3.3) by using T as a matrix performing a change of coordinates in the state space, and we can ensure that the representation (3.4) holds for at least one $\bar{k} \in \{1, \ldots, p\}$.

The problem we are faced with consists in finding a finite-dimensional filter in the form of (1.3), (1.4), such that

$$\widehat{X}(t) = \Pi\Big(X(t)/\mathcal{P}_t^{(\nu)}(Y)\Big), \tag{3.5}$$

where the space $\mathcal{P}_t^{(\nu)}(Y)$ is given by Definition 2.1.

As above mentioned (see point (ii) in the introduction), we will prove that there exists an augmented linear system for which the optimal linear filtering problem is equivalent to the original polynomial filtering problem for system (3.2), (3.3). To this purpose, in the next two sections we state some preliminary results.

4. Optimal linear filtering for BLSSs. Before treating the more general polynomial case, in this section we limit ourselves in considering the optimal linear filtering problem for the BLSS (3.2), (3.3). The reason for considering this particular case in advance is twofold. First of all, as we will see later, the polynomial case reduces to the linear one once a suitable augmented system has been constructed. Moreover, the optimal (finite-dimensional) *linear* filtering problem for a BLSS is interesting by itself, in that it was up to now unsolved in the general case [3]. In this section, we give a solution of this problem, in that we will prove the existence of a *linear* stochastic system with WSW diffusions, which is equivalent to the original BLSS (3.2), (3.3). Indeed, a version of the classical Kalman–Bucy theory [3] solves the optimal linear filtering problem in this case.

Let $M \in \mathbf{R}^{\alpha\times\alpha}$ be a symmetric positive semidefinite matrix, such that $\operatorname{rank}(M) = \rho \leq \alpha$. As is well known, there exists a full rank matrix $N \in \mathbf{R}^{\alpha\times\rho}$ such that $NN^T = M$. We will use the notation

$$M^{(\frac{1}{2})} \triangleq N,$$

that is, a "rectangular square root" of the matrix M. Note that, by definition, the matrix ${M^{(1/2)}}^T M^{(1/2)}$ is nonsingular.

Let ξ be a random vector; in the following we will use the notation $\operatorname{cov}(\xi,\xi) = E\big((\xi - E(\xi))(\xi - E(\xi))^T\big)$. Let us denote $m_X(t) = E(X(t))$, $\Psi_X(t) = \operatorname{cov}(X(t), X(t))$, where X is the state process of system (3.2), (3.3). Moreover, let us denote $\bar{m}_X = E(\bar{X})$ and $\bar{\Psi}_X = \operatorname{cov}(\bar{X}, \bar{X})$, where $\bar{X}$ is the initial state vector of (3.2).

THEOREM 4.1. *Let us consider the system* (3.2), (3.3). *Suppose that the matrix* $\Psi_X(t)$ *is nonsingular for any* $t \in T$. *Let us consider, for* $k = 1, \ldots, p$, *the integers* $\rho_k \leq n$, $\sigma_k \leq q$ *such that*

$$\begin{aligned} \rho_k &=^{\Delta} \operatorname{rank}\Big\{B_k \cdot \Psi_X(t) \cdot {B_k}^T\Big\} \\ \sigma_k &=^{\Delta} \operatorname{rank}\Big\{D_k \cdot \Psi_X(t) \cdot {D_k}^T\Big\} \end{aligned} \qquad \forall\ t \in T. \tag{4.1}$$

Then there exists the representation

$$dX(t) = A(t)X(t)dt + H(t)u(t) + \sum_{k=1}^{2p} \tilde{B}_k(t)d\tilde{W}_{k,1}(t), \quad X(0) = \bar{X}, \tag{4.2}$$

$$dY(t) = C(t)X(t)dt + \sum_{k=1}^{2p} \tilde{D}_k(t)d\tilde{W}_{k,2}(t), \qquad Y(0) = 0, \tag{4.3}$$

where, for $k = 1, \dots, p$: $\tilde{B}_k(t) \in \mathbf{R}^{n \times \rho_k}$ *and* $\tilde{D}_k(t) \in \mathbf{R}^{n \times \sigma_k}$ *are given by*

$$\tilde{B}_k(t) \triangleq \Big(B_k \cdot \Psi_X(t) \cdot B_k^T \Big)^{(\frac{1}{2})}, \tag{4.4}$$

$$\tilde{D}_k(t) \triangleq \Big(D_k \cdot \Psi_X(t) \cdot {D_k}^T \Big)^{(\frac{1}{2})} \tag{4.5}$$

for $k = p+1, \dots, 2p$:

$$\tilde{B}_k(t) \triangleq B_{k-p} E\big(X(t)\big) + F_{k-p}, \tag{4.6}$$

$$\tilde{D}_k(t) \triangleq D_{k-p} E\big(X(t)\big) + G_{k-p}. \tag{4.7}$$

For $i = 1, 2$, *the set* $\{\tilde{W}_{k,i},\ k = 1, \dots, 2p\}$ *is a set of* $2p$ *mutually uncorrelated standard WSW processes. In particular, for* $k = 1, \dots, p$, $\tilde{W}_{k,1}(t) \in \mathbf{R}^{\rho_k}$, $\tilde{W}_{k,2}(t) \in \mathbf{R}^{\sigma_k}$; *for* $k = p+1, \dots, 2p$:

$$\tilde{W}_{k,1}(t) = \tilde{W}_{k,2}(t) = W_{k-p}(t). \tag{4.8}$$

Proof. For $k = 1, \dots, p$, let us define the processes $\tilde{W}_{k,1}, \tilde{W}_{k,2}$ as

$$\tilde{W}_{k,1}(t) = \int_0^t \big(\tilde{B}_k(\tau)^T \tilde{B}_k(\tau) \big)^{-1} \tilde{B}_k(\tau)^T B_k\big(X(\tau) - m_X(\tau)\big) dW_k(\tau), \tag{4.9}$$

$$\tilde{W}_{k,2}(t) = \int_0^t \big(\tilde{D}_k(\tau)^T \tilde{D}_k(\tau) \big)^{-1} \tilde{D}_k(\tau)^T D_k\big(X(\tau) - m_X(\tau)\big) dW_k(\tau), \tag{4.10}$$

where $\tilde{B}_k, \tilde{D}_k$ are given by (4.4), (4.5). Let us show that $\tilde{W}_{k,i}$, $i = 1, 2$, are standard WSW processes. As a matter of fact, using well-known properties of the Ito integral and (4.4), it results, for $s < t$:

$$\begin{aligned} &E\big(\tilde{W}_{k,1}(t) \tilde{W}_{k,1}(s)^T\big) \\ &\quad = \int_0^s \big(\tilde{B}_k(\tau)^T \tilde{B}_k(\tau) \big)^{-1} \tilde{B}_k(\tau)^T \Big(B_k \Psi_X(\tau) B_k^T \Big) \tilde{B}_k(\tau) \big(\tilde{B}_k(\tau)^T \tilde{B}_k(\tau) \big)^{-1} \, d\tau \\ &\quad = \int_0^s \big(\tilde{B}_k(\tau)^T \tilde{B}_k(\tau) \big)^{-1} \tilde{B}_k(\tau)^T \big(\tilde{B}_k(\tau) \tilde{B}_k(\tau)^T \big) \cdot \tilde{B}_k(\tau) \big(\tilde{B}_k(\tau)^T \tilde{B}_k(\tau) \big)^{-1} \, d\tau \\ &\quad = I_{\rho_k} \cdot s. \end{aligned}$$

Similarly, taking again an $s < t$, it can be proved that

$$E\big(\tilde{W}_{k,2}(t) \tilde{W}_{k,2}(s)^T\big) = I_{\sigma_k} \cdot s,$$

and hence, since the Wiener's process components $W_1, \dots, W_p$, are mutually independent, we have that, for $i = 1, 2$, $\{\tilde{W}_{k,i},\ k = 1, \dots, p\}$ is a family of mutually independent (vector) WSW processes with identity covariance.

Now let us show that, for $k = 1, \dots, p$ (almost surely),

$$\tilde{B}_k(t) d\tilde{W}_{k,1}(t) = B_k\big(X(t) - m_X(t)\big) dW_k(t), \tag{4.11}$$

$$\tilde{D}_k(t) d\tilde{W}_{k,2}(t) = D_k\big(X(t) - m_X(t)\big) dW_k(t). \tag{4.12}$$

From the hypotheses the symmetric positive-definite matrix $\Psi(t)^{1/2}$ is well defined. Hence, for any $y(t) \in \mathbf{R}^n$ we can define $\bar{y}(t) \in \mathbf{R}^n$ such that $y(t) = \Psi_x(t) \bar{y}(t)$. Next, let us consider the decomposition $\bar{y}(t) = \bar{y}_1(t) + \bar{y}_2(t)$, where

$$\bar{y}_1(t) \in \mathcal{R}\big(\Psi_X(t)^{1/2} B_k^T\big), \quad \bar{y}_2(t) \in \Big\{ \mathcal{R}\big(\Psi_X(t)^{1/2} B_k^T\big) \Big\}^{\perp} = \mathcal{N}\big(B_k \Psi_X(t)^{1/2}\big), \tag{4.13}$$

where $\mathcal{N}(M)$, $\mathcal{R}(M)$ denote the null-space and the range, respectively, of a matrix M. Using (4.13) and choosing a $\bar{z}(t)$ such that $\bar{y}_1(t) = \Psi_X(t)^{1/2} B_k \bar{z}(t)$, we have

$$B_k y(t) = B_k \Psi_X(t)^{\frac{1}{2}} \bar{y}(t) = B_k \Psi_X(t)^{\frac{1}{2}} \bar{y}_1(t) = B_k \Psi_X(t) B_k^T \bar{z}(t) = \tilde{B}_k(t) \tilde{B}_k(t)^T \bar{z}(t),$$

where the definition of $\tilde{B}_k(t)$, given by (4.4) has been used. It follows that for any $y(t) \in \mathbf{R}^n$ there exists a $z(t) \in \mathbf{R}^{\rho_k}$ (indeed $z(t) = \tilde{B}_k(t)^T \bar{z}(t)$) such that

$$B_k y(t) = \tilde{B}_k(t) z(t) \quad \forall\, t \in T. \tag{4.14}$$

Then, for any $y(t)$ we have

$$\begin{aligned} \tilde{B}_k(t) \Big(\tilde{B}_k(t)^T \tilde{B}_k(t) \Big)^{-1} \tilde{B}_k(t)^T B_k y(t) &= \tilde{B}_k(t) \Big(\tilde{B}_k(t)^T \tilde{B}_k(t) \Big)^{-1} \tilde{B}_k(t)^T \tilde{B}_k(t) z(t) \\ &= \tilde{B}_k(t) z(t) = B_k y(t), \end{aligned}$$

from which, using the definition of $\tilde{W}_{k,1}$ given by (4.9), equality (4.11) follows. A similar argument can be used to prove (4.12).

Finally, by adding and subtracting the state-expectation $m_{_X}(t)$, in the bilinear forms of (3.2), (3.3) and taking into account (4.11), (4.12), we obtain the representation (4.2), (4.3). The thesis follows as soon as it is proven that, for $i = 1, 2$, $\tilde{W}_{k',i}(t)$ $(p+1 \le k' \le 2p)$ is uncorrelated with $\tilde{W}_{k'',i}(t)$ $(1 \le k'' \le p)$. As a matter of fact, from (4.8), for $p+1 \le k' \le 2p$, $k'' \ne k' - p$,

$$E\big(\tilde{W}_{k'',1}(t) \tilde{W}_{k',1}(t)^T\big) = E\big(\tilde{W}_{k'',1}(t) W_{k'-p}(t)^T\big) = 0,$$

and, for $k'' = k' - p$,

$$\begin{aligned} &E\big(\tilde{W}_{k'',1}(t) \tilde{W}_{k',1}(t)^T\big) = E\big(\tilde{W}_{k'',1}(t) W_{k''}(t)^T\big) \\ &= E\left(\int_0^t \big(\tilde{B}_{k''}(\tau)^T \tilde{B}_{k''}(\tau)\big)^{-1} \tilde{B}_{k''}(\tau)^T B_{k''} \big(X(\tau) - m_{_X}(\tau)\big) dW_{k''}(\tau) \cdot \int_0^t dW_{k''}(\tau) \right) \\ &= \int_0^t E\Big(\big(\tilde{B}_{k''}(\tau)^T \tilde{B}_{k''}(\tau)\big)^{-1} \tilde{B}_{k''}(\tau)^T B_{k''} \big(X(\tau) - m_{_X}(\tau)\big) \Big) d\tau = 0. \end{aligned}$$

In the same way, it is possible to show that $E\big(\tilde{W}_{k'',2}(t) \tilde{W}_{k',2}(t)^T\big) = 0$ for $p+1 \le k' \le 2p$. □

In the following theorem a sufficient condition will be given which guarantees the nonsingularity of $\Psi_X(t)$. Let us consider a time-invariant version of the BLSS given by (3.2), (3.3):

$$dX(t) = AX(t)dt + Hu(t)dt + \sum_{k=1}^{p} (B_k X(t) + F_k) dW_k(t), \quad X(t_0) = \bar{X}, \tag{4.15}$$

$$dY(t) = CX(t)dt + \sum_{k=1}^{p} (D_k X(t) + G_k) dW_k(t), \qquad Y(t_0) = 0, \tag{4.16}$$

where $t_0 \in \mathbf{R}$ is any "initial time." We suppose that system (4.15), (4.16) is well defined over the time interval $[t_0\ \infty)$.

THEOREM 4.2. *Let the matrix $\Psi_X(t_0)$ be nonsingular (or the pair (A, F_k) of the state equation* (4.15) *be controllable for at least one $k = 1, \dots, p$); then the state covariance matrix $\Psi_X(t)$ is nonsingular for any $t \ge t_0$, $(t > 0)$.*

Proof. Let us denote $\tilde{X}(t) = X(t) - m_X(t)$. Taking the expectations of (4.15), we have

$$dm_X(t) = Am_X(t)dt + Hu(t)dt, \qquad m_X(0) = \bar{m}_X.$$

Subtracting this from (4.15) results in

$$d\tilde{X}(t) = A\tilde{X}(t)dt + \sum_{k=1}^{p} B_k\tilde{X}(t)dW_k(t) + \sum_{k=1}^{p}(B_k m_X(t) + F_k)dW_k(t), \quad \tilde{X}(t_0) = \bar{X} - \bar{m}_X$$

or

$$\begin{aligned}\tilde{X}(t) = e^{A(t-t_0)}\tilde{X}(t_0) &+ \sum_{k=1}^{p}\int_{t_0}^{t} e^{A(t-\tau)}B_k\tilde{X}(\tau)dW_k(\tau) \\ &+ \int_{t_0}^{t} e^{A(t-\tau)}(B_k m_X(\tau) + F_k)dW_k(\tau).\end{aligned} \tag{4.17}$$

From (4.17) the following equation is easily recognized to hold for $\Psi_X(t)$:

$$\begin{aligned}\Psi_X(t) = e^{A(t-t_0)}\Psi_X(t_0)e^{A^T(t-t_0)} &+ \sum_{k=1}^{p}\int_{t_0}^{t} e^{A(t-\tau)}B_k\Psi_X(\tau)B_k^T e^{A^T(t-\tau)}d\tau \\ &+ \sum_{k=1}^{p}\int_{t_0}^{t} e^{A(t-\tau)}(B_k m_X(\tau) + F_k)(B_k m_X(\tau) + F_k)^T e^{A^T(t-\tau)}d\tau.\end{aligned} \tag{4.18}$$

The thesis follows by noting that the three terms in the right-hand side of (4.18) are at least symmetric nonnegative definite and, in particular, the nonsingularity of $\phi_X(t_0)$ implies the positive definiteness of the first term, whereas the hypothesis of controllability of (A, F_k) for some k implies the positive definiteness of the term $\int e^{A(t-\tau)}F_k F_k^T e^{A^T(t-\tau)}d\tau$. □

Remark 4.3. Note that, when theorem 4.2 holds with $t_0 < 0$, it results that, for any finite time-interval $T \subset [t_0 \ +\infty)$, the state-covariance has the property $\Psi_X(t) > \alpha \cdot I \ \forall \ t \in T$ (I denotes the identity) for some real number $\alpha > 0$, (it is *unifomly nonsingular in* T).

Now, we can state the following theorem, which defines the optimal linear filter for a BLSS.

THEOREM 4.4. *Let the time-invariant BLSS as defined in* (4.15), (4.16) *be given. Let the hypotheses of Theorem* 4.2 *be satisfied. Moreover, let us suppose that*

(H1) $\operatorname{rank}(D_k) = q$ *or* $\operatorname{rank}(G_k) = q$ *for some* k.

Then, with reference to the notations of section 2, *the optimal linear estimate of the state process* X, *namely* $\hat{X}$, *and the error covariance*

$$P(t) = E\big((X(t) - \hat{X}(t))(X(t) - \hat{X}(t))^T\big)$$

satisfy the following system of equations:

$$\begin{aligned}\frac{dm_X(t)}{dt} &= Am_X(t) + Hu(t), \qquad m(0) = \bar{m}, \\ \frac{d\Psi_X(t)}{dt} &= A\Psi_X(t) + \Psi_X(t)A^T + \sum_{k=1}^{p} B_k\Psi_X(t)B_k^T\end{aligned} \tag{4.19}$$

$$+\sum_{k=1}^{p}(B_k m_X(t)+F_k)(B_k m_X(t)+F_k)^T, \qquad \Psi_X(0)=\bar{\Psi}_X, \tag{4.20}$$

$$\tilde{B}_k(t)=\begin{cases}\left(B_k\cdot\Psi_X(t)\cdot B_k^T\right)^{(\frac{1}{2})}, & 1\le k\le p,\\ B_{k-p}m_X(t)+F_{k-p}, & p+1\le k\le 2p,\end{cases} \tag{4.21}$$

$$\tilde{D}_k(t)=\begin{cases}\left(D_k\cdot\Psi_X(t)\cdot D_k^T\right)^{(\frac{1}{2})}, & 1\le k\le p,\\ D_{k-p}m_X(t)+G_{k-p}, & p+1\le k\le 2p,\end{cases} \tag{4.22}$$

$$R(t)=\sum_{i=1}^{2p}\tilde{D}_i(t)\tilde{D}_i(t)^T, \tag{4.23}$$

$$\begin{aligned}\frac{dP(t)}{dt}&=AP(t)+P(t)A^T+R(t),\\ &-\left(\sum_{i=1}^{2p}\tilde{B}_i(t)\tilde{D}_i(t)^T+P(t)C^T\right)R(t)^{-1}\left(\sum_{i=1}^{2p}\tilde{B}_i(t)\tilde{D}_i(t)^T+P(t)C^T\right)^T,\end{aligned} \tag{4.24}$$

$$P(0)=\bar{\Psi}_X, \tag{4.25}$$

$$d\hat{X}(t)=A\hat{X}(t)dt+\left(\sum_{i=1}^{2p+1}\tilde{B}_i(t)\tilde{D}_i(t)^T+P(t)C^T\right)R(t)^{-1}(dY(t)-C\hat{X}(t)dt), \tag{4.26}$$

$$\hat{X}(0)=\bar{m}. \tag{4.27}$$

Proof. (4.19) readily derives by taking the expectations of both sides of (4.15). Moreover, (4.20) is easily obtained by differentiating (4.18). From Theorem 4.2 and Remark 4.3, $\Psi_X(t)$ is uniformly nonsingular in T. Then, we can apply Theorem 4.1 in order to put system (4.15), (4.16) in the form of a linear stochastic system with suitable WSW state and output diffusions, deriving from (4.2), (4.3). Note that such an equivalent system is a time-varying one even if it is derived from the time-invariant BLSS (4.15), (4.16). Now from (4.22), (4.23) it results that

$$R(t)\triangleq\sum_{k=1}^{p}D_k\Psi_X(t)D_k^T+\sum_{k=1}^{p}(D_k m_X(t)+G_k)(D_k m_X(t)+G_k)^T,$$

which is uniformly nonsingular in T, by the hypothesis (H1) (and possibly by the uniform nonsingularity of $\Psi_X(t)$). The thesis easily derives from an application of [3, Thm. 15.3] to the representation (4.2), (4.3). □

Remark 4.5. In the general case, when the BLSS is time-varying the uniform nonsingularity of $\Psi_X(t)$ cannot be guaranteed. Nevertheless, in all the cases of a nonsingular $\Psi_X(t)$, the equations of the optimal linear filter can be still derived using the representation given by Theorem 4.1. The resulting system of equations is formally similar to (4.19)–(4.27), but the constant parameters are replaced with the corresponding time-varying ones.

5. The vector Ito formula in the Kronecker formalism. In this section, by using a formalism derived from the Kronecker algebra, we present a new version of the Ito formula which has, with respect to the classical formulation, the advantage of

being much more compact and will allow us to calculate, for a given stochastic process ϕ, the stochastic differential of the process $\phi^{[h]}$, where $[h]$ is any integer Kronecker power.

Let $x \in \mathbf{R}^n$ and F be any C^2 function in $\mathbf{R}^{m\times p}$; we introduce the matrix $(d/dx)\otimes F(x)$, having dimensions $m \times (n \cdot p)$, defined as

$$\frac{d}{dx} \otimes F(x) \triangleq \Big[\frac{\partial F(x)}{\partial x_1} \ \cdots \ \frac{\partial F(x)}{\partial x_n}\Big], \tag{5.1}$$

where the operator d/dx is given by

$$\frac{d}{dx} \triangleq \Big[\frac{\partial}{\partial x_1} \ \cdots \ \frac{\partial}{\partial x_n}\Big]. \tag{5.2}$$

Note that in (5.1) the rules defining the Kroneker product between matrices (see Definition A.1) are formally satisfied, provided that the "multiplication" between the differential operator $\partial/\partial x_i$ and a matrix function $F(x)$ is conventionally defined as

$$\frac{\partial}{\partial x_i} \cdot F(x) = \frac{\partial F(x)}{\partial x_i},$$

where the right-hand side has the usual meaning. Similarly, we can define the operator:

$$\frac{d}{dx} \otimes \frac{d}{dx} \triangleq \Big[\frac{\partial^2}{\partial x_1^2} \ \frac{\partial^2}{\partial x_1 \partial x_2} \ \cdots \ \frac{\partial^2}{\partial x_n^2}\Big].$$

Also in this case the composition rule of the Kronecker product is satisfied, but the "multiplication" between the differential operators $\partial/\partial x_i$ and $\partial/\partial x_i$ had to be interpreted as resulting in the differential operator $\partial^2/\partial x_i \partial x_j$. In general, we will adopt the convention: the multiplication between a differential operator and a function F results in a function (the derivative of F), whereas the multiplication between two differential operators results in a differential operator (the second order differential operator). Obviously, this convention could be generalized in order to give a precise meaning to the quantity

$$\frac{d^{[h]}}{dx^{[h]}} \otimes F(x)$$

for any integer $h \geq 0$. However, in this paper we are concerned at most with second-order derivatives.

It is easy to recognize that for any matrix, namely M, and for any pair of differentiable matrix functions having suitable dimensions, namely $V(x)$ and $W(x)$, it results that

$$\frac{d}{dx} \otimes (V(x) \otimes W(x)) = \Big(\frac{d}{dx} \otimes V(x)\Big) \otimes W(x) + V(x) \otimes \Big(\frac{d}{dx} \otimes W(x)\Big), \tag{5.3}$$

$$\frac{d}{dx} \otimes (MW(x)) = M\Big(\frac{d}{dx} \otimes W(x)\Big). \tag{5.4}$$

Moreover, the following "associative" property holds:

$$\frac{d}{dx} \otimes \frac{d}{dx} \otimes F(x) = \Big(\frac{d}{dx} \otimes \frac{d}{dx}\Big) \otimes F(x) = \frac{d}{dx} \otimes \Big(\frac{d}{dx} \otimes F(x)\Big).$$

Using the above notation, we can prove the following lemma, which will be very useful in the following sections.

LEMMA 5.1. *For any integer $h \geq 1$ and $x \in \mathbf{R}^n$, it results that*

$$\frac{d}{dx} \otimes x^{[h]} = U_n^h (I_n \otimes x^{[h-1]}), \tag{5.5}$$

and for any $h > 1$,

$$\frac{d}{dx} \otimes \frac{d}{dx} \otimes x^{[h]} = O_n^h (I_{n^2} \otimes x^{[h-2]}), \tag{5.6}$$

where the matrices $C_{u,v}^T$, $u, v \in \mathbf{N}$, are the commutation matrices defined by Theorem A.3 *and*

$$U_n^h \triangleq \left(\sum_{\tau=0}^{h-1} C_{n,n^{h-1-\tau}}^T \otimes I_{n^\tau} \right), \quad O_n^h \triangleq \sum_{\tau=0}^{h-1} \sum_{s=0}^{h-2} (C_{n,n^{h-1-\tau}}^T \otimes I_{n^\tau})(I_n \otimes C_{n,n^{h-2-s}}^T \otimes I_n).$$

Proof. According to the definition of the differential operator (5.1) and using (5.3), we have

$$\begin{aligned} Q^h \triangleq \frac{d}{dx} \otimes x^{[h]} &= \frac{d}{dx} \otimes \left(x \otimes x^{[h-1]} \right) = I_n \otimes x^{[h-1]} + x \otimes \left(\frac{d}{dx} \otimes x^{[h-1]} \right) \\ &= I_n \otimes x^{[h-1]} + x \otimes Q^{(h-1)}, \end{aligned} \tag{5.7}$$

from which, using Theorem A.3, we obtain

$$\frac{d}{dx} \otimes x^{[h]} = \sum_{\tau=0}^{h-1} x^{[h-1-\tau]} \otimes I_n \otimes x^{[\tau]} = \sum_{\tau=0}^{h-1} C_{n,n^{h-1-\tau}}^T \left(I_n \otimes x^{[h-1-\tau]} \right) \otimes x^{[\tau]},$$

from which (5.5) follows, taking into account the property (A.3c).

Similarly, by exploiting (5.3), (5.5), and (A.3c), it results that

$$\begin{aligned} &\frac{d}{dx} \otimes \frac{d}{dx} \otimes x^{[h]} \\ &= \frac{d}{dx} \otimes \left(\left(\sum_{\tau=0}^{h-1} C_{n,n^{h-1-\tau}}^T \otimes I_{n^\tau} \right) \left(I_n \otimes x^{[h-1]} \right) \right) \\ &= \sum_{\tau=0}^{h-1} \left(C_{n,n^{h-1-\tau}}^T \otimes I_{n^\tau} \right) \left(\frac{d}{dx} \otimes \left(I_n \otimes x^{[h-1]} \right) \right) \\ &= \sum_{\tau=0}^{h-1} \left(C_{n,n^{h-1-\tau}}^T \otimes I_{n^\tau} \right) \left(I_n \otimes \left(\frac{d}{dx} \otimes x^{[h-1]} \right) \right) \\ &= \sum_{\tau=0}^{h-1} \left(C_{n,n^{h-1-\tau}}^T \otimes I_{n^\tau} \right) \left(I_n \otimes \left(\left(\sum_{s=0}^{h-2} C_{n,n^{h-2-s}}^T \otimes I_{n^s} \right) \left(I_n \otimes x^{[h-2]} \right) \right) \right) \\ &= \sum_{\tau=0}^{h-1} \sum_{s=0}^{h-2} \left(C_{n,n^{h-1-\tau}}^T \otimes I_{n^\tau} \right) \left(I_n \otimes \left(\left(C_{n,n^{h-2-s}}^T \otimes I_{n^s} \right) \left(I_n \otimes x^{[h-2]} \right) \right) \right) \\ &= \sum_{\tau=0}^{h-1} \sum_{s=0}^{h-2} \left(C_{n,n^{h-1-\tau}}^T \otimes I_{n^\tau} \right) \left(\left(I_n \otimes \left(C_{n,n^{h-2-s}}^T \otimes I_{n^s} \right) \right) \left(I_n \otimes \left(I_n \otimes x^{[h-2]} \right) \right) \right), \end{aligned}$$

so that the proof is completed. □

Now, we are able to rewrite the vector valued version of the Ito formula in the Kronecker formalism.

THEOREM 5.2. *Let $(X_t, \mathcal{F}_t)$ be a vector-continuous semimartingale in $\mathbf{R}^n$ described by the Ito stochastic differential*

$$dX_t = d\beta_t + dM_t, \tag{5.8}$$

where $(\beta_t, \mathcal{F}_t)$ is an almost surely continuous bounded variation process and $(M_t, \mathcal{F}_t)$ is a square integrable martingale. Let

$$F : \mathbf{R}^n \to \mathbf{R}^p$$

be a continuous function endowed with the first and second derivatives. Then the process $Z_t = F(X_t)$ is a square integrable semimartingale, whose differential is given by

$$dZ_t = \left(\frac{d}{dx} \otimes F(x)\right)_{x=X_t} dX_t + \frac{1}{2}\left(\frac{d}{dx} \otimes \frac{d}{dx} \otimes F(x)\right)_{x=X_t} (dM_t)^{[2]}, \tag{5.9}$$

with $(dM_t)^{[2]}$ denoting the associate quadratic variation process whose arguments are

$$(dM_t)^{[2]} = \begin{bmatrix} d < M_1, M_1 >_t \\ d < M_1, M_2 >_t \\ \vdots \\ d < M_n, M_n >_t \end{bmatrix}, \tag{5.10}$$

with obvious meaning of symbols [11, 12].

Proof. Formula (5.10) can be directly verified by using the Ito formula in the scalar case (see for instance [11, Thm. 4.2.1]) and by taking into account the definition of the differential operator d/dx. □

6. Stochastic differential for the Kronecker power of a BLSS solution. Using the Ito formula, in the version given by Theorem 5.2, we can now prove the following theorem, which defines the stochastic differential for the power process of the solution of a bilinear SDE. This will be the fundamental tool in the derivation of the augmented system.

THEOREM 6.1. *Let $\phi(t) \in \mathbf{R}^d$ be the process defined by the following SDE:*

$$d\phi(t) = (\Gamma(t)\phi(t) + \gamma(t))dt + \sum_{k=1}^{p} (\Theta_k \phi(t) + \chi_k) dW_k(t), \tag{6.1}$$

where $\Gamma(t), \Theta_k \in \mathbf{R}^{d\times d}$, $\gamma(t), \chi_k \in \mathbf{R}^d$. Then, defining

$$\Phi_2 \stackrel{\Delta}{=} \sum_{k=1}^{p} \Theta_k^{[2]}, \quad \Phi_1 \stackrel{\Delta}{=} \sum_{k=1}^{p} (\Theta_k \otimes \chi_k + \chi_k \otimes \Theta_k), \quad \Phi_0 \stackrel{\Delta}{=} \sum_{k=1}^{p} \chi_k^{[2]}, \tag{6.2}$$

it results for $i \geq 2$ that

$$\begin{aligned} d\phi^{[i]}(t) = &\left(\mathcal{M}_i^0(t)\phi^{[i]}(t) + \mathcal{M}_i^1(t)\phi^{[i-1]}(t) + \mathcal{M}_i^2\phi^{[i-2]}(t)\right) dt \\ &+ \sum_{k=1}^{p} \left(\mathcal{G}_{k,i}^0 \phi^{[i]}(t) + \mathcal{G}_{k,i}^1 \phi^{[i-1]}(t)\right) dW_k(t), \end{aligned}$$

where

$$\mathcal{M}_i^0(t) = U_d^i(\Gamma(t) \otimes I_{d^{i-1}}) + \frac{1}{2} O_d^i(\Phi_2 \otimes I_{d^{i-2}}),$$
$$\mathcal{M}_i^1(t) = U_d^i(\gamma(t) \otimes I_{d^{i-1}}) + \frac{1}{2} O_d^i(\Phi_1 \otimes I_{d^{i-2}}),$$
$$\mathcal{M}_i^2 = \frac{1}{2} O_d^i\big(\Phi_0 \otimes I_{d^{i-2}}\big),$$
$$\mathcal{G}_{k,i}^0 = U_d^i(\Theta_k \otimes I_{d^{i-1}}),$$
$$\mathcal{G}_{k,i}^1 = U_d^i(\chi_k \otimes I_{d^{i-1}}).$$

Proof. By using property (A.3c) the following formula is easily recognized to hold for any $k = 0, 1, \ldots$, $j = 1, 2, \ldots$, $\psi \in \mathbf{R}^\sigma$, $M \in \mathbf{R}^{r \times \sigma^k}$:

$$\left(I_r \otimes \psi^{[j]}\right) M \psi^{[k]} = (M \otimes I_{\sigma^j}) \psi^{[k+j]}. \tag{6.3}$$

Let us apply Theorem 5.2 for $X = \phi$, $F(\phi) = \phi^{[i]}$, $d\beta = (\Gamma\phi + \gamma)dt$, and $dM = d\Lambda$, where Λ is the martingale:

$$\Lambda(t) \triangleq \int_0^t \sum_{k=1}^p \big(\Theta_k(\tau)\phi(\tau) + \chi_k(\tau)\big) dW_k(\tau).$$

Using formulas (5.5), (5.6), it results that (understanding time dependencies)

$$d\phi^{[i]} = U_d^i \left(I_d \otimes \phi^{[i-1]}\right) (\Gamma\phi dt + \gamma dt + d\Lambda) + \frac{1}{2} O_d^i \left(I_{d^2} \otimes \phi^{[i-2]}\right) (d\Lambda)^{[2]}. \tag{6.4}$$

By exploiting the definition (5.10) it results that

$$(d\Lambda)^{[2]} = \big(\Phi^2 \phi_{[2]} + \Phi_1 \phi + \Phi_0\big)\, dt, \tag{6.5}$$

where Φ_2, Φ_1, and Φ_0 are given by (6.2). By substituting (6.5) in (6.4) and using formula (6.3), the thesis follows. ∎

7. The augmented system. Let us return to consider the BLSS (3.2), (3.3). In this section, by means of a repeated application of Theorem 6.1, we will show that the process (X, Y) and its powers up to a certain degree represent a solution of a suitably defined bilinear SDE. The latter will be next transformed into a linear system with WSW diffusions, generating the powers of the observation Y up to the required degree (the augmented system).

Let $x \in \mathbf{R}^d$ and h be a positive integer. We recall that the following relations hold, linking together the *reduced hth Kroneker power of* x [3], [15], namely $x_{[h]}$ and the (ordinary) hth Kronecker power $x^{[h]}$:

$$x^{[h]} = T_d^h x_{[h]}, \quad x_{[h]} = \tilde{T}_d^h x^{[h]}, \tag{7.1}$$

where T_d^h and $\tilde{T}_d^h$ are suitably dimensioned transformation matrices [3].

Let us define the process Z as

$$Z(t) \triangleq \begin{bmatrix} Y(t) \\ X(t) \end{bmatrix} \tag{7.2}$$

and let $\delta = \dim(Z)$. Moreover, let us define the augmented process:

$$\mathcal{Z}(t) \triangleq \begin{bmatrix} Z(t) \\ Z_{[2]}(t) \\ \vdots \\ Z_{[\nu]}(t) \end{bmatrix}. \tag{7.3}$$

We can derive an SDE for the process $\mathcal{Z}$ in the following way. First, note that, from (3.2), (3.3), Z satisfies the following SDE:

$$dZ(t) = \left(\tilde{A}(t)Z(t) + \alpha(t)\right)dt + \sum_{k=1}^{p}\left(B_k Z(t) + \tilde{\beta}_k\right)dW_k(t), \tag{7.4}$$

where

$$\tilde{A}(t) \triangleq \begin{bmatrix} 0 & C(t) \\ 0 & A(t) \end{bmatrix}; \quad \alpha(t) \triangleq \begin{bmatrix} 0 \\ Hu \end{bmatrix}; \quad B_k \triangleq \begin{bmatrix} 0 & D_k \\ 0 & B_k(t) \end{bmatrix}; \quad \beta_k = \begin{bmatrix} G_k \\ F_k \end{bmatrix}. \tag{7.5}$$

Next, by applying Theorem 6.1 to the process Z, it results for $i = 2, \dots, \nu$ that

$$\begin{aligned} dZ^{[i]}(t) = &\left(L_i^0(t)Z^{[i]}(t) + L_i^1(t)Z^{[i-1]}(t) + L_i^2 Z^{[i-2]}(t)\right)dt \\ &+ \sum_{k=1}^{p}\left(V_{k,i}^0 Z^{[i]}(t) + V_{k,i}^1 Z^{[i-1]}(t)\right)dW_k(t), \end{aligned} \tag{7.6}$$

where

$$L_i^0(t) = U_\delta^i\left(\tilde{A}(t) \otimes I_{\delta^{i-1}}\right) + \frac{1}{2}O_\delta^i\left(\Psi_2 \otimes I_{\delta^{i-2}}\right), \tag{7.7}$$

$$L_i^1(t) = U_\delta^{(i)}(\alpha(t) \otimes I_{\delta^{i-1}}) + \frac{1}{2}O_\delta^i\left(\Psi_1 \otimes I_{\delta^{i-2}}\right), \tag{7.8}$$

$$L_i^2 = \frac{1}{2}O_\delta^i\left(\Psi_0 \otimes I_{\delta^{i-2}}\right), \tag{7.9}$$

$$V_{k,i}^0 = U_\delta^i\left(\tilde{B}_k \otimes I_{\delta^{i-1}}\right), \tag{7.10}$$

$$V_{k,i}^1 = U_\delta^i(\beta_k \otimes I_{\delta^{i-1}}), \tag{7.11}$$

and Ψ_2, Ψ_1, and Ψ_0 are given by

$$\Psi_2 \triangleq \sum_{k=1}^{p}\tilde{B}_k^{[2]}, \quad \Psi_1 \triangleq \sum_{k=1}^{p}\left(\tilde{B}_k \otimes \beta_k + \beta_k \otimes \tilde{B}_k\right), \quad \Psi_0 \triangleq \sum_{k=1}^{p}\beta_k^{[2]}.$$

Observing that, from (7.1) we have

$$Z^{[i]} = T_\delta^i Z_{[i]}, \quad Z_{[i]} = \tilde{T}_\delta^i Z^{[i]},$$

and using (7.6), we can state the following proposition.

PROPOSITION 7.1. *The process $\mathcal{Z}$ defined in* (7.3) *satisfies the bilinear SDE*

$$d\mathcal{Z}(t) = \left(\mathcal{A}(t)\mathcal{Z}(t) + \mathcal{U}(t)\right)dt + \sum_{k=1}^{p}\left(\mathcal{B}_k\mathcal{Z}(t) + \mathcal{V}_k\right)dW_k(t), \tag{7.12}$$

where

$$\mathcal{A}(t) = \begin{bmatrix} \tilde{A}(t) & 0 & \dots & & 0 \\ L_2^1(t) & \tilde{T}_\delta^2 L_2^0(t) T_\delta^2 & 0 & & \\ L_3^2 & \tilde{T}_\delta^3 L_3^1(t) T_\delta^2 & \tilde{T}_\delta^3 L_3^0(t) T_\delta^3 & & \vdots \\ & \ddots & \ddots & \ddots & \\ 0 & \dots & \tilde{T}_\delta^\nu L_\nu^2 T_\delta^{\nu-2} & \tilde{T}_\delta^\nu L_\nu^1(t) T_\delta^{\nu-1} & \tilde{T}_\delta^\nu L_\nu^0(t) T_\delta^\nu \end{bmatrix}, \tag{7.13}$$

$$\mathcal{B}_k = \begin{bmatrix} \tilde{B}_k & 0 & \dots & & 0 \\ V_{k,2}^1 & \tilde{T}_\delta^2 V_{k,2}^0 T_\delta^2 & & & \vdots \\ 0 & \tilde{T}_\delta^3 V_{k,3}^1 T_\delta^2 & \tilde{T}_\delta^3 V_{k,3}^0 T_\delta^3 & & \\ \vdots & & \ddots & \ddots & \\ 0 & \dots & & \tilde{T}_\delta^\nu V_{k,\nu}^1 T_\delta^{\nu-1} & \tilde{T}_\delta^\nu V_{k,\nu}^0 T_\delta^\nu \end{bmatrix}, \tag{7.14}$$

$$\mathcal{U}(t) = \begin{bmatrix} \alpha(t) \\ L_2^2 \\ 0 \\ \vdots \\ 0 \end{bmatrix}, \quad \mathcal{V}_k = \begin{bmatrix} \beta_k \\ 0 \\ \vdots \\ 0 \end{bmatrix}. \tag{7.15}$$

The block matrices in (7.13), (7.14) *are given by* (7.7)–(7.11) *and* (7.5), *and the matrices* $\tilde{T}_\cdot, T_\cdot$, *are the reduction matrices defined in* (7.1).

Now, we can use Theorem 4.1 in order to rewrite the bilinear SDE (7.12) in the form of a *linear* SDE with WSW diffusion term. The underlying hypothesis is that the covariance matrix of the process $\mathcal{Z}$ defined in (7.3), namely $\Phi_Z(t)$, is uniformly nonsingular over T. There are many ways to assure this, starting from some suitable, nonrestrictive hypothesis on the original system. As a matter of fact, since we are here concerned with a finite interval T, it is easy to recognize that the uniform nonsingularity of $\Phi_Z(t)$ is assured as soon as it is assumed that the covariance of the initial original state $X(0)$ is positive definite. Henceforth, we will understand the uniform nonsingularity in T of $\Phi_Z(t)$.

PROPOSITION 7.2. *Let* ρ_k, $k = 1, \dots, p$, *be the ranks of the matrices* $\mathcal{B}_k$, *given in* (7.12). *Then the process* $\mathcal{Z}$ *satisfies the SDE*

$$d\mathcal{Z}(t) = \big(\mathcal{A}(t)\mathcal{Z}(t) + \mathcal{U}(t)\big)dt + \sum_{k=1}^{2p} \tilde{\mathcal{B}}_k(t) d\tilde{W}_k(t), \tag{7.16}$$

where $\tilde{W}_k$, $k = 1, \dots, 2p$ *are independent standard WSW processes,* $\tilde{W}_k \in \mathbf{R}^{\rho_k}$ *for* $k = 1, \dots, p$, $\tilde{W}_k = W_k \in \mathbf{R}$, *for* $k = p+1, \dots, 2p$, *and*

$$\tilde{\mathcal{B}}_k(t) \triangleq \begin{cases} \big(\mathcal{B}_k \Psi_Z(t) \mathcal{B}_k^T\big)^{(\frac{1}{2})}, & 1 \le k \le p, \\ \mathcal{B}_{k-p} m_Z(t) + \mathcal{V}_{k-p}, & p+1 \le k \le 2p, \end{cases} \tag{7.17}$$

with $m_Z = E(\mathcal{Z})$.

In order to write down the equations of the *augmented system* we need to split out the vector SDE (7.12) into two SDEs: one for the *observed* components of $\mathcal{Z}$ and the other one for the remaining entries.

From the definition (7.2) we see that the components of the vector $\mathcal{Z}$ are of the form

$$X_1^{i_1} \cdots X_n^{i_n} \cdots Y_1^{j_1} \cdots Y_q^{j_q}, \tag{7.18}$$

where X_l, Y_l denote the lth component of vectors X, Y, respectively, and $0 \leq i_l, j_r \leq \nu$ for $l = 1, \ldots, n$, $r = 1, \ldots, q$, $\sum_{l=1}^{n} i_l \leq \nu$, $\sum_{r=1}^{q} j_r \leq \nu$. The observed components are those of the form (7.18) with $i_1 = \cdots = i_n = 0$. Denote by $\mathcal{Y}$ the vector of all such components:

$$\mathcal{Y} \triangleq \begin{bmatrix} Y \\ Y_{[2]} \\ \vdots \\ Y_{[\nu]} \end{bmatrix}.$$

Moreover, let us denote by $\mathcal{E}_{\mathcal{Y}}$ the $(0,1)$-matrix such that

$$\mathcal{Y} = \mathcal{E}_{\mathcal{Y}} \mathcal{Z}. \tag{7.19}$$

It is easy to recognize that

$$\mathcal{E}_{\mathcal{Y}} = \begin{bmatrix} \mathcal{E}_{\mathcal{Y}}^1 & 0 & \ldots & 0 \\ 0 & \mathcal{E}_{\mathcal{Y}}^2 & \ddots & \\ \vdots & \ddots & \ddots & \\ 0 & \ldots & 0 & \mathcal{E}_{\mathcal{Y}}^{\nu} \end{bmatrix}, \tag{7.20}$$

where the diagonal blocks $\mathcal{E}_{\mathcal{Y}}^j$, $j = 1, \ldots, \nu$ are defined as

$$\mathcal{E}_{\mathcal{Y}}^j Z_{[j]} = Y_{[j]} \tag{7.21}$$

and have the expressions

$$\mathcal{E}_{\mathcal{Y}}^j = [I_q \;\; 0]^{[j]} T_{\delta}^j, \tag{7.22}$$

where T_{δ}^j is the expansion matrix defined in (7.1). Let us denote with $\mathcal{X}$ the aggregate vector of all the components in $\mathcal{Z}$ which are not components of $\mathcal{Y}$. Moreover, let us denote by $\mathcal{E}_{\mathcal{X}}$ the $(0,1)$-matrix such that

$$\mathcal{X} = \mathcal{E}_{\mathcal{X}} \mathcal{Z}. \tag{7.23}$$

A simple way to compute $\mathcal{E}_{\mathcal{X}}$ is just to remove from the identity matrix $I_{d_{\mathcal{Z}}}$ with $d_{\mathcal{Z}} = \dim(\mathcal{Z})$ (note that $I_{d_{\mathcal{Z}}}$ includes all the rows of $\mathcal{E}_{\mathcal{Y}}$) all those rows which are rows of $\mathcal{E}_{\mathcal{Y}}$.

From the above the aggregate matrix $\mathcal{I}$,

$$\mathcal{I} \triangleq \begin{bmatrix} \mathcal{E}_{\mathcal{Y}} \\ \mathcal{E}_{\mathcal{X}} \end{bmatrix} \tag{7.24}$$

results to be invertible. Let us consider the matrices $\mathcal{I}_1$, $\mathcal{I}_2$ such that

$$\mathcal{Z} = \mathcal{I}_1 \mathcal{Y} + \mathcal{I}_2 \mathcal{X}. \tag{7.25}$$

Note that, from (7.19), (7.23), and because of the invertibility of the matrix $\mathcal{I}$, it results that the matrices $\mathcal{I}_1, \mathcal{I}_2$ defined in (7.25) are obtained by means of a suitable partition of the matrix $\mathcal{I}^{-1} = [\mathcal{I}_1\ \mathcal{I}_2]$.

Using (7.19), (7.23), (7.25), and (7.12), we can now state the following proposition.

PROPOSITION 7.3. *The processes $\mathcal{X}$, $\mathcal{Y}$ defined in* (7.23) *and* (7.19) *satisfy the pair of SDEs (augmented system)*

$$d\mathcal{X}(t) = \big(\mathcal{A}_1(t)\mathcal{Y}(t) + \mathcal{A}_2(t)\mathcal{X}(t) + \mathcal{U}_1(t)\big)dt + \sum_{k=1}^{2p} \mathcal{B}_k^1(t)d\tilde{W}_k(t), \tag{7.26}$$

$$d\mathcal{Y}(t) = \big(\mathcal{C}_1(t)\mathcal{Y}(t) + \mathcal{C}_2(t)\mathcal{X}(t) + \mathcal{U}_2(t)\big)dt + \sum_{k=1}^{2p} \mathcal{D}_k^1(t)d\tilde{W}_k(t), \tag{7.27}$$

where

(7.28)

$$\mathcal{A}_1(t) = \mathcal{E}_{\mathcal{X}}\mathcal{A}(t)\mathcal{I}_1, \quad \mathcal{A}_2(t) = \mathcal{E}_{\mathcal{X}}\mathcal{A}(t)\mathcal{I}_2, \quad \mathcal{U}_1(t) = \mathcal{E}_{\mathcal{X}}\mathcal{U}(t), \quad \mathcal{B}_k^1(t) = \mathcal{E}_{\mathcal{X}}\tilde{\mathcal{B}}_k(t),$$
$$\mathcal{C}_1(t) = \mathcal{E}_{\mathcal{Y}}\mathcal{A}(t)\mathcal{I}_1, \quad \mathcal{C}_2(t) = \mathcal{E}_{\mathcal{Y}}\mathcal{A}(t)\mathcal{I}_2, \quad \mathcal{U}_2(t) = \mathcal{E}_{\mathcal{Y}}\mathcal{U}(t), \quad \mathcal{D}_k^1(t) = \mathcal{E}_{\mathcal{Y}}\tilde{\mathcal{B}}_k(t),$$

$\mathcal{A}$, $\tilde{\mathcal{B}}_k$, $\mathcal{U}$, are the matrix coefficients of (7.16)*, the matrices $\mathcal{E}_{\mathcal{X}}$, $\mathcal{E}_{\mathcal{Y}}$, $\mathcal{I}_1$, $\mathcal{I}_2$, are defined by means of* (7.19), (7.23), (7.25)*, and $\{\tilde{W}_k,\ k = 1, \dots, 2p\}$ is a set of mutually uncorrelated standard WSW processes.*

8. Polynomial filter equations. Proposition 7.3 states that the augmented observation process $\mathcal{Y}$ defined in (7.19) can be generated as the output process of the augmented representation (7.26), (7.27). This implies that the problem of finding the νth degree polynomial filter for the original system (7.26), (3.3) is now reduced to an *optimal linear filtering* problem for the linear system (7.26), (7.27). Indeed, by denoting with $\hat{\mathcal{X}}(t)$ the optimal linear estimate given $\{\mathcal{Y}_s,\ s \le t\}$ of the augmented state $\mathcal{X}(t)$, we have (see section 2)

$$\hat{\mathcal{X}}(t) = \Pi\Big(\mathcal{X}(t)/\mathcal{L}_t(\mathcal{Y})\Big).$$

On the other hand, from Definition 2.1 and taking into account the structure of the augmented observation $\mathcal{Y}$, it results that $\mathcal{L}_t(\mathcal{Y}) = \mathcal{P}_t^{(\nu)}(Y)$, where Y is the original observation process given by (3.3). Hence we have

$$\hat{\mathcal{X}}(t) = \Pi\Big(\mathcal{X}(t)/\mathcal{P}_t^{(\nu)}(Y)\Big),$$

and, as we will see later, we can get $\hat{X}(t)$ (which is given by (3.5)) by extracting a suitable subvector in $\hat{\mathcal{X}}(t)$.

In [3] the optimal linear filter is defined for the class of linear stochastic systems whose noise terms are represented by WSW processes. System (7.26), (7.27) comes within this class of systems, and we can use here the same approach as in [3] in order to obtain the optimal linear filter with respect to the augmented observation process $\mathcal{Y}$ (and, hence the optimal νth degree polynomial filter with respect to the original observed process Y). In order to do this, first of all we state the following theorem, whose proof is given in Appendix B, showing the uniform nonsingularity in T of the output-noise covariance of system (7.26), (7.27), namely

$$\mathcal{R}(t) \triangleq \sum_{k=1}^{2p} \mathcal{D}_k^1(t)\mathcal{D}_k^1(t)^T. \tag{8.1}$$

Indeed, the uniform nonsingularity of (8.1) is required, in order to apply the Kalman–Bucy scheme to system (7.26), (7.27).

THEOREM 8.1. *The noise covariance matrix function of the augmented measurement equation* (7.27)*, given by* (8.1)*, is uniformly nonsingular over* T.

Proof. See Appendix B. □

Now, we can prove the main theorem, defining the νth degree polynomial filter for system (3.2), (3.3). We remind readers that ρ_k is the dimension of the WSW process $\tilde{W}_k$ when $k = 1, \dots, p$, and for $k = p+1, \dots, 2p$, $\tilde{W}_k = W_k \in \mathbf{R}$. Let us denote with γ the dimension of the augmented process $\mathcal{Z}$. Moreover, we shall denote with $\operatorname{cov}(\chi, \eta)$ the cross-covariance between two random variables χ, η. Finally, we shall denote with $M^\dagger$ the Moore–Penrose pseudoinverse of the square matrix M.

THEOREM 8.2. *The νth order polynomial filter for system* (3.2), (3.3) *is described by the following system of equations:*

$$\frac{dm_{_Z}(t)}{dt} = \mathcal{A}(t)m_{_Z}(t) + \mathcal{U}(t), \tag{8.2}$$

$$\bar{\mathcal{B}}_k(t) = \mathcal{B}_k m_{_Z}(t) + \mathcal{V}_k, \quad 1 \le k \le p, \tag{8.3}$$

$$\frac{d\Psi_Z(t)}{dt} = \mathcal{A}(t)\Psi_Z(t) + \Psi_Z(t)\mathcal{A}(t)^T + \sum_{k=1}^{p} \mathcal{B}_k \Psi_Z(t) \mathcal{B}_k^T + \sum_{k=1}^{p} \bar{\mathcal{B}}_k(t)\bar{\mathcal{B}}_k(t)^T, \tag{8.4}$$

$$\tilde{\mathcal{B}}_k(t) = \left(\mathcal{B}_k \Psi_Z(t) \mathcal{B}_k^T\right)^{(\frac{1}{2})}, \quad 1 \le k \le p, \tag{8.5}$$

$$\mathcal{J}(t) = \sum_{k=1}^{p} \mathcal{E}_{\mathcal{X}}\big(\tilde{\mathcal{B}}_k(t)\tilde{\mathcal{B}}_k(t)^T + \bar{\mathcal{B}}_k(t)\bar{\mathcal{B}}_k(t)^T\big)\mathcal{E}_{\mathcal{Y}}^T, \tag{8.6}$$

$$\mathcal{R}(t) = \sum_{k=1}^{p} \mathcal{E}_{\mathcal{Y}}\Big(\tilde{\mathcal{B}}_k(t)\tilde{\mathcal{B}}_k(t)^T + \bar{\mathcal{B}}_k(t)\bar{\mathcal{B}}_k(t)^T\Big)\mathcal{E}_{\mathcal{Y}}^T, \tag{8.7}$$

$$\mathcal{Q}(t) = \sum_{k=1}^{p} \mathcal{E}_{\mathcal{X}}\Big(\tilde{\mathcal{B}}_k(t)\tilde{\mathcal{B}}_k(t)^T + \bar{\mathcal{B}}_k(t)\bar{\mathcal{B}}_k(t)^T\Big)\mathcal{E}_{\mathcal{X}}^T, \tag{8.8}$$

$$\begin{aligned}\frac{d\mathcal{P}(t)}{dt} &= \mathcal{A}_2(t)\mathcal{P}(t) + \mathcal{P}(t)\mathcal{A}_2(t)^T + \mathcal{Q}(t) \\ &\quad - \Big(\mathcal{J}(t) + \mathcal{P}(t)\mathcal{C}_2(t)^T\Big)\mathcal{R}(t)^{-1}\Big(\mathcal{J}(t) + \mathcal{P}(t)\mathcal{C}_2(t)^T\Big)^T,\end{aligned} \tag{8.9}$$

$$\begin{aligned}d\hat{\mathcal{X}}(t) &= \Big(\mathcal{A}_1(t)\mathcal{Y}(t) + \mathcal{A}_2(t)\hat{\mathcal{X}}(t) + \mathcal{U}_1(t)\Big)dt + \Big(\mathcal{J}(t) + \mathcal{P}(t)\mathcal{C}_2(t)^T\Big)\mathcal{R}(t)^{-1} \\ &\quad \cdot\big(d\mathcal{Y}(t) - \big(\mathcal{C}_1(t)\mathcal{Y}(t) + \mathcal{C}_2(t)\hat{\mathcal{X}}(t) + \mathcal{U}_2(t)\big)dt\big),\end{aligned} \tag{8.10}$$

$$\hat{X}(t) = \mathcal{T}_\nu \hat{\mathcal{X}}(t), \tag{8.11}$$

where $\mathcal{T}_\nu$ is the operator extracting the first n entries of a vector, the matrices $\mathcal{A}(t)$, $\mathcal{U}(t)$, $\mathcal{A}_1(t), \mathcal{A}_2(t)$, $\mathcal{B}_1(t), \mathcal{B}_2(t), \mathcal{U}_1(t), \mathcal{U}_2(t)$ are defined in (7.16) *and* (7.28)*, the matrices $\mathcal{B}_k$ are defined in* (7.14)*, $\rho_k = \operatorname{rank}(\mathcal{B}_k)$, and* (8.2), (8.4), (8.9), (8.10) *are endowed with the initial conditions*

$$\begin{aligned}m_{_Z}(0) &= E(\mathcal{X}(0)), \\ \Psi_Z(0) &= \operatorname{cov}(\mathcal{X}(0), \mathcal{X}(0)), \\ \hat{\mathcal{X}}(0) &= E(\mathcal{X}(0)) + \operatorname{cov}(\mathcal{X}(0), \mathcal{Y}(0))\operatorname{cov}(\mathcal{Y}(0), \mathcal{Y}(0))^\dagger\big(\mathcal{Y}(0) - E(\mathcal{Y}(0))\big), \\ \mathcal{P}(0) &= \operatorname{cov}(\mathcal{X}(0), \mathcal{X}(0)) - \operatorname{cov}(\mathcal{X}(0), \mathcal{Y}(0))\operatorname{cov}(\mathcal{Y}(0), \mathcal{Y}(0))^\dagger \operatorname{cov}(\mathcal{X}(0), \mathcal{Y}(0))^T.\end{aligned}$$

Proof. Equations (8.6)–(8.10) easily derive from an application of [3, Thm. 15.3] to the representation (7.26), (7.27). The augmented-state covariance $\Psi_Z(t)$, appearing in the definition of $\tilde{\mathcal{B}}_k$ given by (8.5) (see also (7.17)), satisfies the ODE (8.4). This can be readily proved in the same way of (4.20), but the time-varying BLSS (7.12) is considered now, and the semigroup generated by $\{\mathcal{A}(t),\ t \in T\}$ should be used.

The so-obtained estimate $\hat{\mathcal{X}}_t$ is the optimal one among all the linear transformation of the augmented observation process $\{\mathcal{Y}_s,\ s \le t\}$, and hence it is the νth degree polynomial estimate of the augmented state $\mathcal{X}_t$. In order to obtain the νth degree polynomial estimate of the state X_t of the original system (3.2), (3.3), first of all note that, because $\hat{\mathcal{X}}_t$ is the L^2-projection of $\mathcal{X}_t$ onto the closed subspace linearly spanned by $\{\mathcal{Y}_s,\ s \le t\}$, we have that each entry of $\hat{\mathcal{X}}_t$ agrees with the L^2-projection (onto the same subspace) of the corresponding entry in $\mathcal{X}_t$. Now, by definition, $\mathcal{X}(t)$ includes the components of the original state X_t. From (7.2), (7.3), and by the definition of the extracting operator $\mathcal{E}_{\mathcal{X}}$, it results that these components are placed in the first n entries of the vector $\mathcal{X}$. Hence, $\hat{X}(t)$ can be obtained simply by extracting the first n components of $\hat{\mathcal{X}}_t$, that is (8.11). □

9. Simulation example. In order to test the algorithm described in the previous sections, the filtering problem for the following second-order system has been considered:

$$dX(t) = AX(t)dt + BX(t)dW(t) + UdN(t), \quad X(0) = 0, \tag{9.1}$$

$$dY(t) = CX(t)dt + DX(t)dV(t), \qquad Y(0) = 0, \tag{9.2}$$

where

$$A = \begin{bmatrix} a_1 & 1 \\ 0 & a_2 \end{bmatrix}, \quad B = \begin{bmatrix} b_1 & 0 \\ 0 & b_2 \end{bmatrix}, \quad U = \begin{bmatrix} u_1 \\ u_2 \end{bmatrix},$$
$$C = [\,1 \quad 1\,], \qquad D = [\,g \quad 0\,],$$

and W, N, and V are mutually independent scalar Wiener processes.

The well-known extended Kalman filter (EKF) was up to now the classical tool for the filtering of a nonlinear system in the form of (9.1), (9.2). However, nothing is known about the working conditions or the performances of the EFK. In the present case, for instance, the EKF does not work at all. Indeed the state-expectation is zero and hence the state process is expected to cross the zero. Since the term $D\hat{X}(t)\hat{X}(t)^T D^T$ ($\hat{X}$ denoting the EKF estimation) needs to be inverted in the EKF equations, we should expect a failure of the algorithm. This really happens in the simulations we carried out for several values of the parameters $a_1, a_2, b_1, b_2, u_1, u_2, g$. We have always observed a sudden and strong deviation to infinity. In order to improve the working conditions we have substituted the term $D\hat{X}(t)\hat{X}(t)^T D^T$ with $\epsilon + D\hat{X}(t)\hat{X}(t)^T D^T$, where the number ϵ has been chosen small enough. In these cases we have observed an improvement of the algorithm, in that for a small initial time-interval the EKF shows a good performance, even better than the third-degree polynomial filter below described (no theoretical argument is known about this). However, unavoidably, the EKF diverges in spite of the trick used, whereas the polynomial filters continue to work.

The linear, the second degree (quadratic), and the third degree (cubic) filters have been built up by using (8.2)–(8.11). We remind the reader that the matrices $\mathcal{A}$, $\mathcal{U}$, $\mathcal{B}_k$, and $\mathcal{V}_k$ that appear in the filter equations are the system-matrices of the augmented (7.12). These matrices can be obtained from the original system-matrices

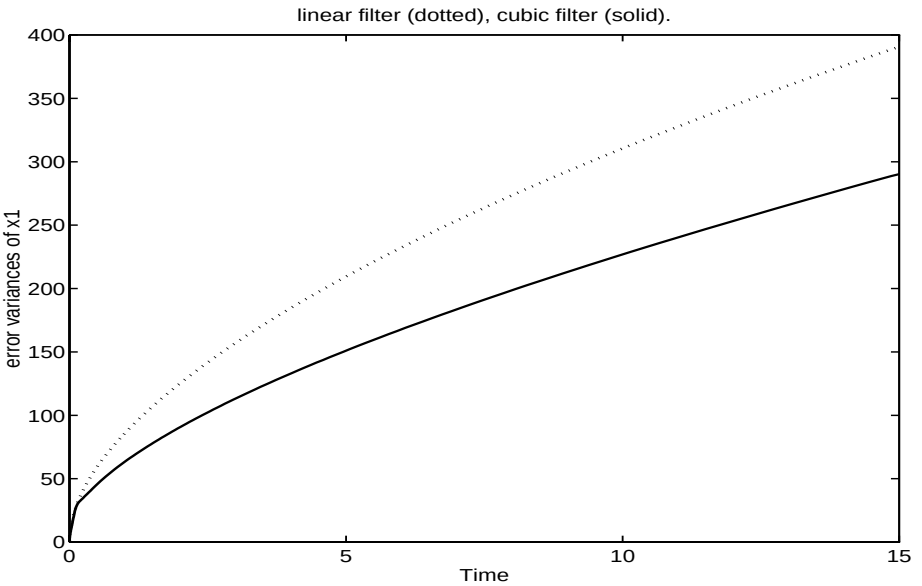


FIG. 9.1

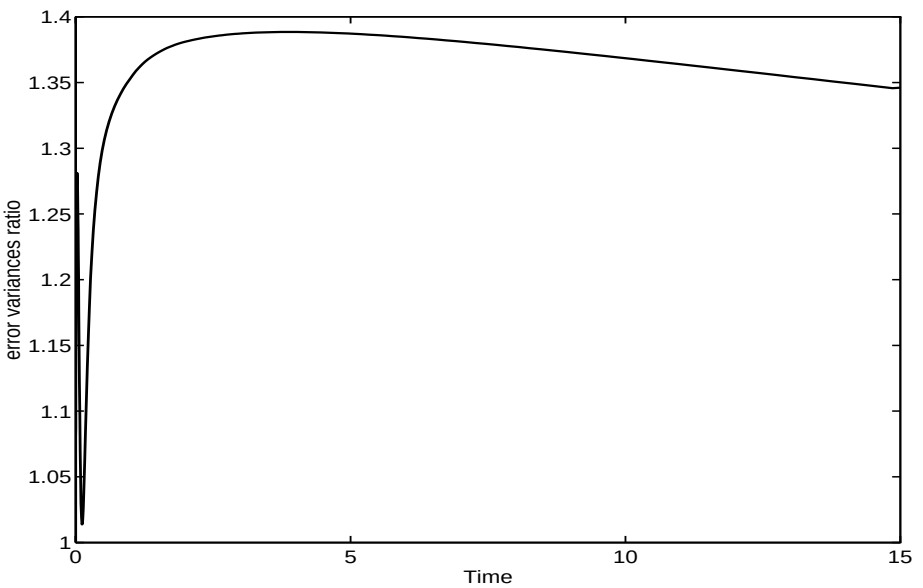


FIG. 9.2

by using the formulas given in section 7 for any polynomial degree. We show below our simulation results for the linear and cubic filters, with the following values of the parameters:

$$a_1 = -0.01, \quad a_2 = -0.5, \quad u_1 = 30, \quad u_2 = 2, \\ b_1 = 0.1, \quad b_2 = 0.1, \quad g = 0.1. \tag{9.3}$$

We do not show graphs related to the quadratic filter simulation because, in our case,

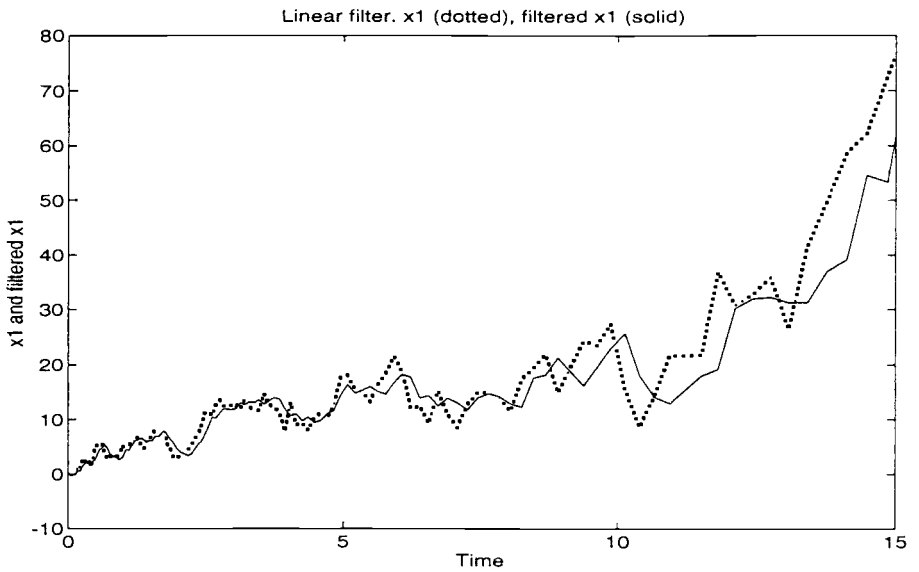


FIG. 9.3

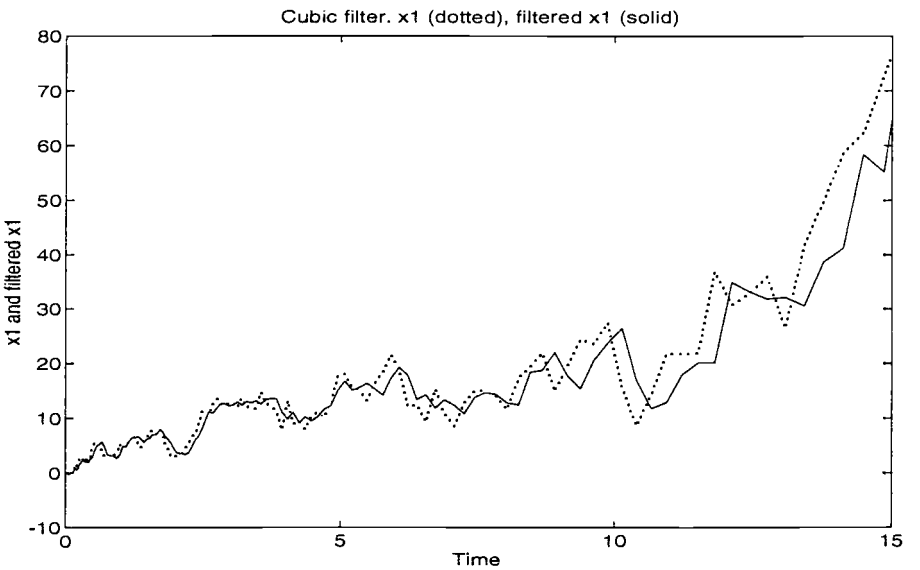


FIG. 9.4

the quadratic filter does not show any valuable improvement with respect to the linear case. Differently from the EKF, for the polynomial filters we are able to compute the a priori state-estimate error variances that are entries of the matrix $\mathcal{P}(t)$ given by (8.9), that is $\mathcal{P}_{1,1}(t)$, $\mathcal{P}_{2,2}(t)$ for the first and second state components, respectively. In our example these values are growing with time. The reason for this is that system (9.1), (9.2), with the values given by (9.3), is unstable. Nevertheless, as shown in Figure 9.1, the time-evolution of $\mathcal{P}_{1,1}(t)$ for the cubic filter (namely $P_C(t)$) is ever less than the $\mathcal{P}_{1,1}(t)$ for the linear filter (namely $P_L(t)$). In Figure 9.2 the evolution of

the ratio $\rho(t) = P_L(t)/P_C(t)$ is shown. We can see that $\rho(t)$ stabilizes over the value $\bar{\rho} = 1.30$. Hence the improvement in the a priori performance of the cubic filter with respect to the linear one can be considered almost 30%.

The time-evolutions of filtered paths for the linear and cubic filters, compared with the corresponding true 1st component state path, are reported in Figures 9.3, 9.4. As we can see, even a visual comparison between the signal time-evolutions shows a valuable improvement in the estimation quality of the cubic filter with respect to the linear one. Several Monte Carlo runs have been carried out. For each one of these runs, the ratio, namely ρ_S, between the sampled error variances of the linear and cubic filters has been computed. We have chosen the paths with $\rho_S = 1.35$.

The simulation of the EKF confirms also in this case its unsatisfactory behavior. Indeed, after almost 0.01 time units the EKF estimate starts up and quickly goes to infinity.

All the simulations have been carried out using the standard functions of the Matlab software package for Windows. The computer was a PC, endowed with a 200 MHz Pentium processor.

10. Conclusions. Equations (8.2)–(8.11) define a finite-dimensional filter for the BLSS (3.2), (3.3) which is optimal in a class of polynomial estimates. Although the considered class does not include all the polynomials of the currently available measurements, it includes the linear estimates, and, moreover, it defines a nondecreasing sequence of spaces for increasing polynomial degree. This implies that the polynomial filter had to improve the estimation performance for increasing polynomial degree.

We underline that the proposed filter is finite-dimensional. Of course, it is always possible to approximate the optimal filter (for instance, by applying a finite-elements method to the Zakai equation, as shown in [13]) with an arbitrary approximation degree. However, the more accurately the approximation level is chosen, the heavier the computational burden of the algorithm is. The computational effort is prohibitive even for small approximation degrees. Moreover, it makes no sense, within this approach, to use a large approximation degree in order to make the filtering algorithm really implementable. Counterwise, our suboptimal approach allows us to get meaningful estimates also for small polynomial degrees, which do not present difficult implementation problems.

In section 4 we have presented the equations of the optimal linear filter for a BLSS. We highlight that this result is interesting by itself in that it was up to now known only for the scalar case. The main tool is given by Theorem 4.1, stating the existence of a linear representation for a general vector BLSS. The optimal linear filter is then obtained by an application of a classical Kalman–Bucy scheme. Nevertheless, in the framework of this paper, the main purpose of Theorem 4.1 remains its application to the bilinear SDE (7.12), which allows us to obtain the linear representation (7.16).

Theorem 8.1 states that the output noise covariance of the augmented system is uniformly nonsingular, as required by the Kalman–Bucy scheme, provided that the output noise covariance of the original system (3.2), (3.3) is nonsingular. The proof is presented in Appendix B.

In section 9 a numerical simulation is shown, where a second-order BLSS has been filtered using the polynomial filters up to the third degree. The EFK has been also simulated, however its performance is resulted to be unsatisfactory. The simulation results show that the estimation quality really improves as polynomial degree grows, and for the cubic filter we obtained an improvement valuable over 30% with respect

to the linear filter.

We stress that, due to the well-known approximation capabilities of the polynomial functions, with the aim to define better and better implementable approximation schemes of the optimal filter, the use of polynomial estimators appears to be very promising.

Appendix A. Kronecker algebra.

Throughout this paper, we have widely used Kronecker algebra [14], [15]. Here, for the sake of completeness, we recall some definitions and properties on this subject.

DEFINITION A.1. *Let M and N be matrices of dimension $r \times s$ and $p \times q$, respectively. Then the Kronecker product $M \otimes N$ is defined as the $(r \cdot p) \times (s \cdot q)$ matrix*

$$M \otimes N = \begin{bmatrix} m_{11}N & \dots & m_{1s}N \\ \dots\dots\dots\dots\dots & & \\ m_{r1}N & \dots & m_{rs}N \end{bmatrix},$$

where the m_{ij} are the entries of M.

Of course this kind of product is not commutative.

DEFINITION A.2. *Let M be the $r \times s$ matrix*

$$M = [\, m_1 \quad m_2 \quad \dots \quad m_s \,], \tag{A.1}$$

where m_i denotes the ith column of M, and then the stack of M is the $r \cdot s$ vector

$$st(M) = [\, m_1^T \quad m_2 \quad \dots \quad m_s \,]^T. \tag{A.2}$$

Observe that a vector as in (A.2) can be reduced to a matrix M as in (A.1) by considering the inverse operation of the stack denoted by st^{-1}. With reference to the Kronecker product and the stack operation, the following properties hold [15]:

$$(A+B) \otimes (C+D) = A \otimes C + A \otimes D + B \otimes C + B \otimes D, \tag{A.3a}$$

$$A \otimes (B \otimes C) = (A \otimes B) \otimes C, \tag{A.3b}$$

$$(A \cdot C) \otimes (B \cdot D) = (A \otimes B) \cdot (C \otimes D), \tag{A.3c}$$

$$(A \otimes B)^T = A^T \otimes B^T, \tag{A.3d}$$

$$st(A \cdot B \cdot C) = (C^T \otimes A) \cdot st(B), \tag{A.3e}$$

$$u \otimes v = st(v \cdot u^T), \tag{A.3f}$$

$$tr(A \otimes B) = tr(A) \cdot tr(B), \tag{A.3g}$$

where A, B, C, and D are suitably dimensioned matrices, u and v are vectors, and $tr(M)$ denotes the trace of a square matrix M. The Kronecker power of the matrix M is defined as

$$M^{[0]} = 1,$$
$$M^{[n]} = M \otimes M^{[n-1]} = M^{[n-1]} \otimes M, \quad n > 0.$$

As an easy consequence of (A.3b) and (A.3g), it follows that

$$tr(A^{[h]}) = \big(tr(A)\big)^h. \tag{A.3h}$$

It is easy to verify that for $u \in \mathbf{R}^r$, $v \in \mathbf{R}^s$, the ith entry of $u \otimes v$ is given by

$$(u \otimes v)_i = u_l \cdot v_m; \quad l = \left[\frac{i-1}{s}\right] + 1, \quad m = |i-1|_s + 1, \tag{A.4}$$

where $[\cdot]$ and $|\cdot|_s$ denote integer part and s-modulo, respectively. Although the Kronecker product is not commutative, the following property holds [9, 15].

THEOREM A.3. *For any given pair of matrices $A \in \mathbf{R}^{r\times s}$, $B \in \mathbf{R}^{n\times m}$, we have*

$$B \otimes A = C_{r,n}^T (A \otimes B) C_{s,m}, \tag{A.5}$$

where the commutation matrix $C_{u,v}$ is the $(u \cdot v) \times (u \cdot v)$ matrix such that its (h,l) entry is given by

$$\{C_{u,v}\}_{h,l} = \begin{cases} 1 & \text{if } l = (|h-1|_v)u + \left(\left[\frac{h-1}{v}\right] + 1\right), \\ 0 & \text{otherwise.} \end{cases} \tag{A.6}$$

Observe that $C_{1,1} = 1$; hence in the vector case when $a \in \mathbf{R}^r$ and $b \in \mathbf{R}^n$, (A.5) becomes

$$b \otimes a = C_{r,n}^T (a \otimes b). \tag{A.7}$$

COROLLARY A.4. *For any given matrices A, B, C, D, having dimensions $n_A \times m_A$, $n_B \times m_B$, $n_C \times m_C$, $n_D \times m_D$, respectively, denoted with $I(l)$, the identity matrix in $\mathbf{R}^l$, we have*

$$A \otimes B \otimes C \otimes D = \left(I(n_A) \otimes C^T_{n_C n_D, n_B}\right)(A \otimes C \otimes D \otimes B)\left(I(m_A) \otimes C_{m_C m_D, m_B}\right).$$

Proof. See [3]. □

Appendix B. Proof of Theorem 8.1. We need to state in advance some preliminary definitions and lemmas.

Let δ and j be two positive integers.

DEFINITION B.1. *Let $r, s \in \{1, 2, \ldots, \delta^j\}$. The pair (r,s) is said to be (δ, j)-redundant ((δ, j)-R for short) if $\forall\ x \in \mathbf{R}^\delta$, it results that $\left(x^{[j]}\right)_r = \left(x^{[j]}\right)_s$, where $\left(x^{[j]}\right)_l$ denotes the lth entry of the vector $x^{[j]}$. Otherwise, the pair (r,s) is said to be (δ, j)-nonredundant ((δ, j)-NR for short).*

Example B.2. The pair $(2,3)$ is $(2,2)$-R; however, it is $(3,2)$-NR. The pairs $(1,1)$, $(2,2), \ldots$ are (δ, j)-NR for any δ and j.

Remark B.3. Let $x \in \mathbf{R}^\delta$. For some $s, r \in \{1, 2, \ldots, \delta^j\}$ let us consider the multi-indexes $s_1, \ldots, s_j$ and $r_1, \ldots, r_j$ in $\{1, \ldots, \delta\}$ defined by the identities

$$\left(x^{[j]}\right)_s = x_{s_1} x_{s_2} \cdots x_{s_j}, \quad \left(x^{[j]}\right)_r = x_{r_1} x_{r_2} \cdots x_{r_j}.$$

Then, we immediately realize that (r,s) is (δ, j)-R if and only if there exists a permutation of indexes transforming $s_1, \ldots, s_j$ in $r_1, \ldots, r_j$ (and vice versa).

Remark B.4. It is easy to verify that the (δ, j)-R condition defines an equivalence relation in the set $\{1, 2, \ldots, \delta^j\}$. We shall denote with $\rho(s; \delta, j)$ the equivalence class generated by $s \in \{1, \ldots, \delta^j\}$ via the (δ, j)-R relation

$$\rho(s; \delta, j) \triangleq \left\{ r \in \mathbf{N} \ :\ 1 \le r \le \delta^j,\ (s,r)\ \textit{is}\ (\delta, j)\text{–R} \right\}. \tag{B.1}$$

We shall denote with δ_j the number of equivalence classes of the (δ, j)-R relation, partitioning the set $\{1, 2, \ldots, \delta^j\}$. Moreover, we introduce the sets $\rho'(s; \delta, j)$, $\rho''(s; \delta, j) \subset \rho(s; \delta, j)$ defined as

$$\rho'(s;\delta,j) \triangleq \left\{ i \in \rho(s;\delta,j) \;\Big/\; \left[\frac{i}{\delta^{j-1}}\right] = \left[\frac{s}{\delta^{j-1}}\right] \right\}, \tag{B.2}$$

$$\rho''(s;\delta,j) \triangleq \rho(s;\delta,j) \setminus \rho'(s;\delta,j), \tag{B.3}$$

where we have used in (B.2) the notation $[\,\cdot\,]$ to indicate the integer part. The above defined sets have the following meaning. Let $x \in \mathbf{R}^\delta$ and note that

$$x^{[j]} = \begin{bmatrix} x_1 \cdot x^{[j-1]} \\ x_2 \cdot x^{[j-1]} \\ \vdots \\ x_\delta \cdot x^{[j-1]} \end{bmatrix}, \tag{B.4}$$

where every subvector $x_i x^{[j-1]}$ has dimension δ^{j-1}. By setting $l = [s/\delta^{j-1}]$ and observing in (B.4) the structure of $x^{[j]}$, we realize that the set defined in (B.2) is composed with the integers i such that (i, s) is (δ, j)-R and $\big(x^{[j]}\big)_i \in x_l x^{[j-1]}$. Counterwise, the set defined in (B.3) is composed with the integers i such that (i, s) is (δ, j)-R and $\big(x^{[j]}\big)_i$ does not belong to $x_l x^{[j-1]}$. Let us denote by $|n_1|_{n_2}$ the remainder of the integer division n_1/n_2. Then, again from (B.4), it is easily recognized that

$$\big(x^{[j]}\big)_s = x_l\big(x^{[j-1]}\big)_r, \quad r \triangleq |s|_{\delta^{j-1}}. \tag{B.5}$$

Remark B.5. Note that the number δ_j agrees with the number of entries of $x_{[j]}$ for $x \in \mathbf{R}^\delta$.

LEMMA B.6. *Let $r, s \in \{1, \ldots, \delta^{j-1}\}$ such that (r, s) is $(\delta, j-1)$-R. Then, for any $l = 0, 1, \ldots, \delta - 1$, the pair $(r + l\delta^{j-1}, s + l\delta^{j-1})$ is (δ, j)-R. Counterwise, if $r, s \in \{1, \ldots, \delta^j\}$ are (δ, j)-R and $r' = s'$ with*

$$r' \triangleq \left[\frac{r}{\delta^{j-1}}\right], \quad s' \triangleq \left[\frac{s}{\delta^{j-1}}\right],$$

then, denoting $r'' = |r|_{\delta^{j-1}}$, $s'' = |s|_{\delta^{j-1}}$, it results that (r'', s'') is $(\delta, j-1)$-R.

Proof. From Definition B.1 it results that

$$\big(x^{[j-1]}\big)_r = \big(x^{[j-1]}\big)_s \quad \forall\, x \in \mathbf{R}^\delta. \tag{B.6}$$

From (B.4) we see that

$$\big(x^{[j]}\big)_{r+l\delta^{j-1}} = x_l\big(x^{[j-1]}\big)_r, \quad \big(x^{[j]}\big)_{s+l\delta^{j-1}} = x_l\big(x^{[j-1]}\big)_s,$$

and hence, from (B.6),

$$\big(x^{[j]}\big)_{r+l\delta^{j-1}} = \big(x^{[j]}\big)_{s+l\delta^{j-1}}.$$

Counterwise, if $r, s \in \{1, \ldots, \delta^j\}$ are (δ, j)-R, then, taking into account (B.5), we have

$$\big(x^{[j]}\big)_r = x_{r'}\big(x^{[j-1]}\big)_{r''} = x_{s'}\big(x^{[j-1]}\big)_{s''} = \big(x^{[j]}\big)_s \quad \forall\, x \in \mathbf{R}^\delta. \tag{B.7}$$

Since, by hypothesis, $r' = s'$, (B.7) implies that $\left(x^{[j-1]}\right)_{r''} = \left(x^{[j-1]}\right)_{s''}$. □

Let $\mathcal{I} \subset \mathbf{N}$ and $n \in \mathbf{N}$. In the following, we will use the notation $\mathcal{I} - n$ to indicate the *translated* set:

$$\mathcal{I} - n = \{i \ / \ i \in \mathbf{N},\ \exists i' \in \mathcal{I},\ \text{such that } i = i' - n\}. \tag{B.8}$$

LEMMA B.7. *Suppose that*

$$\left[\frac{s}{\delta^{j-1}}\right] = l < \delta. \tag{B.9}$$

Then, for any $q < \delta - l$ it results that

$$\rho'(s; \delta, j) = \rho'(s + q\delta^{j-1}; \delta, j) - q\delta^{j-1},$$

where ρ' is the set defined in (B.2).

Proof. It suffices to show that for any $r \in \{1, \ldots, \delta^j\}$ such that $[r/\delta^{j-1}] = l$ and such that (r, s) is (δ, j)-NR ((δ, j)-R), the pair $(r + q\delta^{j-1}, s + q\delta^{j-1})$ is (δ, j)-NR ((δ, j)-R).

Suppose first that (r, s) is (δ, j)-NR. Let $x \in \mathbf{R}^\delta$ and $z = x^{[j]}$. From the structure (B.4) of the vector z and taking into account (B.9), we see that $z_s, z_r \in x_l \cdot x^{[j-1]}$. Hence since (s, r) is (δ, j)-NR, we have that, there exist integers $h_1, \ldots, h_\delta$ and $h'_1, \ldots, h'_\delta$, $h_1 + \cdots + h_\delta = h'_1 + \cdots + h'_\delta = j - 1$ such that it results that

$$\begin{aligned} z_s &= x_l \cdot x_1^{h_1} \cdots x_\delta^{h_\delta}, \\ z_r &= x_l \cdot x_1^{h'_1} \cdots x_\delta^{h'_\delta}. \end{aligned} \tag{B.10}$$

Since $z_r \neq z_s$ it follows that

$$x_1^{h_1} \cdots x_\delta^{h_\delta} \neq x_1^{h'_1} \cdots x_\delta^{h'_\delta}. \tag{B.11}$$

Again, looking at (B.4), we readily realize that

$$z_{s+q\delta^{j-1}} = x_{l+q} \cdot x_1^{h_1} \cdots x_\delta^{h_\delta}, \tag{B.12}$$

and

$$z_{r+q\delta^{j-1}} = x_{l+q} \cdot x_1^{h'_1} \cdots x_\delta^{h'_\delta}, \tag{B.13}$$

and hence, taking into account (B.11), it follows that $z_{s+q\delta^{j-1}} \neq z_{r+q\delta^{j-1}}$; that is $(s + q\delta^{j-1}, r + q\delta^{j-1})$ is (δ, j)-NR.

Next, suppose that (r, s) is (δ, j)-R. Then $z_s, z_r \in x_l \cdot x^{[j-1]}$, $z_s = z_r$, and by (B.10) it follows that $h_i = h'_i$, $i = 1, \ldots, \delta$. This in turn implies, taking into account (B.12), (B.13), that $z_{s+q\delta^{j-1}} = z_{r+q\delta^{j-1}}$; that is $(s + q\delta^{j-1}, r + q\delta^{j-1})$ is (δ, j)-R. □

LEMMA B.8. *Let (r, s) be a (δ, j)-R pair such that*

$$\left[\frac{r}{\delta^{j-1}}\right] = l, \qquad \left[\frac{s}{\delta^{j-1}}\right] = m, \quad l < m < \delta. \tag{B.14}$$

Then for any $q < \delta - l$ the pair $(r + q\delta^{j-1}, s + q\delta^{j-1})$ is (δ, j)-NR.

Proof. As in the proof of Lemma B.7 it is readily verified that, for some integers $h_1, \ldots, h_\delta$ such that $h_1 + \cdots + h_\delta = j - 1$, it results that

$$z_r = x_l \cdot x_1^{h_1} \cdots x_l^{h_l} \cdots x_m^{h_m} \cdots x_\delta^{h_\delta}. \tag{B.15}$$

Since $z_s = z_r$, (B.15) implies that

$$z_s = x_m \cdot x_1^{h_1} \cdots x_l^{h_l+1} \cdots x_m^{h_m-1} \cdots x_\delta^{h_\delta}.$$

Hence we have

$$z_{r+q\delta^{j-1}} = x_{l+q} \cdot x_1^{h_1} \cdots x_l^{h_l} \cdots x_m^{h_m} \cdots x_\delta^{h_\delta} = x_{l+q}\frac{z_r}{x_l},$$
$$z_{s+q\delta^{j-1}} = x_{m+q} \cdot x_1^{h_1} \cdots x_l^{h_l+1} \cdots x_m^{h_m-1} \cdots x_\delta^{h_\delta} = x_{m+q}\frac{z_s}{x_m}.$$

From this, since $z_r = z_s$ and $l \neq m$, it follows that $z_{r+q\delta^{j-1}} \neq z_{s+q\delta^{j-1}}$. ∎

Let us consider the state and output processes X, Y, of system (3.2), (3.3). We remind the reader that q and δ are the dimensions of the vectors Y and $Z = [Y^T\ X^T]^T$, respectively. Note that the components of $Z^{[j]}$ can be divided into two groups: the one including monomials composed only with components of the vector Y, and the other one including the remaining monomials. We shall call the components belonging to the former group *the Y-monomials*.

Let us consider the extraction matrix $\mathcal{E}_{\mathcal{Y}}$ defined in (7.19), and recall that the diagonal blocks $\mathcal{E}_{\mathcal{Y}}^j$, $j = 1, \ldots, \nu$, appearing there are such that (7.21) holds. According to the above defined notation (see Remark B.5), we shall denote by q_j the dimension of the vector $Y_{[j]}$. Finally, let us consider the reduction matrix $\tilde{T}_\delta^j$ defined in (7.1) and the matrix U_δ^j defined in (5.5). We can prove the following lemma.

LEMMA B.9. *There exists a full (row) rank matrix, namely L_δ^j, having dimensions $q_j \times q\delta^{j-1}$, such that*

$$\mathcal{E}_{\mathcal{Y}}^j \tilde{T}_\delta^j U_\delta^j = [\, L_\delta^j \quad 0\,].$$

Proof. Using (7.21) and property (5.4) we have

$$\begin{aligned}\mathcal{E}_{\mathcal{Y}}^j\Big(\frac{d}{dZ} \otimes Z_{[j]}\Big) &= \frac{d}{dZ} \otimes \mathcal{E}_{\mathcal{Y}}^j Z_{[j]} = \frac{d}{dZ} \otimes Y_{[j]} \\ &= \Big[\frac{\partial}{\partial Y} \quad \frac{\partial}{\partial X}\Big] \otimes Y_{[j]} = \Big[\frac{\partial}{\partial Y} \otimes Y_{[j]} \quad 0\Big].\end{aligned} \tag{B.16}$$

On the other hand, by (7.1), (5.3), and using formula (5.5),

$$\begin{aligned}\mathcal{E}_{\mathcal{Y}}^j\Big(\frac{d}{dZ} \otimes Z_{[j]}\Big) &= \mathcal{E}_{\mathcal{Y}}^j\Big(\frac{d}{dZ} \otimes \tilde{T}_\delta^j Z^{[j]}\Big) = \mathcal{E}_{\mathcal{Y}}^j \tilde{T}_\delta^j U_\delta^j \big(I_\delta \otimes Z^{[j-1]}\big) \\ &= \mathcal{E}_{\mathcal{Y}}^j \tilde{T}_\delta^j U_\delta^j \begin{bmatrix} I_q \otimes Z^{[j-1]} & 0 \\ 0 & I_{\delta-q} \otimes Z^{[j-1]} \end{bmatrix}.\end{aligned} \tag{B.17}$$

Using (B.16), (B.17), and defining L_δ^j as the matrix composed by the first $q\delta^{j-1}$ columns of $\mathcal{E}_{\mathcal{Y}}^j \tilde{T}_\delta^j U_\delta^j$, it results that

$$[\, \tfrac{\partial}{\partial Y} \otimes Y_{[j]} \quad 0\,] = [\, L_\delta^j \quad S\,] \begin{bmatrix} I_q \otimes Z^{[j-1]} & 0 \\ 0 & I_{\delta-q} \otimes Z^{[j-1]} \end{bmatrix},$$

from which it follows that $S = 0$ and

$$\frac{d}{dY} \otimes Y_{[j]} = L_\delta^j\Big(I_q \otimes Z^{[j-1]}\Big). \tag{B.18}$$

Let $V = (d/dY) \otimes Y_{[j]}$. Note that the components of the matrix V are either zero or they are monomials of $j-1$st degree. It results that V has linearly independent rows (in the sense of linear independence of monomial functions). As a matter of fact, any row is different from zero, and there cannot exist two (nonzero) similar monomials on the same column, because $Y_{[j]}$ has not repeated entries. Hence L^j_δ necessarily has linearly independent rows. Indeed, suppose there exists $u \neq 0$ such that $u^T L^j_\delta = 0$; then we would have $u^T V = 0 \ \ \forall\, Y \in \mathbf{R}^q$, which is a contradiction. □

LEMMA B.10. *Let $s \in \{1, \dots, q\delta^{j-1}\}$ and denote with λ_i, $i = 1, \dots, q\delta^{j-1}$, the ith column of the matrix L^j_δ. The following properties hold:*

(A) *$\forall\ i \in \{1, \dots, q\delta^{j-1}\}$, λ_i has zero entries, but possibly one, nonnegative;*

(B) *the set $\{\lambda_i /\ i \in \rho'(s; \delta, j)\}$, with $\rho'(s; \delta, j)$ given by* (B.2)*, is a set of linearly dependent vectors;*

(C) *if the sth component of $Z^{[j]}$ is not a Y-monomial, then $\lambda_s = 0$.*

Proof. Let us define l and r as

$$l \triangleq \left[\frac{s}{\delta^{j-1}}\right], \quad r \triangleq |s|_{\delta^{j-1}}. \tag{B.19}$$

Consider again the relation (B.18):

$$\frac{d}{dY} \otimes Y_{[j]} = \left[\ \tfrac{\partial}{\partial Y_1} Y_{[j]} \quad \dots \quad \tfrac{\partial}{\partial Y_q} Y_{[j]}\ \right] = L^j_\delta \big(I_q \otimes Z^{[j-1]}\big). \tag{B.20}$$

From (B.20) it results that

$$\frac{\partial}{\partial Y_l} Y_{[j]} = \tilde{L}^{(l)} Z^{[j-1]}, \tag{B.21}$$

where

$$\tilde{L}^{(l)} \triangleq \left[\, \lambda_{(l-1)\delta^{j-1}+1} \quad \lambda_{(l-1)\delta^{j-1}+2} \quad \dots \quad \lambda_{l\delta^{j-1}} \,\right].$$

Now, from (B.21) we see that each component of $(\partial/\partial Y_l) Y_{[j]}$ either is equal to zero or is equal (unless an *integer positive* coefficient) to some component of $Z^{[j-1]}$. Let h be the position of a nonzero entry of $(\partial/\partial Y_l) Y_{[j]}$, and let $r \in \{1, \dots, \delta^{j-1}\}$ be a position for which it appears (unless a coefficient, and possibly repeated) in $Z^{[j-1]}$. Then it results that the hth row of $\tilde{L}$ has, possibly, nonzero (hence positive) elements in the set $\rho(r; \delta, j-1)$. Indeed, this set of positions is determined by the position (r) of the component to be extracted in $Z^{[j-1]}$, endowed with all its $(\delta, j-1)$-R positions.

Let $i \in \{1, \dots, l\delta^{j-1}\}$ such that $\lambda_{(l-1)\delta^{j-1}+i}$ has a nonzero component, namely the hth. Then $\big(\lambda_{(l-1)\delta^{j-1}+i}\big)_k = 0$ for $k = 1, \dots, q_j$ and $k \neq h$. As a matter of fact, if $\big(\lambda_{(l-1)\delta^{j-1}+i}\big)_k \neq 0$, and $k \neq h$, then some monomial, equal to the ith, would be taken in $Z^{[j-1]}$, and hence we would have two equal components in $(\partial/\partial Y_l) Y_{[j]}$, which is impossible because $Y_{[j]}$ has no redundancies. This proves part (A) of the lemma.

From the above it follows that all the columns $\{\lambda_{(l-1)\delta^{j-1}+i},\ i \in \rho(r; \delta, j-1)\}$ have zero entries, but possibly one, placed in the same position h for any $i \in \rho(r; \delta, j-1)$. Hence, they constitute a set of linearly dependent vectors. Part (B) of the lemma follows as soon as it is noticed that, using Lemma B.6 and taking into account (B.19), it results that $\{\lambda_{(l-1)\delta^{j-1}+i},\ i \in \rho(r; \delta, j-1)\} = \{\lambda_i,\ i \in \rho'(s; \delta, j)\}$.

Finally, in order to prove part (C), note that, since $l \leq q$ (and hence, by recalling the structure of Z, given by (7.2), it results that $Z_l = Y_l$), we have that the sth component of $Z^{[j]}$ is of the form $Y_l Z_1^{h_1} \cdots Z_\delta^{h_\delta}$, where the powers $h_1, \dots, h_\delta$ are such

that $h_1 + \cdots + h_\delta = j - 1$, and it is not a Y-monomial by the hypothesis. Hence the monomial $Z_1^{h_1} \cdots Z_\delta^{h_\delta}$ is not a Y-monomial of $Z^{[j-1]}$, and then it cannot belong to the left-hand side of (B.21). This in turn implies, again by (B.21), that the rth column of $\tilde{L}^{(l)}$ (that is the sth column of L_δ^j, because l, r are defined by (B.19)) must be zero. □

Before proving Theorem 8.1, we need to give the following definition.

DEFINITION B.11. *We define the (δ, j)-Kronecker space, namely $\mathbf{K}(\delta, j)$, as the following subspace of $\mathbf{R}^{\delta^j}$:*

$$\mathbf{K}(\delta, j) = \operatorname{span}\left(\left\{ z \in \mathbf{R}^{\delta^j} \;\middle/\; \exists x \in \mathbf{R}^\delta \text{ such that } z = x^{[j]} \right\}\right).$$

Remark B.12. From Definition B.11 it follows that

$$\mathbf{K}(\delta, j) = \left\{ z \in \mathbf{R}^{\delta^j} \;\middle/\; z_r = z_s \text{ if } (r, s) \text{ is } (\delta, j) - R \right\}.$$

Proof of Theorem 8.1. By exploiting the definition of $\mathcal{D}_k^1$ given in (7.28), and the definition of $\tilde{\mathcal{B}}_k$ given by (7.17), we can rewrite the matrix $\mathcal{R}(t)$, defined in (8.1), as

$$\mathcal{R}(t) = \sum_{k=1}^{p} \mathcal{E}_{\mathcal{Y}} \mathcal{B}_k \Phi_Z(t) \mathcal{B}_k^T \mathcal{E}_{\mathcal{Y}}^T + \sum_{k=1}^{p} \mathcal{E}_{\mathcal{Y}} \big(\mathcal{B}_k m_Z(t) + \mathcal{V}_k\big) \big(\mathcal{B}_k m_Z(t) + \mathcal{V}_k\big)^T \mathcal{E}_{\mathcal{Y}}^T. \tag{B.22}$$

We will prove the theorem by showing that the matrix $\mathcal{E}_{\mathcal{Y}} \mathcal{B}_k \Phi_Z \mathcal{B}_k^T \mathcal{E}_{\mathcal{Y}}^T$ is uniformly nonsingular for some $k = 1, \ldots, p$, or (which is the same because $\Phi_Z(t)$ is uniformly nonsingular over T) that $\mathcal{E}_{\mathcal{Y}} \mathcal{B}_k$ is a full (row) rank matrix for some k.

In order to verify this, first of all note that, from Assumption 3.1 and Remark 3.2, it results that there exists a $\bar{k}$ such that $\operatorname{rank}(D_{\bar{k}}) = q$ (we remind the reader that q is the dimension of the original observation Y). Indeed, we have

$$D_{\bar{k}} = [I_q \; 0]. \tag{B.23}$$

For such a $\bar{k}$, let us show that

$$\operatorname{rank}\big(\mathcal{E}_{\mathcal{Y}} \mathcal{B}_{\bar{k}}\big) = q + q_2 + \cdots + q_\nu; \tag{B.24}$$

that is, it is a full (row) rank matrix (remember that q_i is the dimension of $Y_{[i]}$). From the definition of $\mathcal{B}_k$ and $\mathcal{E}_{\mathcal{Y}}$, given in (7.14) and (7.20), respectively, using (7.10) and taking into account the block triangular structure of $\mathcal{B}_k$, it results that condition (B.24) is equivalent to:

$$\operatorname{rank}\big(\mathcal{E}_{\mathcal{Y}}^1 \tilde{B}_{\bar{k}}\big) = q, \tag{B.25}$$

$$\operatorname{rank}\big(\mathcal{E}_{\mathcal{Y}}^j \tilde{T}_\delta^j U_\delta^j (\tilde{B}_{\bar{k}} \otimes I_{\delta^{j-1}}) T_\delta^j\big) = q_j \quad \forall\, j = 2, \ldots, \nu. \tag{B.26}$$

Now, from (7.22) we see that $\mathcal{E}_{\mathcal{Y}}^1 \in \mathbf{R}^{q \times \delta}$, $\mathcal{E}_{\mathcal{Y}}^1 = [I_q \; 0]$. Hence, by the definition of $\tilde{B}_k$ given in (7.5) and taking into account (B.23), it results that $\mathcal{E}_{\mathcal{Y}}^1 \tilde{B}_{\bar{k}} = [0 \; I_q \; 0]$, and hence condition (B.25) is verified.

It remains to prove (B.26). In order to do this, first note that, from the definition of $\tilde{B}_k$ given in (7.5) and taking into account (B.23), we can consider the following partition of the matrix $\tilde{B}_{\bar{k}} \otimes I_{\delta^{j-1}}$:

$$\tilde{B}_{\bar{k}} \otimes I_{\delta^{j-1}} = \begin{bmatrix} M_1 \\ M_2 \end{bmatrix}, \tag{B.27}$$

where M_1 has dimensions $q\delta^{j-1} \times \delta^j$ and has the following structure:

$$M_1 = [\,0 \quad I_{q\delta^{j-1}} \quad 0\,], \tag{B.28}$$

where the first null-block has dimensions $q\delta^{j-1} \times q\delta^{j-1}$. Using Lemma B.9 and (B.27), we have

$$\mathcal{E}_{\mathcal{Y}}^j \tilde{T}_\delta^j U_\delta^j (\tilde{B}_{\bar{k}} \otimes I_{\delta^{j-1}}) T_\delta^j = [\,L_\delta^j \quad 0\,] \begin{bmatrix} M_1 \\ M_2 \end{bmatrix} T_\delta^j = L_\delta^j M_1 T_\delta^j. \tag{B.29}$$

Now note that the range of the *expansion* matrix T_δ^j is equal to the Kronecker space $\mathbf{K}(\delta, j)$ (we remind the reader that T_δ^j performs the operation $Z^{[j]} = T_\delta^j Z_{[j]}$). Then by (B.29), we have that (B.26) is implied by the following condition: the operator $L_\delta^j M_1 : \mathbf{R}^{\delta^j} \to \mathbf{R}^{q_j}$, restricted to $\mathbf{K}(\delta, j)$ is surjective.

Let $y \in \mathbf{R}^{q_j}$, and we will prove that there exists a $z \in \mathbf{K}(\delta, j)$ such that $y = L_\delta^j M_1 z$. By Lemma B.9, $\operatorname{rank}(L_\delta^j) = q_j$; then there exist q_j indexes. $1 \le i_1, i_2, \ldots, i_{q_j} \le q\delta^{j-1}$, such that the columns $\lambda_{i_1}, \lambda_{i_2}, \ldots, \lambda_{i_{q_j}}$ (λ_i denotes as usual the ith column of L_δ^j) are linearly independent. For every i_s, $s = 1, \ldots, q_j$, let us consider the sets $\rho'(i_s; \delta, j) \subset \rho(i_s; \delta, j)$ defined in (B.2). Let us define $\bar{\lambda}_{i_s}$ as

$$\bar{\lambda}_{i_s} \triangleq \sum_{i \in \rho'(i_s;\delta,j)} \lambda_i. \tag{B.30}$$

From Lemma B.10, parts (A) and (B), we have that the set $\{\bar{\lambda}_{i_s},\ s = 1, \ldots, q_j\}$ is a set of linearly independent vectors, and hence there exist real numbers $\alpha_{i_1}, \alpha_{i_2}, \ldots, \alpha_{i_{q_j}}$, such that

$$y = \alpha_{i_1} \bar{\lambda}_{i_1} + \cdots + \alpha_{i_{q_j}} \bar{\lambda}_{i_{q_j}}. \tag{B.31}$$

Now let us show that the elements of the set $\{i_s + q\delta^{j-1}\ s = 1, \ldots, q_j\}$ are pairwise (δ, j)-NR. To this purpose, for any pair (i_r, i_s), $r \neq s$, $r, s = 1, \ldots, q_j$, we can distinguish the following two cases.

(i)

$$\left[\frac{i_r}{\delta^{j-1}}\right] = \left[\frac{i_s}{\delta^{j-1}}\right].$$

In this case, since λ_{i_r} and λ_{i_s} are linearly independent, it follows that (i_r, i_s) is (δ, j)-NR. Indeed, if (i_r, i_s) were (δ, j)-R, then Lemma B.10, part (B) would imply that λ_{i_r} and λ_{i_s} are linearly dependent vectors. Hence, since (i_r, i_s) is (δ, j)-NR, Lemma B.7 implies that $(i_r + q\delta^{j-1}, i_s + q\delta^{j-1})$ is (δ, j)-NR.

(ii)

$$i'_r \triangleq \left[\frac{i_r}{\delta^{j-1}}\right] \neq \left[\frac{i_s}{\delta^{j-1}}\right] \triangleq i'_s. \tag{B.32}$$

In this case, if (i_r, i_s) is (δ, j)-R then Lemma B.8 directly implies the same conclusion of (i). Else, if (i_r, i_s) is (δ, j)-NR, then we can show that $(i_r + q\delta^{j-1}, i_s + q\delta^{j-1})$ is again (δ, j)-NR. For, let $h_1, \ldots, h_\delta$, $h'_1, \ldots, h'_\delta$ such that $h_1 + \cdots + h_\delta = h'_1 + \cdots + h'_\delta = j - 1$ and

$$\begin{aligned} \left(Z^{[j]}\right)_{i_r} &= Z_{i'_r} Z_1^{h_1} \cdots Z_\delta^{h_\delta}, \\ \left(Z^{[j]}\right)_{i_s} &= Z_{i'_s} Z_1^{h'_1} \cdots Z_\delta^{h'_\delta}, \end{aligned} \tag{B.33}$$

where i'_r, i'_s are given by (B.32). Since λ_{i_r} and λ_{i_s} are linearly independent (hence nonzero), Lemma B.10, part (C) implies that both the monomials in (B.33) are Y-monomials. If $(i_r + q\delta^{j-1}, i_s + q\delta^{j-1})$ were (δ, j)-R, we should have

$$Z_{i'_r+q} Z_1^{h_1} \cdots Z_\delta^{h_\delta} = Z_{i'_s+q} Z_1^{h'_1} \cdots Z_\delta^{h'_\delta},$$

which is possible if and only if

$$\begin{aligned} &h_{i'_r+q} = h'_{i'_r+q} - 1, \quad h_{i'_s+q} = h'_{i'_s+q} + 1, \\ &h_i = h'_i \quad \forall\ i \neq i'_r + q, i'_s + q. \end{aligned} \tag{B.34}$$

Now $i'_r, i'_s \leq q$, and then we have that $Z_{i'_r+q}$ and $Z_{i'_s+q}$ are not components of the vector Y, hence condition (B.34) can be verified if and only if both the monomials in (B.33) are not Y-monomials, which is a contradiction.

Since the elements of the set $\{i_s + q\delta^{j-1},\ s = 1, \dots q_j\}$ are pairwise (δ, j)-NR, we have that

$$\rho(i_s + q\delta^{j-1}; \delta, j) \cap \rho(i_r + q\delta^{j-1}; \delta, j) = \emptyset \quad \forall\ r, s = 1, \dots, q_j,\ r \neq s. \tag{B.35}$$

From (B.35) it results that the following vector $z \in \mathbf{R}^{\delta^{j-1}}$ is well defined:

$$z_l = \begin{cases} \alpha_{i_s} & \text{if } l \in \rho(i_s + q\delta^{j-1}; \delta, j), \\ 0 & \text{otherwise.} \end{cases} \tag{B.36}$$

Noting that, by construction, $z \in \mathbf{K}(\delta, j)$, the theorem is proven as soon as it is shown that $y = L_\delta^j M_1 z$ with y given by (B.31).

Let $z' = M_1 z$. By the structure of the matrix M_1 (B.28), it follows that

$$z'_l = \begin{cases} \alpha_{i_s} & \text{if } l \in \rho(i_s + q\delta^{j-1}; \delta, j) - q\delta^{j-1}, \\ 0 & \text{otherwise,} \end{cases} \tag{B.37}$$

where the definition of translated set, given by (B.8), has been used. Observing (B.37), (B.31), and the definition of the $\bar{\lambda}_{i_s}$s (B.30), we see that the equality $y = L_\delta^j z'$, and hence the theorem is implied by the condition

$$\sum_{i \in \rho'(i_s; \delta, j)} \lambda_i = \sum_{i \in \rho(i_s + q\delta^{j-1}; \delta, j) - q\delta^{j-1}} \lambda_i \quad \forall\ s = 1, \dots, q_j. \tag{B.38}$$

Now, from Lemma B.7 we have $\rho'(i_s; \delta, j) = \rho'(i_s + q\delta^{j-1}; \delta, j) - q\delta^{j-1}$; moreover, by (B.3)

$$\rho(i_s + q\delta^{j-1}; \delta, j) = \rho'(i_s + q\delta^{j-1}; \delta, j) \cup \rho''(i_s + q\delta^{j-1}; \delta, j),$$

and hence (B.38) becomes

$$\sum_{i \in \rho''(i_s + q\delta^{j-1}; \delta, j) - q\delta^{j-1}} \lambda_i = 0 \quad \forall\ s = 1, \dots, q_j,$$

which is implied by

$$\lambda_i = 0 \quad \forall\ i \in \rho''(i_s + q\delta^{j-1}; \delta, j) - q\delta^{j-1}. \tag{B.39}$$

In order to prove (B.39), first note that by Lemma B.8, for any $i \in \rho''(i_s+q\delta^{j-1};\delta,j)-q\delta^{j-1}$, we must have that (i,i_s) is (δ,j)-NR and such that $(i+q\delta^{j-1}, i_s+q\delta^{j-1})$ is (δ,j)-R. Now, let $h_1,\dots,h_\delta$, $h'_1,\dots,h'_\delta$ such that $h_1+\cdots+h_\delta = h'_1+\cdots+h'_\delta = j-1$ and

$$\begin{aligned} \left(Z^{[j]}\right)_i &= Z_{i'} Z_1^{h_1}\cdots Z_\delta^{h_\delta}, \\ \left(Z^{[j]}\right)_{i_s} &= Z_{i'_s} Z_1^{h'_1}\cdots Z_\delta^{h'_\delta}, \end{aligned} \tag{B.40}$$

with $i' = [i/\delta^{j-1}]$, $i'_s = [i_s/\delta^{j-1}]$. Since $(i+q\delta^{j-1}, i_s+q\delta^{j-1})$ is (δ,j)-R, it results that

$$Z_{i'+q} Z_1^{h_1}\cdots Z_\delta^{h_\delta} = Z_{i'_s+q} Z_1^{h'_1}\cdots Z_\delta^{h'_\delta},$$

which in turn implies the following condition:

$$\begin{aligned} &h_{i'+q} = h'_{i'+q} - 1, \quad h_{i'_s+q} = h'_{i'_s+q} + 1, \\ &h_i = h'_i \quad \forall\, i \neq i'+q, i'_s+q. \end{aligned} \tag{B.41}$$

Since $i', i'_s \leq q$, $Z_{i'+q}$ and $Z_{i'_s+q}$ are not components of the vector Y. Hence, condition (B.41) implies that both the monomials in (B.40) are not Y-monomials. In particular, since $\left(Z^{[j]}\right)_i$ is not a Y-monomial, Lemma B.10, part (C) gives $\lambda_i = 0$, that is (B.39). □

REFERENCES

[1] R. E. Kalman, *A new approach to linear filtering and prediction problems*, J. Basic. Engrg., 1 (1960), pp. 35–45.

[2] R. E. Kalman and R. S. Bucy, *New results in linear filtering and prediction theory*, Trans. ASME Ser. D. J. Basic Engrg., 83 (1961), pp. 95–108.

[3] R. S. Liptser and A. N. Shiryayev, *Statistics of Random Processes*, Vol. 2, Springer-Verlag, New York, 1978.

[4] R. R. Mohler, *Bilinear Control Processes*, Academic Press, New York, 1970.

[5] C. Bruni, G. Di Pillo, and G. Koch, *Bilinear systems: An appealing class of "nearly linear" systems in theory and applications*, IEEE Trans. Automat. Control, 19 (1974), pp. 334–348.

[6] C. Bruni and G. Koch, *Suboptimal filtering for linear-in-control systems with small observation noise*, in IV International Federation of Automatic Control Symposium on Identification and System Parameter Estimation, Tblisi, SSSR, Sept. 1976.

[7] X. Yang, R. R. Mohler, and R. M. Burton, *Adaptive suboptimal filtering of bilinear systems*, Internat. J. Control, 52 (1990), pp. 135–158.

[8] E. Yaz, *Full and reduced order observer design for discrete stochastic bilinear systems*, IEEE Trans. Automat. Control, 37 (1992), pp. 503–505.

[9] F. Carravetta, A. Germani, and M. Raimondi, *Polynomial filtering for linear discrete-time non-Gaussian systems*, SIAM J. Control Optim., 34 (1996), pp. 1666–1690.

[10] F. Carravetta, A. Germani, and M. Raimondi, *Polynomial filtering of discrete-time stochastic linear systems with multiplicative state noise*, IEEE Trans. Automat. Control, 42 (1997), pp. 1106–1126.

[11] G. Kallianpur, *Stochastic Filtering Theory*, Springer-Verlag, New York, 1980.

[12] V. Krishnan, *Nonlinear Filtering and Smoothing*, John Wiley, New York, 1984.

[13] A. Germani and M. Piccioni, *Semi-discretization of stochastic partial differential equations on $\mathbf{R}^d$ by a finite-element technique*, Stochastics, 23 (1988), pp. 131–148.

[14] R. Bellman, *Introduction to Matrix Analysis*, McGraw-Hill, New York, 1970.

[15] G. S. Rodgers, *Matrix Derivatives*, Lecture Notes in Statist. 2, Marcel Dekker, New York, 1980.

SIAM J. CONTROL OPTIM.
Vol. 38, No. 4, pp. 1204–1222

CONTINUITY OF OPTIMAL VALUES AND SOLUTIONS FOR CONTROL OF MARKOV CHAINS WITH CONSTRAINTS*

MABEL M. TIDBALL†, ARIEL LOMBARDI‡, ODILE POURTALLIER§, AND EITAN ALTMAN§

Abstract. We consider in this paper constrained Markov decision processes. This type of control model has many applications in telecommunications and other fields [E. Altman and A. Shwartz, *IEEE Trans. Automat. Control*, 34 (1989), pp. 1089–1102, E. A. Feinberg and M. I. Reiman, *Probab. Engrg. Inform. Sci.*, 8 (1994), pp. 463–489, A. Hordijk and F. Spieksma, *Adv. in Appl. Probab.*, 21 (1989), pp. 409–431, A. Lazar, *IEEE Trans. Automat. Control*, 28 (1983), pp. 1001–1007, P. Nain and K. W. Ross, *IEEE Trans. Automat. Control*, 31 (1986), pp. 883–888, K. W. Ross and B. Chen, *IEEE Trans. Automat. Control*, 33 (1988), pp. 261–267]. We address the issue of the convergence of the value and optimal policies of the problem with discounted costs, to the ones for the problem with expected average cost. We consider the general multichain ergodic structure. We present two stability results in this paper. We establish the continuity of optimal values and solutions of as well as some type of robustness of some suboptimal solutions in the discount factor. Our proof relies on same general theory on continuity of values and solutions in convex optimization that relies on well-known notions of Γ-convergence.

Key words. optimization, sensitivity analysis, constrained Markov decision processes

AMS subject classifications. 93E20, 93B35, 90C40

PII. S0363012997280294

1. Introduction.

We consider a sequence $\mathbf{MP}_n$, $n = 1, 2, \ldots$ of constrained Markov decision processes (CMDPs), and a "limit" one, denoted by $\mathbf{MP}_\infty$, or simply by $\mathbf{MP}$. These are defined on some vector spaces, possibly infinite dimensional ones. $\mathbf{MP}$ is assumed to be feasible (it has at least one solution). However, for any given n, $\mathbf{MP}_n$ need not be feasible, and even if it is, it need not possess an optimal solution (i.e., it may only have ϵ-optimal solutions). We are interested in the following questions.

(i) Do the values of $\mathbf{MP}_n$ converge to the value of $\mathbf{MP}$?
(ii) Do optimal (or almost optimal) policies converge in some sense?
(iii) Given an (almost) optimal policy for $\mathbf{MP}_n$, will it be an almost optimal policy for $\mathbf{MP}$ if n is sufficiently large?
(iv) Conversely, given an optimal policy for $\mathbf{MP}$, will it be an almost optimal policy for $\mathbf{MP}_n$ for all n sufficiently large?

By reducing our control problem to equivalent mathematical programs, we show that the epigraph theory and the Γ-convergence theory provide sufficient conditions for having convergence in the sense of (i) and (ii) above. It turns out that the answers for (iii) and for (iv) are in general negative, unlike the unconstrained case. The reason is that an optimal policy for $\mathbf{MP}_n$ may be unfeasible for $\mathbf{MP}$, and vice versa. We shall, however, establish sufficient conditions for the following slightly weaker versions of (iii) and (iv).

*Received by the editors December 2, 1997; accepted for publication (in revised form) August 23, 1999; published electronically April 18, 2000.
http://www.siam.org/journals/sicon/38-4/28029.html

†Institut National de Recherche en Agronomie, Economie et Sociologie Rurales, 2 Place Viala 34060, Montpellier Cedex 1, France (tidball@ensam.inra.fr).
‡Facultad de Ciencias Exactas y Naturales, Universidad de Buenos Aires. Ciudad Universtaria, 1428 Capital Federal, Argentina (aldoc7@mate.dm.uba.ar).
§INRIA, Centre Sophia-Antipolis, 2004 Route des Lucioles, B.P. 93, 06902 Sophia-Antipolis Cedex, France (pourtali@sophia.inria.fr, altman@sophia.inria.fr).

(iii′) Given an optimal policy for $\mathbf{MP}_n$, can we perturb it "slightly" so that it becomes almost optimal for $\mathbf{MP}$ if n is sufficiently large?

(iv′) Given an optimal policy for $\mathbf{MP}$, can we perturb it "slightly" so that it becomes almost optimal for $\mathbf{MP}_n$ for all n sufficiently large?

The answers for (iii′) and (iv′) follow from continuity properties of the solutions and values of mathematical programming. We use in particular notions of Γ-convergence (see [16] and [37]). We extend and adapt existing theorems and tools for the sensitivity of solutions of convex optimization to obtain the appropriate framework for analyzing CMDPs.

We focus here on the convergence in the discount factor α_n as it converges to some limit within the unit interval. We present sufficient conditions for the convergence of the values and of optimal policies, as well as some robustness properties of suboptimal policies. Related results were already obtained in [1] but had a strong restriction on the ergodic structure of the controlled model. It was required to have a single ergodic class under any stationary policy. This condition enabled us to restrict our problem to stationary policies, for which some general theorem on approximation [2] could be used. In the present paper we make no assumption on the ergodic structure, thus allowing a multichain situation. For such ergodic structure, it is known that stationary policies need not be optimal (nor even ϵ-optimal) for the expected average cost criterion, and one has to use either Markov policies (see [25, 27]) or mixed-stationary policies (this term was raised by Feinberg [21] in a similar context; it refers to policies that are highly nonstationary). We use the latter approach to establish, with the help of the results from the first part of the paper, the convergence of the values and policies.

We briefly mention some related work on the continuity and sensitivity analysis of mathematical programs and of control problems. Many papers and books studied similar problems in the case of finite dimensional state, e.g., [17, 24, 36]. Several special issues of scientific journals were devoted to these questions, as well as other related sensitivity, stability, and parametric analysis: *Mathematical Programming* 21, 1984, *Annals of Operations Research* 27, 1990. Convergence results for constrained dynamic control problems were obtained in [1, 2, 4, 6, 8]. Conditions were obtained there for the convergence in the transition probabilities, in the horizon, and in the immediate cost. These results were applied to adaptive control problems [6] and to problems of finite state approximations of CMDPs. Similar questions to those addressed in this paper were studied in [23] and in [43], and some of the results there are close to those in the first part of our paper. The main difference lies in the types of assumptions made. In [23] the central assumption is stated in terms of constraints set convergence making the use of a metric. In [43] less regularity on the constraints sets are required (convergence of the constraint sets in the Hausdorff topology). Nevertheless only points (i) and (ii) are studied there. Some related questions but in the context of min-max problems and Stakelberg equilibrium are studied in [31, 32, 34]. Some other references related to the current paper are [13, 16, 28, 33, 42, 44].

2. CMDPs: The convergence in the discount factor. We consider CMDPs, known also as controlled Markov chains, with a general ergodic structure (multichain). We consider the discounted cost and the average cost. We shall obtain new results on the convergence of the values and optimal policies of the discounted cost, as the discount factor tends to one, to the value and to optimal policies corresponding to the expected average cost. A similar result for the special unichain case was obtained

in [1].

Consider a Markov decision process (MDP) with a finite state space $\mathbf{X} = \{0, 1, \ldots, N\}$ and a finite *action space* $\mathbf{A}$. Without loss of generality, we assume that in any state x all actions in $\mathbf{A}$ are available. The probability to go from state x to state y given that action a is used is given by the transition probability P_{xay}. A policy u in the *policy space* $\mathbf{U}_h$ is described as $u = (u_1, u_2, \ldots)$, where u_t, applied at time epoch t, is a probability measure over $\mathbf{A}$ conditioned on the whole history of actions and state prior to t, as well as the state at time t. Given an initial distribution β on $\mathbf{X}$, each policy u induces a probability measure denoted by P^u_β on the space of sample paths of states and actions (which serves as the canonical sample space Ω). The corresponding expectation operator is denoted by E^u_β. On this probability space are defined the state and action processes, X_t, A_t, $t = 1, 2, \ldots$.

A *Markov policy* $u \in \mathbf{U}_M$ is characterized by the dependence of u_{t+1} on the current state and the time only. A *stationary policy* $g \in \mathbf{U}_S$ is characterized by a single conditional probability measure $p^g_{\cdot|x}$ over $\mathbf{A}$, so that $p^g_{\mathbf{A}|x} = 1$; under g, X_t becomes a Markov chain with stationary transition probabilities, given by $P^g_{xy} = \sum_{a\in\mathbf{A}} p^g_{a|x} P_{xay}$. The class of *stationary deterministic policies* $\mathbf{U}_D$ is a subclass of $\mathbf{U}_S$, and every $g \in \mathbf{U}_D$ is characterized by a mapping g:$\mathbf{X} \to \mathbf{A}$, so that $p^g_{\cdot|x} = \delta_{g(x)}(\cdot)$ is concentrated at the point $g(x)$ for each x. Let L be the number of stationary deterministic policies among $\mathbf{U}_D$, and enumerate the policies in $\mathbf{U}_D$ such that $\mathbf{U}_D = \{u^1, \ldots, u^L\}$.

It will often be useful to extend the definition of a policy $u = (u_1, u_2, \ldots)$ so as to allow u_t to depend not only on the history, but also on some additional randomizing mechanism. In particular, for any finite class of policies $G \subset \mathbf{U}_h$, we define $\overline{M}(G)$ to be the class of mixed policies generated by G. We call these mixed-G policies. A mixed-G policy $\hat{q}$ is identified with a distribution q over G; the controller first uses q to choose some policy $u \in G$ and then proceeds with that policy from time 1 onwards. Define $\mathcal{U} := \overline{M}(U_D)$. Finally, we denote $\mathbf{U} = \mathbf{U}_h \cup \mathcal{U}$.

DEFINITION 2.1. *For any initial distribution q over the set $\mathbf{U}_D$, we shall identify the policy $m(q) \in \mathcal{U}$ to be the one that chooses initially the policy u^j with probability q_j.*

Remark 2.2. Although we consider here a more general definition of policies than $\mathbf{U}_h$, these policies are in fact equivalent to those in $\mathbf{U}_h$ (details are given in Remark 3.1). A randomization over actual classes of policies has already been considered in [20, 29]. An alternative for the randomization over policies is the randomization over the strategic measures they generate. If we identify a policy with the strategic measure it generates, then it follows that the large class $\overline{M}(\mathbf{U}_h)$ is equivalent to the class of Markov policies $\mathbf{U}_M$, as was shown in [19].

For any given distribution β for the initial state (at time 1) and a policy u, define a probability measure P^u_β on which the stochastic processes X_t and A_t of the states and actions are defined. When β is concentrated on some state x (i.e., $\beta = \delta_x$), we shall use the notation P^u_x instead of P^u_β.

Let $c : \mathbf{X} \times \mathbf{A} \to \mathbb{R}$ and $d : \mathbf{X} \times \mathbf{A} \to \mathbb{R}^K$ be immediate cost functions, $d = (d^1, d^2, \ldots, d^k)$.

Fix some discount factor $\alpha \in [0, 1)$, and define the normalized discounted costs corresponding to an initial distribution β and a policy u by

$$C_\alpha(\beta, u) = \sum_{t=1}^{\infty} (1-\alpha) E^u_\beta \alpha^{t-1} c(X_t, A_t),$$

$$D_\alpha^k(\beta,u)=\sum_{t=1}^{\infty}(1-\alpha)E_\beta^u\alpha^{t-1}d^k(X_t,A_t),\qquad k=1,\dots,K.$$

Define the average costs associated with a policy u and with an initial distribution β on $\mathbf{X}$:

$$C_{ea}(\beta,u)=\overline{\lim_{t\to\infty}}\,\frac{1}{t}E_\beta^u\left[\sum_{s=1}^{t}c(X_s,A_s)\right],$$

$$D_{ea}^k(\beta,u)=\overline{\lim_{t\to\infty}}\,\frac{1}{t}E_\beta^u\left[\sum_{s=1}^{t}d^k(X_s,A_s)\right],\qquad k=1,\dots,K.$$

Given a vector $V\in\mathbb{R}^K$, we consider the subset $\Pi_V\subset\mathbf{U}$ of policies satisfying the constraints

$$D(\beta,u)\le V. \tag{2.1}$$

A policy $u\in\Pi_V$ is called feasible. We introduce the following constrained problem (COP): find a policy u^* that achieves

$$C(\beta):=\inf_{u\in\Pi_V}C(\beta,u). \tag{2.2}$$

In (2.1) and (2.2), the costs stand for either the discounted or the average cost. The COP is said to be feasible if Π_V is nonempty.

We are now ready to state the first main result.

THEOREM 2.3 (convergence of the value). *Assume that the following Slater condition holds: there exists some policy $v\in\mathbf{U}$ such that*

$$D^k(\beta,v)\le V^k-\eta\qquad\forall k=1,\dots,K \tag{2.3}$$

for some $\eta>0$. Then, the value converges in the discount factor:

$$\lim_{\alpha\to1}C_\alpha(\beta)=C_{ea}(\beta).$$

The proof of this theorem, as well as the convergence of optimal policies, are delayed to the next section.

3. Convergence of the values and the policies. In order to be able to define the convergence of optimal policies, we shall show that one may restrict the search of optimal policies to the simple subclasses of policies $\mathcal{U}$, without loss of optimality. Moreover, we should relate the solutions of the COP to solutions of mathematical programming, in order to be able to apply the tools that we developed. Note that the control problem is already of the form of a mathematical program, but the cost is not convex in the policies. We shall show that when restricting to $\mathcal{U}$, the costs are convex functions.

There are several ways to solve (2.2). For the discounted cost, the solution was given by Kallenberg in [27] using a linear program (LP) approach. For the expected average cost, there are several possible LP approaches: the one by Hordijk and Kallenberg [25, 27], the one by Feinberg [21], and a related one by Altman and Shwartz [9]; for a definition slightly different than in (2.2), an efficient LP method for computing ϵ-optimal solutions was obtained by Ross and Varadarajan; see [40, 41]. Lagrangian

techniques have also been used to solve CMDPs with a single constraint; see Beutler and Ross [11, 12] and Sennott [45, 46]. The relation between the Lagrange and the LP approaches was pointed out in [10].

Remark 3.1. All the references above considered the solution of the constrained MDP among the policies $\mathbf{U}_h$. However, a standard argument due to Derman and Strauch [18] shows (in a constructive way) that for any policy $u \in \mathbf{U}$, there exists an equivalent policy $\chi \in \mathbf{U}_M$ under which the marginal probabilities of the states and actions are the same as those under u, and in particular, both the discounted and the expected average costs are the same. Thus below, whenever we obtain an optimal policy among $\mathcal{U}$, one may consider the policy χ instead without loss of optimality. Note, however, that one cannot in general restrict further to the stationary policies $\mathbf{U}_S$. Indeed, it is shown in [25, 27] that they are not sufficient for the expected average cost with a general multichain structure.

We would like to prove the convergence results by showing that there is a correspondence between values and optimal solutions of the control problem, and values and optimal solutions of related LPs, and then use the general results from the previous section. While this is possible for the unichain case, it turns out that for the multichain case, the LP introduced by Kallenberg [27] for the discounted cost is completely different than any of the LPs for the expected average cost (e.g., the number of decision variables is different). Therefore, as a first step, we shall introduce a new LP method for computing the value and optimal policies for the discounted cost problem, which is an adaptation of the one for the expected average cost in Feinberg [21] and Altman and Shwartz [9]. This will allow us to have the same type of LP for both the discounted and the average cost.

Denote the simplex $S(L) := \{\gamma \in R^L : \gamma_i \geq 0, i = 1, \ldots, L, \ \sum_{i=1}^L \gamma_i = 1\}$. Introduce the following $\mathbf{LP}_\alpha$: Find $\gamma^* \in S(L)$ that achieves

$$\mathcal{C}_\alpha^* := \min_{\gamma \in S(L)} \sum_{i=1}^L \gamma_i C_\alpha(\beta, u^i), \qquad \textit{such that (s.t.)} \tag{3.1}$$

$$\mathcal{D}_\alpha^k(\gamma) := \sum_{i=1}^L \gamma_i D_\alpha^k(\beta, u^i) \leq V^k, \ k = 1, \ldots, K. \tag{3.2}$$

Define $\mathcal{C}_\alpha(\gamma) := \sum_{i=1}^L \gamma_i C_\alpha(\beta, u^i)$. We say that the LP is feasible if the subset of $S(L)$ satisfying the constraints (3.2) is nonempty.

THEOREM 3.2 (relation between LP and the CMDP, the discounted cost).

(i) *For any $\gamma \in S(L)$, the policy $m(\gamma) \in \mathcal{U}$ (see Definition 2.1) satisfies*

$$C_\alpha(\beta, m(\gamma)) = \mathcal{C}_\alpha(\gamma), \qquad D_\alpha^k(\beta, m(\gamma)) = \mathcal{D}_\alpha^k(\gamma), \qquad k = 1, \ldots, K.$$

(ii) *For any vector of costs*

$$\{C_\alpha(\beta, u), D_\alpha^k(\beta, u), k = 1, \ldots, K\},$$

achievable by some policy $u \in \mathbf{U}$, there exists some $v \in \mathcal{U}$ achieving the same vector of costs.

(iii) *COP_α is feasible if and only if LP_α is and the optimal values are the same: $\mathcal{C}_\alpha^* = C_\alpha(\beta)$. Moreover, if γ^* is optimal for LP_α, then $m(\gamma^*)$ is optimal for COP_α.*

Proof. Denote

$$f_\alpha(\beta,u;y,a) := (1-\alpha)\sum_{t=1}^{\infty}\alpha^{t-1}P_\beta^u(X_t=y,A_t=a), \qquad x\in\mathbf{X}, a\in\mathbf{A},$$

and let $f_\alpha(\beta,u)$ be the vector whose (y,a)th elements are given by $f_\alpha(\beta,x;y,a)$. Then (see, e.g., [14])

$$C_\alpha(\beta,u) = c\cdot f_\alpha(\beta,u) = \sum_{y\in\mathbf{X}}\sum_{a\in\mathbf{A}} c(y,a)f_\alpha(\beta,u;y,a), \tag{3.3}$$

with a similar representation for the costs $D_\alpha^k(\beta,u)$. For any class of policies $\overline{U}$, denote $\mathbf{L}_\alpha(\beta,\overline{U}) = \cup_{u\in\overline{U}} f_\alpha(\beta,u)$. It is known that the set $\mathbf{L}_\alpha(\beta,\mathbf{U}_h)$ is convex, compact, and its extreme points are contained in $\{f_\alpha(\beta,u), u\in\mathbf{U}_D\}$; see [14, 27]. For any probability γ over $\mathbf{U}_D$, we clearly have

$$f_\alpha(\beta,m(\gamma)) = \sum_{i=1}^{L}\gamma_i f_\alpha(\beta,u^i). \tag{3.4}$$

Consequently, $\mathbf{L}_\alpha(\beta,\mathbf{U})$ is convex, compact, and its extreme points are $\{f_\alpha(\beta,u), u\in\mathbf{U}_D\}$.

Combining this with (3.3), we conclude that the set of achievable costs

$$\cup_{u\in\mathbf{U}}\{C_\alpha(\beta,u), D_\alpha^k(\beta,u), k=1,\dots,K\} \tag{3.5}$$

is also convex, compact, and its extreme points are

$$\{C_\alpha(\beta,u), D_\alpha^k(\beta,u), k=1,\dots,K, u\in\mathbf{U}_D\}. \tag{3.6}$$

By combining (3.3) with (3.4), we get, for any probability γ over $\mathbf{U}_D$,

$$C_\alpha(\beta,m(\gamma)) = \sum_{i=1}^{L}\gamma_i C_\alpha(\beta,u^i), \qquad D_\alpha^k(\beta,m(\gamma)) = \sum_{i=1}^{L}\gamma_i D_\alpha(\beta,u^i),\ k=1,\dots,K.$$

Hence, the set of performance measures achievable by $u\in\mathcal{U}$ is also convex, compact, with the extreme points in the set (3.6), and thus, equal to the set (3.5) achievable by all policies. This establishes (i) and (ii) and implies (iii). □

The LP method corresponding to (3.1) for the expected average cost, due to Feinberg [21], is the same: $\mathbf{LP_{ea}}$: Find $\gamma^*\in S(K)$ that achieves

$$\mathcal{C}_{ea}^* := \min_{\gamma\in S(L)}\sum_{i=1}^{L}\gamma_i C_{ea}(\beta,u^i), \qquad s.t. \tag{3.7}$$

$$\mathcal{D}_{ea}^k(\gamma) := \sum_{i=1}^{L}\gamma_i D_{ea}^k(\beta,u^i) \le V^k,\ k=1,\dots,K. \tag{3.8}$$

Define $\mathcal{C}_{ea}(\gamma) := \sum_{i=1}^{L}\gamma_i C_{ea}(\beta,u^i)$.

THEOREM 3.3 (relation between LP and the CMDP, the expected average cost).

(i) *For any $\gamma\in S(L)$, the policy $m(\gamma)\in\mathcal{U}$ (see Definition 2.1) satisfies*

$$C_{ea}(\beta,m(\gamma)) = \mathcal{C}_{ea}(\gamma), \qquad D_{ea}^k(\beta,m(\gamma)) = \mathcal{D}_{ea}^k(\gamma), \qquad k=1,\dots,K.$$

(ii) *For any vector of costs*

$$\{C_{ea}(\beta,u), D^k_{ea}(\beta,u), k=1,\dots,K\}$$

achievable by some policy $u \in \mathbf{U}$*, there exists a dominating* $v \in \mathcal{U}$*, i.e., such that*

$$C_{ea}(\beta,v) \le C_{ea}(\beta,u), \quad D^k_{ea}(\beta,v) \le D^k_{ea}(\beta,u), k=1,\dots,K.$$

(iii) COP_{ea} *is feasible if and only if* LP_{ea} *is and the optimal values are the same:* $\mathcal{C}^*_{ea} = C_{ea}(\beta)$*. Moreover, if* γ^* *is optimal for* LP_{ea}*, then* $m(\gamma^*)$ *is optimal for* COP_{ea}*.*

Proof. Denote

$$f^t_{ea}(\beta,u;y,a) := t^{-1}\sum_{s=1}^{t} P^u_\beta(X_s=y, A_s=a), \qquad x\in\mathbf{X}, a\in\mathbf{A},$$

and let $f^t_{ea}(\beta,u)$ be the vector whose (y,a)th elements are given by $f^t_{ea}(\beta,x;y,a)$. For any $u \in U_D$, since the state process is a Markov chain, it is known that

$$f_{ea}(\beta,u) = \lim_{t\to\infty} f^t_{ea}(\beta,u) \text{ exists}, \tag{3.9}$$

and it is straightforward to show that

$$C_{ea}(\beta,u) = c\cdot f_{ea}(\beta,u) = \sum_{y\in\mathbf{X}}\sum_{a\in\mathbf{A}} c(y,a) f_{ea}(\beta,u;y,a), \tag{3.10}$$

with a similar representation for the costs $D^k_{ea}(\beta,u)$. It is then clear that (3.9) and (3.10) hold in fact for any $u \in \mathcal{U}$, which establishes (i).

For a fixed initial distribution β, for any policy $v \in \mathbf{U}_h$, and for any accumulation point f of the sequence $f^t_{ea}(\beta,u)$, there exists some $u \in \mathcal{U}$ such that $f_{ea}(\beta,u) = f$. This is a direct consequence of Theorem 2 in [25] and is a special case of the result in [21] (who studies the semi-Markov case). Combining this with the fact that (3.9) and (3.10) hold for any $u \in \mathcal{U}$ establishes (ii), by using Corollary 2.5 in [7]. Finally, (iii) is a consequence of (i) and (ii). □

For a given $u \in \mathcal{U}$ we shall understand below $\pi_\delta(u) = m(\pi_\delta(\gamma))$, where γ is such that $u = m(\gamma)$ and π_δ is defined in (B.17). We are now ready to state the second main result for MDPs.

THEOREM 3.4. *Assume that the Slater condition* (2.3) *holds. Consider a sequence* α_n *converging to* 1*, and let* COP_n *be the constrained optimal control problem corresponding to the discount factor* α_n*. Let* δ *be such that* $\eta > \delta > 0$*. Then we have the following.*

(i) *Let* $u_n \in \mathcal{U}$ *be* ϵ_n*-optimal for* COP_n*,* $\limsup_n \epsilon_n \le \epsilon$*. Then there exists* $N(\epsilon,\delta)$ *such that* $\forall n \ge N(\epsilon,\delta)$*,* $\pi_\delta(u_n)$ *is* $O(\epsilon+\delta)$*-optimal for* COP_{ea}*.*

(ii) *Let* $u \in \mathcal{U}$ *be optimal for* COP_{ea}*. Then there exists* $N(\epsilon,\delta)$ *such that* $\forall n \ge N(\epsilon,\delta)$*,* $\pi_\delta(u)$ *is* $O(\epsilon+\delta)$*-optimal for* COP_n*.*

(iii) *Let* $u_n \in \mathcal{U}$ *be optimal for* COP_n *and let* $\gamma_n \in S(L)$ *be such that* $u_n = m(\gamma_n)$*. Assume that* γ_n *converges to some* γ*. Then* $m(\gamma)$ *is optimal for* COP_{ea}*.*

Proof of Theorems 2.3 *and* 3.4. We apply below Theorems B.1, B.4, and B.15, which present sensitivity results for general convex optimization problems, to obtain the convergence of the optimal values and the convergence and robustness of policies

LP_α to the optimal value of LP_{ea}, and consequently, by Theorems 3.2(ii), (iii), and 3.3(ii), (iii), the convergence for the original constrained optimal control problems. It remains to show that the required conditions in Theorems B.1, B.4, and B.15 hold.

It is well known that for any $u \in U_D$,

$$\lim_{\alpha\to 1} C_\alpha(\beta,u) = C_{ea}(\beta,u), \qquad \lim_{\alpha\to 1} D^k_\alpha(\beta,u) = D^k_{ea}(\beta,u),\ k=1,\dots,K. \tag{3.11}$$

Choose an arbitrary $u \in \mathcal{U}$, and let γ be such that $u = m(\gamma)$. Then, due to Theorems 3.2(i) and 3.3(i),

$$\begin{aligned} &|C_\alpha(\beta,u) - C_{ea}(\beta,u)| \\ &\le \left| \sum_{j=1}^{L} \gamma_j \left[C_\alpha(\beta,u^j) - C_{ea}(\beta,u^j) \right] \right| \\ &\le \max_{1\le j\le L} |C_\alpha(\beta,u^j) - C_{ea}(\beta,u^j)|, \end{aligned}$$

which does not depend on u.

With the same applied to the costs D^k, this implies that (3.11) holds for any $u \in \mathcal{U}$, and that the convergence is uniform over $\mathcal{U}$. Equivalently, for any $\gamma \in S(L)$,

$$\lim_{\alpha\to 1} \mathcal{C}_\alpha(\gamma) = \mathcal{C}_{ea}(\gamma), \qquad \lim_{\alpha\to 1} \mathcal{D}^k_\alpha(\gamma) = \mathcal{D}^k_{ea}(\gamma),\ k=1,\dots,K \tag{3.12}$$

uniformly in γ. This establishes conditions (A.2) and (A.3).

The set $S(L)$ is clearly convex. As the costs are linear in γ, conditions (A.4) and (A.5) hold. Since γ is bounded in a simplex, this implies condition (A.6).

It follows from condition (2.3) and from Theorem 3.3 that there exists some $\eta > 0$ and some $\gamma \in S(L)$ such that the policy $m(\gamma)$ satisfies

$$\mathcal{D}^k_{ea}(\gamma) = D^k_{ea}(\beta, m(\gamma)) \le V^k - \eta, \qquad k = 1,\dots,K.$$

This establishes condition (A.7). Finally, condition (A.8) trivially holds, as $V = V_n$ do not depend on n. □

4. Concluding remarks. We have presented sufficient conditions for the continuity of the optimal values of constrained optimization problems and established several results on convergence of optimal solutions. Using these general tools, we obtained a new result for the convergence of discounted MDPs to the expected average cost one, under a general multichain ergodic structure. This was done by showing that one could restrict without loss of optimality to some subclass of policies, and then that an equivalent LP can be used to compute the values and the optimal policies. Since our results in the first part of the paper hold for convex programs and not just for LPs, this suggests that they could be used for establishing convergence properties in control problems with more complex cost functions.

One can further use the continuity of the optimal values and solutions of constrained opimtization to obtain other important features in constrained control of Markov chains. One can obtain convergence of the values and optimal policies of finite horizon problems to infinite ones, and one can establish the convergence of problems with finite state spaces to those with infinite state spaces (see [1, 3]). Another interesting question is on the structure of optimal policies in CMDPs. Ross [38] has shown that CMDPs with finite state and action spaces have optimal stationary policies that require at most $K+1$ randomizations, where K is the number of

constraints. Borkar [15] has extended this to a countable state space. Our approach allows us to obtain the same result as Borkar using the fact that limits of policies that are optimal for truncated (finite state) problems (which can be chosen with no more than $K+1$ randomization, according to [38]) are optimal for the original countable state problem (see details in [1, 3]).

Appendix A. Solutions of convex optimization. Let X be a convex subset of a topological vector space $V_n = (V_n^1, \dots, V_n^K) \in \mathbb{R}^K$, $n = 1, 2, \dots, \infty$, $(V_\infty = V)$ and

$$C_n : X \to \mathbb{R}, \quad n = 1, \dots, \infty \quad (C_\infty = C),$$

$$D_n : X \to \mathbb{R}^K, \quad n = 1, \dots, \infty \quad (D_\infty = D),$$

$$\Delta_n = \{x \in X : D_n(x) \le V_n\}, \quad n = 1, 2, \dots, \infty \quad (\Delta_\infty = \Delta).$$

We define the values of the constrained problems:

$$R_n = \inf_{u \in \Delta_n} C_n(u); \qquad R = \inf_{u \in \Delta} C(u).$$

We want to answer the following questions.

(i) Does $R_n \to R$ when $n \to \infty$?

(ii) Convergence of policies: Let $\pi : X \to \Delta$, and fix some $\epsilon \ge 0$. Let ϵ_n be a sequence of positive real numbers such that $\overline{\lim}_{n\to\infty} \epsilon_n \le \epsilon$. Assume that u_n^* is an ϵ_n-optimal policy for the nth optimal cost function R_n. Is $\pi(u_n^*)$ "almost" optimal for the limit optimal cost function R, for n large enough?

(iii′) Robustness of the optimal policy: If u^* is ϵ-optimal for the limit optimal cost function, can we derive of it an "almost" optimal policy for the nth approximating optimal cost function for all n large enough?

(iv′) Let $\overline{u} \in X$ be some limit point of u_n^*, defined above. Is $\overline{u}$ ϵ-optimal for the limit optimal cost function?

We can give an answer to questions (i) and (iv′) using the notion of Γ-convergence [16]. In fact, let X be a topological space, $\mathcal{N}(u)$ the set of all open neighborhoods of $u \in X$, and F_n a sequence of functions from X to $\mathbb{R} \cup \infty$.

We define

$$\Gamma - \underline{\lim}_{n\to\infty} F_n(u) = \sup_{V \in \mathcal{N}(u)} \underline{\lim}_{n\to\infty} \inf_{y \in V} F_n(y),$$

$$\Gamma - \overline{\lim}_{n\to\infty} F_n(u) = \sup_{V \in \mathcal{N}(u)} \overline{\lim}_{n\to\infty} \inf_{y \in V} F_n(y).$$

We say that F_n Γ-converge to F if and only if

$$\Gamma - \overline{\lim}_{n\to\infty} F_n(u) \le F(u) \le \Gamma - \underline{\lim}_{n\to\infty} F_n(u). \tag{A.1}$$

It is known that if there exists F that verifies (A.1) we obtain properties about minima and minimizers of function F (see [16]). As these properties are related with questions (i) and (iv′) of our problem, we are going to rewrite them in the context of Γ-convergence and ask for some assumptions on the data problem in order to obtain (A.1) for an appropriate definition of F_n and F.

We assume there exists $M_1 > 0$ such that

$$\lim_{n\to\infty} D_n = D, \quad \text{uniformly in} \quad \bigcup_{k\geq M_1} \Delta_k \cup \Delta, \tag{A.2}$$

$$\lim_{n\to\infty} C_n = C, \quad \text{uniformly in} \quad \bigcup_{k\geq M_1} \Delta_k \cup \Delta, \tag{A.3}$$

$$D : X \to \mathbb{R}^K \quad \text{is a lower semicontinuous and convex function,} \tag{A.4}$$

$$C : X \to \mathbb{R} \quad \text{is a lower semicontinuous and convex function,} \tag{A.5}$$

$$\exists M > 0 \quad \text{such that} \quad -M \leq C(x) \quad \forall x \in X, \tag{A.6}$$

$$\exists v \in U, \exists \eta > 0 \quad \text{such that} \quad D^k(v) \leq V^k - \eta \ \forall k = 1, \ldots, K, \tag{A.7}$$

$$V_n \to V, \quad n \to \infty. \tag{A.8}$$

We shall denote by (**H**) the set of hypotheses (A.2)–(A.8).

For any vector $\mathcal{V} \in \mathbb{R}^K$ and any constant $v \in \mathbb{R}$, we shall understand below $\mathcal{V} + v$ to mean the vector in $\mathbb{R}^K$ obtained by adding the constant v to each of the components of $\mathcal{V}$. We shall say that $x \in X$ is ϵ-optimal for R_n if $x \in \Delta_n$ and $C_n(x) \leq R_n + \epsilon$.

Appendix B. Key theorems for approximations. In this section we shall prove the approximation theorems.

Let us define

$$F_n(u) = \begin{cases} C_n(u) & \text{if} \quad u \in \Delta_n, \\ \infty & \text{if} \quad u \notin \Delta_n \end{cases} \tag{B.1}$$

and

$$F(u) = \begin{cases} C(u) & \text{if} \quad u \in \Delta, \\ \infty & \text{if} \quad u \notin \Delta. \end{cases} \tag{B.2}$$

We first establish the Γ-converge of F_n.

THEOREM B.1. *If* (**H**) *holds, then* F_n Γ*-converge to* F *as* $n \to \infty$.

Proof. Let x be such that $D(x) < V$ (then $F(x) = C(x)$). Let ϵ be a positive real number. Because of the lower semicontinuity (l.s.c.) of C there exists $U \in \mathcal{N}(x)$ such that

$$C(x) - \epsilon < C(y) \qquad \forall y \in U. \tag{B.3}$$

Now, given $\delta > 0$, we can take $N_1 \geq M_1$ such that if $n \geq N_1$, then, from (A.3),

$$|C_n(y) - C(y)| < \delta \qquad \text{when} \quad y \in \bigcup_{k\geq M_1} \Delta_k \cup \Delta, \tag{B.4}$$

and finally, by (A.2) and (A.8), it is possible to choose $N \geq N_1$ such that $x \in \Delta_n$ for all $n \geq N$. Then we have

$$\inf_{y \in U} F_n(y) = \inf_{y \in U \cap \Delta_n} C_n(y) \qquad \text{when} \quad n \geq N. \tag{B.5}$$

Taking into account the relations

$$U \cap \Delta_n \subset \bigcup_{k \geq M_1} \Delta_k \cup \Delta,$$

$$U \cap \Delta_n \subset U,$$

we obtain, from (B.3) and (B.4) for all $n \geq N$,

$$C(x) - \epsilon - \delta \leq \inf_{y \in U \cap \Delta_n} C(y) - \delta \leq \inf_{y \in U \cap \Delta_n} C_n(y),$$

and from the arbitrariness of δ it follows that

$$C(x) - \epsilon \leq \liminf_{n \to \infty} \inf_{y \in U} F_n. \tag{B.6}$$

On the other hand, if $V \in \mathcal{N}(x)$ and $n \geq N$, then

$$\inf_{y \in V} F_n(y) \leq C_n(x),$$

and taking into account that $C_n(x)$ converge to $C(x)$, we obtain

$$\limsup_{n \to \infty} \inf_{y \in V} F_n(x) \leq C(x) \qquad \forall V \in \mathcal{N}(x). \tag{B.7}$$

Finally, from (B.6) and (B.7) and because ϵ is also arbitrarily chosen, we deduce

$$\sup_{V \in \mathcal{N}(x)} \liminf_{n \to \infty} \inf_{y \in V} F_n(y) = C(x),$$

$$\sup_{V \in \mathcal{N}(x)} \limsup_{n \to \infty} \inf_{y \in V} F_n(y) = C(x);$$

that is,

$$(\Gamma - \lim F_n)(x) = F(x).$$

Now, we consider the second case: let x be such that $D(x) > V$ (then $F(x) = \infty$). Because of the l.s.c. of D, given $\lambda > 0$, there exists $U \in \mathcal{N}(x)$ such that if ϵ is sufficiently small, then

$$V + \epsilon < D(x) - \lambda < D(y) \qquad \forall y \in U.$$

Let $N \geq M_1$ be such that (by hypotheses (A.8) and (A.2))

$$|V - V_n| < \frac{\epsilon}{2} \qquad \text{and} \qquad |D(y) - D_n(y)| < \frac{\epsilon}{2} \qquad \forall y \in \bigcup_{k \geq M_1} \Delta_k \cup \Delta$$

whenever $n \geq N$. Therefore, if $n \geq N$ and $y \in U \cap (\bigcup_{k \geq M_1} \Delta_k \cup \Delta)$, then $D_n(y) > V_n$ and then $F_n(y) = \infty$. On the other hand, if $y \in U \setminus (\bigcup_{k \geq M_1} \Delta_k \cup \Delta)$, we also have $F_n(y) = \infty$. Thus, for all $n \geq N$

$$F_n(y) = \infty \qquad \forall y \in U,$$

and from that it follows that

$$\liminf_{n \to \infty} \inf_{y \in U} F_n(y) = \limsup_{n \to \infty} \inf_{y \in U} F_n(y) = \infty;$$

that is,

$$(\Gamma - \lim F_n)(x) = F(x).$$

Now, we consider the last case. Let x be such that $D(x) = V$. We will use the following property: If $U \in \mathcal{N}(x)$ and $\delta > 0$, there exists $y \in U$ such that

$$D(y) < V \qquad \text{and} \qquad C(y) < C(x) + \delta. \tag{B.8}$$

In fact, because of (A.7) there exists $v \in \Delta$ such that $D(v) < V$. From the convexity of C and D it follows that the segment $[v, x]$ is contained in Δ, and we have

$$D((1-t)v + tx) \leq (1-t)D(v) + tD(x) < V,$$

$$C((1-t)v + tx) \leq (1-t)C(v) + tC(x)$$

for all $t \in [0, 1]$. Then we can choose t sufficiently close to 1 and $y = (1-t)v + tx$ such that $y \in U$ and $C(y) < C(x) + \delta$.

Now, let $V \in \mathcal{N}(x)$, and let $\delta > 0$. We can choose $y \in V$ such that (B.8) holds. Then, because of (A.2) and (A.8), there exists $k \in \mathbb{N}$ such that $y \in \Delta_n$ for all $n \geq k$. Then

$$\liminf_n \inf_{x \in V} F_n(x) \leq \liminf_n C_n(y) = C(y) \leq C(x) + \delta,$$

and from the arbitrariness of δ it follows that

$$\liminf_n \inf_{x \in V} F_n(x) \leq C(x). \tag{B.9}$$

Let ϵ be a positive real number. Let $U \in \mathcal{N}(x)$ be such that (because of the l.s.c. of C)

$$C(x) - \frac{\epsilon}{2} < C(y) \qquad \forall y \in U, \tag{B.10}$$

and let $x_0 \in U$ be such that (by the property proved above)

$$D(x_0) < V.$$

Then x_0 satisfies $C(x) - \epsilon < C(x_0)$. There exists (by assumptions (A.2) and (A.8)) $N_1 \geq M_1$ such that $D_n(x_0) < V_n$ for all $n \geq N_1$. Finally, taking into account (B.10) and (A.3), we can take $N \geq N_1$ such that if $n \geq N$, then

$$C(x) - \epsilon < C_n(y) \qquad \forall y \in U \cap \Big(\bigcup_{k \geq M_1} \Delta_k \cup \Delta \Big). \tag{B.11}$$

Then we have

$$C(x) - \epsilon \leq \liminf_{n\to\infty} \inf_{y\in U} F_n(y),$$

where the last inequality follows from (B.11) and the relation

$$x_0 \in U \cap \Delta_n \subset U \cap \Big(\bigcup_{k\geq M_1} \Delta_k \cup \Delta \Big) \qquad \forall n \geq N.$$

Now, because of the arbitrariness of ϵ, we have

$$\sup_{V\in\mathcal{N}(x)} \liminf_{n} \inf_{x\in V} F_n(x) \geq C(x). \tag{B.12}$$

From (B.9) it holds that for all $V \in \mathcal{N}(x)$ and (B.12) we have

$$(\Gamma\text{-}\liminf_n F_n)(x) = F(x). \qquad \square$$

The Γ-convergence of F_n to F is proved. With some additional hypotheses we can answer (i) (Theorem B.3) and (iv) (Theorem B.5). Theorem B.4 gives an answer to (iv) for $\epsilon = 0$.

DEFINITION B.2. *We say that the sequence (G_n) is equicoercive if for all $t \in \mathbb{R}$ there exists a closed countably compact subset K_t of X such that $\{G_n \leq t\} \subseteq K_t$ for every $n \in \mathbb{N}$.*

THEOREM B.3 (see Dal Maso [16]). *Suppose that (G_n) is equicoercive in X and that Γ-converge to a function G in X. Then*

$$\min_{x\in X} G(x) = \lim_{n\to\infty} \inf_{x\in X} G_n(x).$$

THEOREM B.4 (see Dal Maso [16]). *Assume that (F_h) Γ-converge to a function F in X. For every $h \in \mathbb{N}$, let x_h be a minimizer of F_h in X (or, more generally, an ϵ_h-minimizer, where (ϵ_h) is a sequence of real numbers converging to 0). If x is a cluster point of (x_h), then x is a minimizer of F in X, and*

$$F(x) = \limsup_{h\to\infty} F_h(x_h).$$

If (x_h) converges to x in X, then x is a minimizer of F in X, and

$$F(x) = \lim_{h\to\infty} F_h(x_h).$$

THEOREM B.5 (see Rockafellar–Wets [37]). *Let $X = \mathbb{R}^N$. Suppose that (f_n) Γ-converge to f with $-\infty < \inf f < \infty$. Then we have the following.*

1. *$\inf f_n \to \inf f$ if and only if there exist for every $\epsilon > 0$ a compact set $B \subset \mathbb{R}^N$ and $k \in \mathbb{N}$ such that*

$$\inf_B f_n \leq \inf f_n + \epsilon \qquad \forall n \geq k.$$

2. *For all $\epsilon \geq 0$,*

$$\limsup_n (\epsilon\text{-}\operatorname{argmin} f_n) \subset \epsilon\text{-}\operatorname{argmin} f,$$

and consequently,

$$\limsup_n (\epsilon_n\text{-}\operatorname{argmin} f_n) \subset \operatorname{argmin} f \quad \textit{whenever} \quad \epsilon_n \searrow 0.$$

3. *Under the assumption that* $\inf f_n \to \inf f$, *there exists a sequence* $\epsilon_n \searrow 0$ *such that* $\epsilon_n-\operatorname{argmin} f_n \to \operatorname{argmin} f$. *Conversely, if such a sequence exists, and if* $\operatorname{argmin} f \neq \emptyset$, *then* $\inf f_n \to \inf f$.

COROLLARY B.6. *Suppose* (f_n) *is an equicoercive sequence of real functions in* $\mathbb{R}^n$ *that* Γ*-converges to* f. *Let* $\epsilon > 0$ *and* $\delta > 0$, *and let* $x_n \in \epsilon-\operatorname{argmin} f_n$. *Then there exists* $N \in \mathbb{N}$ *such that*

$$B(x_n, \delta) \cap \epsilon-\operatorname{argmin} f \neq \emptyset$$

whenever $n \geq N$.

Proof. If it is not true, then there exists a subsequence (x_{n_k}) such that

$$B(x_{n_k}, \delta) \cap \epsilon-\operatorname{argmin} f = \emptyset. \tag{B.13}$$

Now, if $t = \epsilon + \sup(\inf_{x\in\mathbb{R}_n} f_n, n \in \mathbb{N})$ (it is finite), then there exists a compact set K_t such that $\{f_n \leq t\} \subseteq K_t$ for every $n \in \mathbb{N}$, and then the sequence (x_{n_k}) has a limit point x. By statement 2 of Theorem B.5, $x \in \epsilon-\operatorname{argmin} f$. But, if k is sufficiently large, then $x \in B(x_{n_k}, \delta)$, and this contradicts (B.13). □

We cannot answer (iii), as we cannot directly construct the function of the type of π from the Γ-convergence. But we can answer the following question:

(iii″) If x_0 is ϵ-optimal for R, is it possible to find an arbitrarily small neighborhood of x_0 and an ϵ-optimal policy $x(n)$ for R_n for n sufficiently large?

THEOREM B.7. *Suppose that the sequence* (F_n) Γ*-converges to a function* F *and*

$$\min_{x\in X} F(x) = \lim_n \inf_{x\in X} F_n(x). \tag{B.14}$$

If $x_0 \in \epsilon-\operatorname{argmin} F$, $U \in \mathcal{N}(x_0)$, *and* $\bar{\epsilon} > \epsilon$, *then there exists* $N \in \mathbb{N}$ *such that*

$$U \cap \bar{\epsilon}-\operatorname{argmin} F_n \neq \emptyset \qquad \textit{when} \quad n \geq N.$$

Proof. By definition,

$$F(x_0) = \sup_{V\in\mathcal{N}(x_0)} \liminf_n \inf_{y\in V} F_n(y) = \sup_{V\in\mathcal{N}(x_0)} \limsup_n \inf_{y\in V} F_n(y).$$

Then, if $U \in \mathcal{N}(x_0)$, we have

$$\limsup_n \inf_{y\in U} F_n(y) \leq F(x_0).$$

Thus, given $\lambda > 0$, there exists N_1 such that

$$\inf_{y\in U} F_n(y) \leq F(x_0) + \lambda \qquad \text{when} \quad n \geq N_1.$$

For each $n \geq N_1$ we can choose $y_n \in U$ such that

$$F_n(y_n) \leq F(x_0) + 2\lambda \qquad \text{when} \quad n \geq N_1. \tag{B.15}$$

On the other hand, it follows from (B.14) that there exists $N \geq N_1$ such that

$$F(x_0) - \epsilon - \lambda \leq \inf_{x\in X} F_n(x) \qquad \text{when} \quad n \geq N. \tag{B.16}$$

Finally, from (B.15) and (B.16), we obtain

$$0 \leq F_n(y_n) - \inf_{x\in X} F_n(x) \leq \epsilon + 3\lambda$$

for all $n \geq N$, and then, by choosing λ sufficiently small, we have

$$y_n \in U \cap \bar{\epsilon}\text{−argmin}F_n \qquad \forall n \geq N. \qquad \square$$

To answer questions (ii) and (iii), we present the following definitions and results.

DEFINITION B.8. *For each δ such that $0 < \delta < \eta$ (where η is defined in* (A.7)*) and for each $u \in X$ we set*

$$\epsilon_\delta(u) = \min\{\lambda \ : \ \lambda \in [0,1], \text{and } \lambda D(v) + (1-\lambda)D(u) \leq V - \delta\},$$

and

$$\pi_\delta(u) = \epsilon_\delta(u)v + (1 - \epsilon_\delta(u))u. \tag{B.17}$$

Remark B.9. $\pi_\delta(u) \in \Delta$, because of the convexity of X we have $\pi_\delta(u) \in X$, and by the definition of $\epsilon_\delta(u)$ and the fact that D is a convex function (A.4), we have

$$D(\pi_\delta(u)) \leq V - \delta. \tag{B.18}$$

Remark B.10. Let δ be such that $0 < \delta < \eta$. If (A.2), (A.4), (A.7), and (A.8) hold, there exists $\overline{N}(\delta)$ such that $\pi_\delta(u) \in \Delta_n$ if $n \geq \overline{N}(\delta)$ $\forall u \in X$. In fact, by (A.2) there exists $N(\delta)$ such that if $n \geq N(\delta)$, $D_n(\pi_\delta(u)) \leq D(\pi_\delta(u)) + \delta/2 \ \ \forall u \in X$, and by (B.18)

$$D_n(\pi_\delta(u)) \leq D(\pi_\delta(u)) + \frac{\delta}{2} \leq V - \delta + \frac{\delta}{2} \leq V - \frac{\delta}{2}.$$

Then, by (A.8) we have that there exists $\hat{N}$ such that $V - \delta/2 \leq V_n$ if $n \geq \hat{N}$, so, we have $D_n(\pi_\delta(u)) \leq V_n$ for all $n \geq \overline{N}(\delta) = \max(N(\delta), \hat{N})$.

LEMMA B.11. *Let δ be such that $0 < \delta < \eta$. If* (A.2)*,* (A.7)*, and* (A.8) *hold, then*

$$\limsup_{n\to\infty} \left\{ \sup_{u\in\Delta_n} \epsilon_\delta(u) \right\} \leq \frac{\delta}{\eta}.$$

Proof. By definition of Δ_n we have that $\forall u \in \Delta_n$, $D_n(u) \leq V_n$. By (A.2) and (A.8) we have

$$\forall \hat{\epsilon} > 0 \quad \exists N(\hat{\epsilon}) \text{ such that } \quad \forall n \geq N(\hat{\epsilon}), \forall u \in \Delta_n, \quad D(u) \leq V + 2\hat{\epsilon}. \tag{B.19}$$

By (A.7) and (B.19) we have $\forall \hat{\epsilon} > 0$, $u \in \Delta_n$, $\lambda \in [0,1]$,

$$\lambda D(v) + (1-\lambda)D(u) \leq \lambda(V - \eta) + (1-\lambda)(V + 2\hat{\epsilon}). \tag{B.20}$$

But

$$\lambda(V-\eta) + (1-\lambda)(V+2\hat{\epsilon}) \leq V - \delta \iff \lambda \geq \frac{2\hat{\epsilon} + \delta}{2\hat{\epsilon} + \eta} \quad \forall \hat{\epsilon} > 0. \tag{B.21}$$

By (B.20) and (B.21) we obtain

$$\left[\frac{2\hat{\epsilon} + \delta}{2\hat{\epsilon} + \eta}, 1\right] \subseteq \{\lambda \in [0,1] : \lambda D(v) + (1-\lambda)D(u) \leq V - \delta\} \quad \forall \hat{\epsilon} > 0 \ \ \forall u \in \Delta_n,$$

and then, by definition of $\epsilon_\delta(u)$, we have

$$\epsilon_\delta(u) \le \frac{2\hat{\epsilon}+\delta}{2\hat{\epsilon}+\eta} \qquad \forall u \in \Delta_n, \quad n \ge N(\hat{\epsilon}) \quad \forall \hat{\epsilon} > 0.$$

From this last inequality we deduce

$$\sup_{u\in\Delta_n} \epsilon_\delta(u) \le \frac{2\hat{\epsilon}+\delta}{2\hat{\epsilon}+\eta} \qquad n \ge N(\hat{\epsilon}) \qquad \forall \hat{\epsilon} > 0,$$

and this last inequality implies

$$\limsup_{n\to\infty} \left[\sup_{u\in\Delta_n} \epsilon_\delta(u) \right] \le \frac{\delta}{\eta}. \qquad \square$$

Remark B.12. Let δ be such that $\eta > \delta > 0$. In the same way of Lemma B.11, and if (A.7) holds, we can prove that

$$\sup_{u\in\Delta} \epsilon_\delta(u) \le \frac{\delta}{\eta}.$$

For the proof it is only necessary to remark that (B.20) and (B.21) become

$$\lambda D(v) + (1-\lambda)D(u) \le \lambda(V-\eta) + (1-\lambda)V,$$

$$\lambda(V-\eta) + (1-\lambda)V \le V - \delta \Longleftrightarrow \lambda \ge \frac{\delta}{\eta}.$$

LEMMA B.13. *Let δ be such that $\eta > \delta > 0$. If* (A.2), (A.5), (A.6), (A.7), *and* (A.8) *hold, then*

$$\limsup_{n\to\infty} \left[\sup_{u\in\Delta_n} \{C(\pi_\delta(u)) - C(u)\} \right] \le (C(v)+M)\frac{\delta}{\eta}.$$

Proof. $\forall u \in \Delta_n$ by (A.5) and (B.17) we obtain

$$C(\pi_\delta(u)) - C(u) \le \epsilon_\delta(u)C(v) + (1-\epsilon_\delta(u))C(u) - C(u) = \epsilon_\delta(u)[C(v) - C(u)].$$

Then by this last equation, (A.6), and Lemma B.11 we have

$$\limsup_{n\to\infty} \sup_{u\in\Delta_n} \epsilon_\delta(u)[C(v)-C(u)] \le (C(v)+M) \limsup_{n\to\infty} \sup_{u\in\Delta_n} \epsilon_\delta(u) \le (C(v)+M)\frac{\delta}{\eta}. \qquad \square$$

Remark B.14. Let δ be such that $0 < \delta < \eta$. By Remark B.12 and if (A.5) and (A.7) hold, we can obtain

$$\sup_{u\in\Delta} \{C(\pi_\delta(u)) - C(u)\} \le (C(v)+M)\frac{\delta}{\eta}.$$

THEOREM B.15. *Suppose that* (**H**) *holds and let F_n and F defined by* (B.1) *and* (B.2), *respectively. We also suppose*

$$\lim_n \inf_{x\in X} F_n(x) = \inf_{x\in X} F(x).$$

Then,

1. *If* $u \in \epsilon-\text{argmin}F$ *and* $\lambda > 0$*, there exist* $N \in \mathbb{N}$ *and* $\delta > 0$ *such that if* $n \geq N$*, then*

$$\pi_\delta(u) \in (\epsilon+\lambda)-\text{argmin}F_n. \tag{B.22}$$

2. *Let* $u_n \in \epsilon-\text{argmin}F_n$ *and* $\lambda > 0$*. There exist* $N \in \mathbb{N}$ *and* $\delta_0 > 0$ *such that if* $n \geq N$ *and* $\delta < \delta_0$*, then*

$$\pi_\delta(u_n) \in (\epsilon+\lambda)-\text{argmin}F. \tag{B.23}$$

Proof.

1. We have

$$C_n(\pi_\delta(u)) - C(u) = [C_n(\pi_\delta(u)) - C(\pi\delta(u))] + [C(\pi_\delta(u)) - C(u)].$$

From the convergence of (C_n) to C, Remark B.14, and taking into account that for each $\delta > 0$ if n is sufficiently large then $\pi_\delta(u) \in \Delta$, it follows that if $\lambda > 0$ is given, we can choose $N_1 \in \mathbb{N}$ and $\delta > 0$ sufficiently small, such that if $n > N$, then each term in the right side of the last equation is less than $\frac{\lambda}{3}$, and the equation becomes

$$C_n(\pi_\delta(u)) - C(u) < \frac{2}{3}\lambda.$$

Then, keeping in mind the ϵ-optimality of u, we have

$$C_n(\pi_\delta(u)) < C(u) + \frac{2}{3}\lambda < \inf_{x\in X} F(x) + \epsilon + \frac{2}{3}\lambda < \inf_{x\in X} F_n(x) + \epsilon + \lambda$$

if $n \geq N$ for some $N \geq N_1$, and that is (B.22).

2. We have

$$C(\pi_\delta(u_n)) - C_n(u_n) = [C(\pi_\delta(u_n)) - C(u_n)] + [C(u_n) - C_n(u_n)].$$

From Lemma B.13 and the convergence of (C_n) to C, if $\lambda > 0$ is given, we can choose $N_1 \in \mathbb{N}$ and $\delta > 0$ sufficiently small, such that if $n > N_1$, then each term in the right side of the last equation is less than $\frac{\lambda}{3}$, and the equation becomes

$$C(\pi_\delta(u_n)) - C_n(u_n) < \frac{2}{3}\lambda.$$

Then, keeping in mind the ϵ-optimality of u_n, we have

$$C(\pi_\delta(u_n)) < C_n(u_n) + \frac{2}{3}\lambda < \inf_{x\in X} F_n(x) + \epsilon + \frac{2}{3}\lambda < \inf_{x\in X} F(x) + \epsilon + \lambda$$

if $n \geq N$ for some $N \geq N_1$, and that is (B.23). ∎

PROPOSITION B.16. *Suppose* **(H)** *holds. We set* $K_{\epsilon,n} = \epsilon-\text{argmin}F_n$*,* $K_\epsilon = \epsilon-\text{argmin}F$*,* $R_n = \inf_{x\in X} C_n(x)$*, and* $R = \inf_{x\in X} C(x)$*. Then*

$$K_{\epsilon,n} \cap K_\epsilon \neq \emptyset.$$

Proof. Let $x \in K_{\frac{\epsilon}{4}} \subset K_\epsilon$; then $D(x) \le V$. We will consider the case $D(x) = V$. The case $D(x) < V$ is similar and simpler. Let $w \in \Delta$ be such that $D(w) < V$. Then the segment $[w, x]$ is contained in Δ. Besides,

$$C((1-t)w + tx) \le (1-t)C(w) + tC(x) < C(x) + \frac{\epsilon}{4}$$

if $t \in [T, 1]$, for some T with $0 < T < 1$. We take a such t, and we put $x_0 = (1-t)w + tx$. Then $x_0 \in K_{\frac{\epsilon}{2}} \subset K_\epsilon$, and $D(x_0) < V$, from which it easily follows that $x_0 \in \Delta_n$ if $n \ge N_1$ for some $N_1 \in \mathbb{N}$. We can take $N \ge N_1$ such that for every $n \ge N$

$$|C_n(x_0) - C(x_0)| < \frac{\epsilon}{4} \qquad \text{and} \qquad |R_n - R| < \frac{\epsilon}{4}.$$

Therefore, if $n \ge N$, we have

$$C_n(x_0) < C(x_0) + \frac{\epsilon}{4} < R + \frac{3}{4}\epsilon < R_n + \epsilon,$$

that is $x_0 \in K_{\epsilon,n}$, and then

$$K_\epsilon \cap K_{\epsilon,n} \neq \emptyset \qquad \forall n \ge N. \qquad \square$$

REFERENCES

[1] E. Altman, *Asymptotic properties of constrained Markov decision processes*, Z. Oper. Res., 37 (1993), pp. 151–170.

[2] E. Altman, *Denumerable constrained Markov decision problems and finite approximations*, Math. Oper. Res., 19 (1994), pp. 169–191.

[3] E. Altman, *Constrained Markov Decision Processes*, in Stochastic Modeling, Chapman and Hall/CRC, Boca Raton, FL, 1999.

[4] E. Altman and V. A. Gaitsgory, *Stability and singular perturbations in constrained Markov decision problems*, IEEE Trans. Automat. Control, 38 (1993), pp. 971–975.

[5] E. Altman and A. Shwartz, *Optimal priority assignment: A time sharing approach*, IEEE Trans. Automat. Control, AC-34 (1989), pp. 1089–1102.

[6] E. Altman and A. Shwartz, *Adaptive control of constrained Markov chains*, IEEE Trans. Automat. Control, 36 (1991), pp. 454–462.

[7] E. Altman and A. Shwartz, *Markov decision problems and state-action frequencies*, SIAM J. Control Optim., 29 (1991), pp. 786–809.

[8] E. Altman and A. Shwartz, *Sensitivity of constrained Markov decision problems*, Ann. Oper. Res., 32 (1991), pp. 1–22.

[9] E. Altman and A. Shwartz, *Time-sharing policies for controlled Markov chains*, Oper. Res., 41 (1993), pp. 1116–1124.

[10] E. Altman and F. Spieksma, *The linear program approach in Markov decision problems revisited*, Z. Oper. Res., 42 (1995), pp. 169–188.

[11] F. J. Beutler and K. W. Ross, *Optimal policies for controlled Markov chains with a constraint*, J. Math. Anal. Appl., 112 (1985), pp. 236–252.

[12] F. J. Beutler and K. W. Ross, *Time-average optimal constrained semi-Markov decision processes*, Adv. Appl. Probab., 18 (1986), pp. 341–359.

[13] J. R. Birge and R. J. Wets, *Designing approximatin schemes for stochastic optimization problems*, Math. Programming Stud., 27 (1986), pp. 54–102.

[14] V. S. Borkar, *A convex analytic approach to Markov decision processes*, Probab. Theory Related Fields, 78 (1988), pp. 583–602.

[15] V. S. Borkar, *Ergodic control of Markov chains with constraints—the general case*, SIAM J. Control Optim., 32 (1994), pp. 176–186.

[16] G. Dal Maso, *An Introduction to Γ-Convergence*, Progr. Nonlinear Differential Equations Appl. 8, Birkhäuser, Basel, Boston, 1993.

[17] G. B. Dantzig, J. Folkman, and N. Shapiro, *On the continuity of the minimum set of a continuous function*, J. Math. Anal. Appl., 17 (1967), pp. 519–548.

[18] C. Derman and R. E. Strauch, *A note on memoryless rules for controlling sequential control processes*, Ann. Math. Statist., 37 (1966), pp. 276–278.
[19] E. Dynkin and A. Yushkevich, *Controlled Markov Processes*, Springer-Verlag, Berlin, 1979.
[20] E. A. Feinberg, *Sufficient classes of strategies in discrete dynamic programming* I: *Decomposition of randomized strategies and embedded models*, SIAM Theory Probab. Appl., 31 (1986), pp. 658–668.
[21] E. A. Feinberg, *Constrained semi-Markov decision processes with average rewards*, Z. Oper. Res., 39 (1995), pp. 257–288.
[22] E. A. Feinberg and M. I. Reiman, *Optimality of randomized trunk reservation*, Probab. Engrg. Inform. Sci., 8 (1994), pp. 463–489.
[23] A. V. Fiacco, *Convergence properties of local solutions of convex optimization problems*, J. Optim. Theory Appl., 13 (1974), pp. 1–12.
[24] V. A. Gaitsgory and A. A. Pervozvanskii, *Perturbation theory for mathematical programming problems*, J. Optim. Theory Appl., 49 (1986), pp. 389–410.
[25] A. Hordijk and L. C. M. Kallenberg, *Constrained undiscounted stochastic dynamic programming*, Math. Oper. Res., 9 (1984), pp. 276–289.
[26] A. Hordijk and F. Spieksma, *Constrained admission control to a queuing system*, Adv. in Appl. Probab., 21 (1989), pp. 409–431.
[27] L. C. M. Kallenberg, *Linear Programming and Finite Markovian Control Problems*, Mathematical Centre Tracts 148, Amsterdam, 1983.
[28] P. Kanniappan and S. M. A. Sastry, *Uniform convergence of convex optimization problems*, J. Math. Anal. Appl., 96 (1974), pp. 1–12.
[29] N. V. Krylov, *The construction of an optimal strategy for a finite controlled chain*, Theory Probab. Appl., 10 (1965), pp. 45–54.
[30] A. Lazar, *Optimal flow control of a class of queuing networks in equilibrium*, IEEE Trans. Automat. Control, 28 (1983), pp. 1001–1007.
[31] M. B. Lignota and J. Morgan, *Topological existence and stability of min sup problems*, J. Math. Anal. Appl., 151 (1990), pp. 164–180.
[32] M. B. Lignota and J. Morgan, *Convergences of marginal functions with dependent constraints*, Optimization, 23 (1992), pp. 189–213.
[33] R. Lucchetti and R. J. B. Wets, *Convergence of minima of integral functionals, with applications to optimal control and stochastic optimization*, Statist. Decisions, 11 (1993), pp. 69–84.
[34] J. Morgan and R. Raucci, *Continuity properties of ϵ-solutions for generalized parametric saddle point problems and application to hierarchical games*, J. Math. Anal. Appl., 211 (1997), pp. 30–48.
[35] P. Nain and K. W. Ross, *Optimal priority assignment with hard constraint*, IEEE Trans. Automat. Control, 31 (1986), pp. 883–888.
[36] A. A. Pervozvanskii and V. A. Gaitsgory, *Theory of Suboptimal Decision: Decomposition and Aggregation*, Kluwer Academic Publishers, Dordrecht, The Netherlands, 1988.
[37] R. T. Rockafellar and R. J. B. Wets, *Variational Analysis*, Grundlehren Math. Wiss. 317, Springer-Verlag, Berlin, 1998.
[38] K. W. Ross, *Randomized and past-dependent policies for Markov decision processes with multiple constraints*, Oper. Res., 37 (1989), pp. 474–477.
[39] K. W. Ross and B. Chen, *Optimal scheduling of interactive and non interactive traffic in telecommunication systems*, IEEE Trans. Automat. Control, 33 (1988), pp. 261–267.
[40] K. Ross and R. Varadarajan, *Markov decision processes with sample path constraints: The communicating case*, Oper. Res., 37 (1989), pp. 780–790.
[41] K. Ross and R. Varadarajan, *Multichain Markov decision processes with a sample path constraint: A decomposition approach*, Math. Oper. Res., 16 (1991), pp. 195–207.
[42] M. Schal, *A selection theorem for optimization problems*, Arch. Math., 25 (1974), pp. 219–224.
[43] I. E. Schochetman, *Pointwise versions of the maximum theorem with applications to optimization*, Appl. Math. Lett., 3 (1990), pp. 89–92.
[44] I. E. Schochetman and R. L. Smith, *Convergence of selections with applications in optimization*, J. Math. Anal. Appl., 155 (1991), pp. 278–242.
[45] L. I. Sennott, *Constrained discounted Markov decision chains*, Probab. Engrg. Inform. Sci., 5 (1991), pp. 463–475.
[46] L. I. Sennott, *Constrained average cost Markov decision chains*, Probab. Engrg. Inform. Sci., 7 (1993), pp. 69–83.

SIAM J. CONTROL OPTIM.
Vol. 38, No. 4, pp. 1223–1240

ADAPTIVE IMAGE RECONSTRUCTION USING INFORMATION MEASURES*

ULRIKE HERMANN† AND DOMINIKUS NOLL‡

Abstract. We present a class of nonlinear adaptive image restoration filters which may be steered to preserve sharp edges and contrasts in the restorations. From a theoretical point of view we discuss the associated variational problems and prove existence of solutions in certain Sobolev spaces $W^{1,p}$ or in a BV-space. The degree of regularity of the solution may be understood as a mathematical explanation of the heuristic properties of the designed filters.

Key words. adaptive image restoration filter, image restoration, variational methods

AMS subject classifications. 49J52, 49N15, 94A17

PII. S0363012997324338

1. Motivation and purpose. Various inverse problems require reconstructing an unknown density function $u(x)$, $x \in \Omega \subset \mathbf{R}^n$, from a finite number of measurements of the form

$$\int_\Omega \big(a_k(x)u(x) + b_k(x) \cdot \nabla u(x)\big)\, dx \;=\; c_k, \qquad k = 1, \dots, N. \tag{1.1}$$

Examples of particular interest are in medical imaging, where the data c_k represent attenuation coefficients of transmission x-rays, or in image restoration, where the c_k are gray levels at pixels k of a blurred version of the true image $u(x)$. Restoring the original $u(x)$ is usually an ill-posed problem, and the inevitable measurement noise may make this a difficult task. One way to restore $u(x)$ in the presence of noise is to stabilize inversion of (1.1) by introducing a regularizing functional of the form

$$I[u] \;=\; \int_\Omega h\big(u(x), \nabla u(x)\big)\, dx, \tag{1.2}$$

closely related to the specific restoration problem. Introducing linear operators A, B by

$$(Au)_k = \int_\Omega a_k(x)u(x)\, dx, \qquad (Bv)_k = \int_\Omega b_k(x) \cdot v(x)\, dx, \tag{1.3}$$

we consider the following inverse methods which we call the *tolerance* and the *penalization approaches,* respectively:

$$(P)_{\mathrm{tol}} \qquad \begin{array}{ll} \text{minimize} & I[u] \\ \text{subject to} & \big|Au + B\nabla u - c\big| \le \varepsilon, \\ & \displaystyle\int_\Omega u(x)\, dx = 1 \end{array}$$

*Received by the editors July 11, 1997; accepted for publication (in revised form) June 1, 1999; published electronically April 20, 2000.
http://www.siam.org/journals/sicon/38-4/32433.html
†Mathematisches Institut A, Universität Stuttgart, Pfaffenwaldring 57, 70569 Stuttgart, Germany.
‡Université Paul Sabatier, Mathématiques pour l'Industrie et la Physique, 118, route de Narbonne, 31062 Toulouse, France (noll@dumbo@ups-tlse.fr).

and

$$(P)_{\text{pen}} \qquad \begin{array}{ll} \text{minimize} & I[u] + \frac{\alpha}{2}|Au + B\nabla u - c|^2 \\ \text{subject to} & \displaystyle\int_\Omega u(x)\,dx = 1 \end{array}$$

($|\cdot|$ = Euclidean norm). A well-known method based on the scheme $(P)_{\text{pen}}$ is *Tychonov-regularization*, where the functional (1.2) is a square norm of $u(x)$ or $\nabla u(x)$ (cf. [16, 12, 13]). However, in image restoration this choice is known to produce poor results, and more sophisticated functionals $I[u]$ are required. In our present work, we shall consider a class of functionals $I[u]$ of information type that are particularly suited for image restoration problems and that we motivate by a heuristic argument. The remaining parts of the paper address the mathematical problems arising from this choice.

The values $u(x)$ are relative gray levels of the unknown image, hence the normalization $\int u\,dx = 1$. Since gray levels are nonnegative, we require reconstructions $u(x) \geq 0$, and this may be guaranteed by our choice of $I[u]$. For the moment consider the model $(P)_{\text{tol}}$. The data being noisy, we should not force equality $Au + B\nabla u = c$, but allow for a tolerance $\varepsilon > 0$, typically estimated using a χ^2-statistics (cf. [18]). The role of the functional $I[u]$ is now to avoid picking highly irregular objects u which would fit the tolerance condition. In other terms, minimizing $I[u]$ subject to the constraint $|Au + B\nabla u - c| \leq \varepsilon$ to some degree means filtering the unknown object $u(x)$. However, as mentioned before, default choices like $I[u] = \int_\Omega |\nabla u|^2\,dx$ tend to smooth away sharp edges in the image. Smoothing while retaining edges is needed, and this requires adapting the filter to the image.

Consider the class of functionals (1.2) defined through the integrands

$$h(u,\xi) = \begin{cases} u\phi\big(\xi/u\big) & \text{if } u > 0, \\ \phi^{0+}(\xi) & \text{if } u = 0, \\ +\infty & \text{otherwise,} \end{cases} \tag{1.4}$$

where $\phi : \mathbf{R}^n \to \mathbf{R}$ is a convex functional and ϕ^{0+} denotes its recession function, needed to render the functional h lower semicontinuous (lsc),

$$\phi^{0+}(\xi) = \sup_{t>0} \frac{\phi(\eta + t\xi) - \phi(\eta)}{t},$$

for an arbitrary fixed η in $\text{dom}\phi$ (cf. [21, p. 66ff]). Then h is jointly convex in (u,ξ), and (1.4) will be called *Csiszár information measures*. An important special case is $\phi(t) = |t|^2$, which is Fisher's information (cf. [19]). Notice that since $h(u,\xi) = +\infty$ for $u < 0$, the objectives (1.4) force nonnegative solutions, as required.

In order to motivate the inverse approach based on (1.4), let us specialize even further by considering functionals of the form $\phi(\xi) = \psi(|\xi|)$ for convex $\psi : \mathbb{R} \to \mathbb{R}$. Since $|\nabla u|$ is invariant under rigid motions, so is $h(u, \nabla u)$ defined through (1.4); hence, this choice will lead to methods invariant under rigid motions of the image. Proceeding in a purely formal way, we first do a change of variables $u(x) = e^{v(x)}$ to account for the condition $u(x) > 0$. The Euler–Lagrange equation for the transformed problem $(P)_{\text{pen}}$ is then

$$\begin{aligned} -\operatorname{div}\left(\frac{\psi'\big(|\nabla v|\big)}{|\nabla v|}\nabla v\right) - \psi'\big(|\nabla v|\big)|\nabla v| + \psi(|\nabla v|) \\ + \alpha(A^T - \operatorname{div}B^T)(Ae^v + B(\nabla e^v) - c) \;=\; 0 \end{aligned} \tag{1.5}$$

with adjoints A^T, B^T, and div B^T defined as

$$A^T(\lambda) = \sum_{k=1}^{N} \lambda_k a_k, \qquad B^T(\lambda) = \sum_{k=1}^{N} \lambda_k b_k, \qquad \mathrm{div}B^T(\lambda) = \sum_{k=1}^{N} \lambda_k \mathrm{div} b_k.$$

Consider the case $\Omega \subset \mathbb{R}^2$. Following an idea originating from [8] and extended in [2, 26], sharp edges (contrasts) in the image $v(x, y)$ occur along level curves $v(x, y) = c$, the indication being that the gradient $\nabla v(x, y)$ becomes large. In this case, smoothing across the edge $v(x, y) = c$ should be dispensed with, while smoothing along the edge is still needed to suppress irregular behavior.

For a point (x, y) on the level curve $v = c$, consider the adapted cartesian coordinates $T(x, y), N(x, y)$ meaning tangential and normal directions to the level curve $v = c$ at (x, y):

$$N(x, y) = \frac{\nabla v(x, y)}{|\nabla v(x, y)|}, \qquad T(x, y) \perp N(x, y).$$

Expanding the divergence term in (1.5) gives

$$-\mathrm{div}\left(\frac{\psi'\big(|\nabla v|\big)}{|\nabla v|}\,\nabla v\right) = -\frac{\psi'(|\nabla v|)}{|\nabla v|}\,\Delta v - \left(\frac{\psi''(|\nabla v|)}{|\nabla v|^2} - \frac{\psi'(|\nabla v|)}{|\nabla v|^3}\right)\nabla v \cdot \nabla^2 v \cdot \nabla v.$$

Observe that the Laplacian is invariant under orthogonal transformations, $\Delta v = v_{xx} + v_{yy} = v_{TT} + v_{NN}$, and secondly that $\frac{\nabla v}{|\nabla v|} \cdot \nabla^2 v \cdot \frac{\nabla v}{|\nabla v|} = v_{NN}$. Then the Euler equation in (T, N)-coordinates reads

$$-\left(\frac{\psi'(|\nabla v|)}{|\nabla v|}\right) v_{TT} - \psi''(|\nabla v|)\, v_{NN} - \psi'(|\nabla v|)|\nabla v| + \psi(|\nabla v|) + \alpha(A^T - \mathrm{div}B^T)(Ae^v + B(\nabla e^v) - c) = 0.$$

Suppose $|\nabla v|$ is small, indicating that $v = c$ is not an edge, and hence smoothing should be encouraged. Assuming (i) $\psi'(0) = 0$ and (ii) $\psi''(0) > 0$, in a neighborhood of (x, y), the Euler equation is qualitatively of the form

$$-\psi''(0)\big(v_{TT} + v_{NN}\big) + \psi(0) + \alpha(A^T - \mathrm{div}B^T)(Ae^v + B(\nabla e^v) - c) = 0.$$

Due to $v_{TT} + v_{NN} = v_{xx} + v_{yy} = \Delta v$, this may be considered as having a strong smoothing effect around (x, y).

Assume, on the other hand, that $|\nabla v|$ is large at (x, y), indicating an edge. Then we wish to smooth in T-direction but not in N-direction. This is achieved, for instance, by having

$$\text{(iii)} \qquad \frac{\psi'(t)}{t} \gg \psi''(t)$$

for large t. The coefficient of v_{NN} then being negligible in a neighborhood of (x, y), the differential equation is qualitatively of the form

$$-C\, v_{TT} - \psi'(|\nabla v|)|\nabla v| + \psi(|\nabla v|) + \alpha(A^T - \mathrm{div}B^T)(Ae^v + B(\nabla e^v) - c) = 0,$$

indicating a preference for smoothing in T-direction, since as before the tendency to smoothing is governed by the second order terms. As an example for (iii), consider a

ψ which for large t behaves like $\psi(t) = t^p$ for some $p > 1$. This gives $\frac{\psi'(t)}{t}/\psi''(t) = 1/(p-1)$, which could be made as large as desired by choosing p close to 1.

The conditions (i)–(iii) do not entirely fix the function ψ, so further evidence (theoretical and numerical) is needed to propose a best choice. The present paper, rather, addresses the theoretical aspects of the models $(P)_{\text{tol}}$ and $(P)_{\text{pen}}$, particularly the question of existence of solutions. Our method of proving existence may be considered a fairly general scheme including a large variety of possible applications. It does not rely on compactness arguments but exploits the convexity of the problems.

A second problem associated with the variational methods $(P)_{\text{pen}}$ and $(P)_{\text{tol}}$ is to justify the Euler–Lagrange equation (1.5), formally derived above. This problem, which is difficult, is treated in [10].

It is intuitively clear that the choices $\psi(t) = |t|^p$ discussed above should lead to image restorations exhibiting more and more sharp edges when $p > 1$ approaches 1. One way to corroborate this in the variational context is by showing that the solutions of the corresponding programs $(P)_{\text{pen}}$ and $(P)_{\text{tol}}$ are in a Sobolev space $W^{1,1+\varepsilon(p)}$, with $\varepsilon(p) \to 0$ as $p \to 1$ (cf. Example 3 at the end of section 6). In the limiting case $p = 1$, we would get solutions which degrade to BV-functions, allowing even for discontinuities. We mention that the latter is sometimes considered as a natural setting for image processing, particularly if the purpose is segmentation or edge detection (cf. [17, 5, 24, 27]).

Numerical experiments for special choices $\psi(t)$ have been presented in [19, 20]. The authors of [2, 26] report experiments with objectives of the form $h(u, \xi) = \phi(|\xi|)$ built on a related philosophy. A comparative study of adaptive filters will be presented elsewhere. We mention that the class of functionals (1.4) has various other applications. See in particular [18] for variational problems involving Fisher's information $(p = 2)$.

2. Outline of the method. We start by giving an outline of our method of proving existence of solutions and then point to the steps which cause particular difficulties. Our approach may be called a *bidual relaxation* scheme: Writing (P) for any of the formulations $(P)_{\text{pen}}$ or $(P)_{\text{tol}}$ and proceeding in a formal way, we first obtain a concave dual program (P^*). Formal means that we do not try to find a dual pair of Banach spaces in which the duality may be justified rigorously. In a second step we repeat the same for (P^*), but this time we use the full convexity machinery. This means we prove a Lagrange multiplier theorem for (P^*). The multiplier $\bar{u}$ is an element of the dual Banach space $M(\bar{\Omega})$ of signed Radon measures and an optimal solution to a properly defined convex bidual (P^{**}). We may therefore interpret $\bar{u}$ as a generalized solution to the original program (P). In a third step we show that under mild additional conditions, we get a solution $\bar{u}$ in a Sobolev space or even in a classical space $C^1(\bar{\Omega})$.

Notice that this scheme has been used various times. However, the difficulties are in the details; in particular, technical problems arise if we are not satisfied with solutions in a BV-space, but wish to prove regularity results (cf. section 4). For complementary literature we refer to [7, 3, 4, 2].

Let us now consider some of the details. First, dualizing (P^*) requires a Lagrange multiplier theorem. This type of result typically needs a constraint qualification hypothesis, which should not be artificial in the light of the original problem (P). The existence result Proposition 4.1 in fact avoids any such hypothesis by providing a solution $u \in M(\bar{\Omega})$, the space of Radon measures.

The second step in our scheme is to show that the generalized solution $u \in M(\bar{\Omega})$

is a BV-function. This is done in Proposition 4.3 and requires a richness hypothesis $(A1)$. Condition $(A1)$ excludes objectives (1.2) where $h(x,\xi)$ is linear in ξ. With $h(x,\xi) = g(x) + \eta \cdot \xi$ linear in ξ, it is possible to construct examples where the generalized solution $\bar{u}$ is not a BV-function, although this may be guaranteed under coercivity assumptions on $g(x)$. We consider objectives (1.2) linear in ξ as of minor importance for possible applications and therefore do not pursue their analysis here.

In a third step of Proposition 4.6, we show that the solution $\bar{u}$, so far a BV-function, is an element of the Sobolev space $W^{1,1}(\Omega)$ if a slightly stronger regularity hypothesis $(A2)$ is satisfied. Hypothesis $(A2)$ may be understood as a weak coercivity condition on h, implying in particular that for fixed x, $h(x,\xi)$ grows stronger than linearly in ξ as $|\xi| \to \infty$.

In practice, it is often enough to have solutions in $W^{1,1}(\Omega)$, in particular, if the natural domain of the functional I_h is a better Sobolev space $W^{1,p}(\Omega)$ for some $p > 1$. Here the solution will automatically be an element of $W^{1,p}(\Omega)$. In section 6 we present an extended version of this observation, showing that under a stronger hypothesis $(A3)$, the solution $\bar{u}$ is improved to be of class $W^{1,p}(\Omega)$ for some $p > 1$, with the possibility to having classical solutions if p is large enough. Hypothesis $(A3)$ is seen to be a coercivity condition on h, satisfied, e.g., if h^* grows at most polynomially (see section 6).

We mention that bidual relaxation as presented here is not aimed at image restoration exclusively. In fact, the hypotheses $(A1)$–$(A3)$ are fairly general and ensure a broad applicability. Nonetheless, in image enhancement, $(A2)$ and $(A3)$ might be considered too strong, in particular under the agreement that images be best represented as BV-functions. This point of view, initiated by Osher and Rudin [23, 24, 25], is widely accepted if the aim is edge detection or segmentation (cf. [13, 15, 6]), although it is clear that many images continue to be modeled as continuous or even smooth functions. This is particularly so in cases where the physical image generating process is taken into account (astronomy, medical imaging). We hold that our approach of modeling images in Sobolev spaces may offer a compromise.

3. Lagrangian formulation for $(P)_{\mathrm{pen}}$. In this section we present the first part of the scheme for program $(P)_{\mathrm{pen}}$. We provide a suitable Lagrangian formulation and a corresponding concave dual program (P^*). The second step of the relaxation scheme, dualizing the dual to obtain the bidual, will be presented in section 4.

For the following, let us fix some notations and definitions. Let Ω be a bounded open subset of $\mathbb{R}^N$, and suppose $a_k \in C(\overline{\Omega})$, $b_k \in C^1(\overline{\Omega})^n$ for $k = 1, \dots, N$. (It would be sufficient to require piecewise continuity of a_k and piecewise continuous differentiability of b_k.) Then the linear operators A and B defined by (1.3) are bounded on $L_1(\Omega)$ and $L_1(\Omega)^n$, respectively.

Let $h : \mathbb{R} \times \mathbb{R}^n \to \mathbb{R} \cup \{\infty\}$ be a proper convex lsc function with nonempty domain $\operatorname{dom} h$ (cf. [21]). Then

$$I_h[u] = \int_\Omega h(u(x), \nabla u(x))\, dx$$

is a proper convex lsc functional defined for all $u \in W^{1,1}(\Omega)$. Notice that we do not exclude the possibility $I_h[u] = +\infty$, as would, for instance, occur for a classical functional like $\int_\Omega |\nabla u|^2\, dx$, whose natural domain is $W^{1,2}(\Omega)$. A value $I_h[u] = +\infty$ simply means that u does not contribute to the minimization process. On the other hand, $I_h[u] = -\infty$ is impossible as a consequence of the lower semicontinuity of I_h (cf. [22] for this and other facts about convex integral functionals).

In cases when we wish to force positivity of the solution, we require $h(u,\xi) = +\infty$ whenever $u < 0$. Then $I_h[u] = +\infty$ unless $u \geq 0$ almost everywhere (a.e.) on Ω.

To avoid trivial situations, we will generally assume that $(P)_{\text{pen}}$ and $(P)_{\text{tol}}$ are feasible. More precisely, we assume existence of a function $u \in C^1(\overline{\Omega})$ with $I_h[u] < \infty$, $\int_\Omega u(x)\,dx = 1$, respectively, $|Au + B\nabla u - c| \leq \varepsilon$ in the case of the tolerance model. The value of program (P) will be denoted by $V(P)$, and we will require $V(P) > -\infty$ because otherwise no optimal solution exists. So altogether we adopt $-\infty < V(P) < \infty$ as our standing hypothesis. In the present section we consider $(P)_{\text{pen}}$. Analogous results for $(P)_{\text{tol}}$ will be presented in section 5.

We proceed to give a Lagrangian formulation of $(P)_{\text{pen}}$. By introducing dummy variables $v = \nabla u$ and $e = Au + Bv - c$ and by defining

$$J_h(u,v) := \int_\Omega h(u(x), v(x))\,dx$$

we rewrite $(P)_{\text{pen}}$ in the form

$$(P)_{\text{pen}} \qquad \begin{aligned} &\text{minimize} && J_h(u,v) + \frac{\alpha}{2}|e|^2 \\ &\text{subject to} && \nabla u = v,\ Au + Bv - c = e, \\ & && \int_\Omega u(x)\,dx = 1 \end{aligned}$$

with $e \in \mathbb{R}^N$ and $u \in C^1(\overline{\Omega}), v \in C(\overline{\Omega})^n$. This suggests using the Lagrangian

$$\begin{aligned} L(u,v,e;w,\lambda,\mu) = J_h(u,v) + \frac{\alpha}{2}|e|^2 + \langle w, \nabla u - v\rangle \\ + \lambda\cdot(Au + Bv - c - e) + \mu\bigg(\int_\Omega u(x)\,dx - 1\bigg), \end{aligned}$$

where $\langle .,. \rangle$ denotes the dual form either between $C(\bar{\Omega})^n$ and $M(\bar{\Omega})^n$ or between $L_1(\Omega)^n$ and $L_\infty(\Omega)^n$. We can now write $(P)_{\text{pen}}$ in the equivalent form

$$\inf_{u,v,e} \sup_{w,\lambda,\mu} L(u,v,e;w,\lambda,\mu). \tag{3.1}$$

As usual, the corresponding concave dual program is then defined by switching the inf and sup:

$$\sup_{w,\lambda,\mu} \inf_{u,v,e} L(u,v,e;w,\lambda,\mu). \tag{3.2}$$

We do not attempt to prove directly that (P) and (P^*) are equivalent or at least have equal values, since this will follow later as a consequence of the bidual relaxation scheme. Instead, we investigate (3.2) a little further by explicitly calculating the inner infimum.

To do this, we start by calculating the partial Legendre–Fenchel transform of L in its first three variables, defined as

$$L^*(y,z,d;w,\lambda,\mu) = \sup_{u,v,e} \big(\langle u,y\rangle + \langle v,z\rangle + e\cdot d - L(u,v,e;w,\lambda,\mu)\big),$$

and then recognize $-L^*(0,0,0;w,\lambda,\mu)$ as the objective of the dual (3.2), to be maximized over (w,λ,μ). While [21] is the basic reference for notions from finite dimensional convexity, a rigorous justification of (P^*) as obtained below would call for

methods as used in section 4 or in [7]. In particular, it would require calculating the conjugate L^* with respect to the space $C(\bar{\Omega})$ and its dual $M(\bar{\Omega})$, the space of signed Radon measures on $\bar{\Omega}$. Instead of calculating $(J_h)^*$, defined on a space of measures, we restrict the dual to the classical spaces $C(\bar{\Omega})$ and $C^1(\bar{\Omega})^n$, where it suffices to calculate J_{h^*} with the approach described as formal in section 2.

Written as a convex program, the dual is of the following form:

$$(P^*) \qquad \begin{aligned} \text{minimize} \quad & J_{h^*}(y,z) + \frac{1}{2\alpha}|\lambda|^2 + \lambda \cdot c + \mu \\ \text{subject to} \quad & y = \operatorname{div} z + \operatorname{div} B^T\lambda - A^T\lambda - \mu, \\ & y \in C(\overline{\Omega}), z \in C^1(\overline{\Omega})^n, \lambda \in \mathbb{R}^N, \mu \in \mathbb{R}. \end{aligned}$$

Here h^* is the Legendre–Fenchel conjugate of h and $J_{h^*}(y,z) := \int_\Omega h^*(y(x), z(x))\,dx$.

Example 1. For the class of Csiszár information measures (1.4) we have

$$h^*(y,z) = \begin{cases} 0, & y + \phi^*(z) \le 0, \\ \infty, & y + \phi^*(z) > 0. \end{cases}$$

As is easy to see, (P^*) has feasible points, so the value $V(P^*) < +\infty$. Also, the fact that the dual was obtained by flipping sup and inf gives $V(P^*) \ge -V(P) > -\infty$, so $V(P^*)$ is finite. The relation $V(P^*) \ge -V(P)$ is often referred to as weak duality (cf. [7]).

4. Existence of solutions for $(P)_{\text{pen}}$. The second part of our scheme now requires dualizing (P^*) again to obtain what we call a bidual relaxation (P^{**}) of the original program $(P)_{\text{pen}}$. As opposed to the formal way we employed to derive (P^*), we shall now have to rigorously dualize (P^*). As a consequence, the bidual (P^{**}) will be formulated in a dual Banach space, a space of measures. In a third step, also presented in the section, we will show that under reasonable conditions, the generalized solutions are functions in the Sobolev space $W^{1,1}(\Omega)$. A fourth step, to be presented in section 6, will examine under what circumstances a classical solution in $C^1(\overline{\Omega})$ may be obtained.

As before, duality requires an appropriate Lagrangian formulation, which we obtain by attaching a multiplier $u \in M(\bar{\Omega})$ to the equality constraint in (P^*). The dual Lagrangian is then

$$L_D(y,z,\lambda,\mu;u) = J_{h^*}(y,z) + \frac{1}{2\alpha}|\lambda|^2 + \lambda \cdot c + \mu + \langle u, \operatorname{div} z + \operatorname{div} B^T\lambda - A^T\lambda - \mu - y\rangle,$$

and an equivalent way of writing (P^*) is the minimax form:

$$(P^*) \qquad \inf_{\substack{y \in C(\overline{\Omega}) \\ z \in C^1(\overline{\Omega})^n \\ \lambda \in \mathbb{R}^N, \mu \in \mathbb{R}}} \sup_{u \in M(\overline{\Omega})} L_D(y,z,\lambda,\mu;u).$$

Switching inf and sup leads to the corresponding bidual program,

$$(P^{**}) \qquad \sup_u \inf_{y,z,\lambda,\mu} L_D(y,z,\lambda,\mu;u),$$

and immediately gives $V(P^{**}) \le V(P^*)$ (weak duality). Proving equality requires more work.

PROPOSITION 4.1. *We have $V(P^{**}) = V(P^*)$, and (P^{**}) admits an optimal solution $\overline{u} \in M(\overline{\Omega})$.*

Proof. (a) Let us first consider the case where h is not affine. The function $f : C(\overline{\Omega}) \to \mathbb{R} \cup \{+\infty\}$, defined by

$$f(\eta) = \inf \Big\{ J_{h^*}(y,z) + \frac{1}{2\alpha}|\lambda|^2 + \lambda \cdot c + \mu : y \in C(\overline{\Omega}), z \in C^1(\overline{\Omega})^n, \lambda \in \mathbb{R}^N, \mu \in \mathbb{R}, \\ \operatorname{div} z + \operatorname{div} B^T\lambda - A^T\lambda - \mu - y = \eta \Big\},$$

is proper convex lsc and we have $f(0) = V(P^*)$, which is finite. According to the general theory (cf. [7]) it remains to show that $\partial f(0)$, the subdifferential of f at 0, is nonempty, since every $\overline{u}$ with $-\overline{u} \in \partial f(0)$ is a solution to (P^{**}), showing in addition $V(P^*) = V(P^{**})$. Notice here that f, being defined on $C(\overline{\Omega})$, has subgradients in the dual space $M(\overline{\Omega})$. Proving $\partial f(0) \neq \emptyset$ requires two arguments. First we establish the existence of a supporting functional. Then we argue that the latter must be continuous since f is lsc.

(b) By the Hahn–Banach theorem, existence of a supporting functional will follow if we show that 0 is an algebraic interior point of $\operatorname{dom} f$. That means for every $\eta \in C(\overline{\Omega})$ we have to find $\rho > 0$ such that $\rho\eta \in \operatorname{dom} f$. Equivalently, we have to show that for every $\eta \in C(\overline{\Omega})$ we can find $\varrho > 0$ such that the equation

$$\operatorname{div} z + \operatorname{div} B^T\lambda - A^T\lambda - \mu - y = \varrho\eta$$

admits a solution (y, z, λ, μ) with $(y, z) \in \operatorname{dom} J_{h^*}$.

As h is not affine, $\operatorname{dom} h^*$ consists of at least two points. By convexity this means that either the projection $\Pi_y(\operatorname{dom}\ h^*)$ of $\operatorname{dom} h^*$ on the first coordinate contains a ball $|y - y_0| \leq \varepsilon$, or that $\Pi_z(\operatorname{dom} h^*)$ contains a segment.

(c) First consider the case where $\Pi_y(\operatorname{dom} h^*)$ has nonempty interior. By convexity there exists an affine function $y \mapsto z(y)$ such that $(y, z(y)) \in \operatorname{dom} h^*$ for all $|y - y_0| \leq \varepsilon$ and some fixed y_0. Let $z(y) = ay + b$, with $a, b \in \mathbb{R}^n$. Setting $y(x) = y_0 + \tilde{y}(x), \mu = -y_0$, and $\lambda = 0$, we have to solve the linear equation

$$a \cdot \nabla\tilde{y} - \tilde{y} = \varrho\eta$$

for $\|\tilde{y}\|_\infty \leq \epsilon$. Assuming without loss that $a_1 \neq 0$, a possible solution is the smooth function

$$\tilde{y}(x) = \varrho\, c(x)\, e^{x_1/a_1}, \quad \text{where} \quad c(x) = \frac{1}{a_1} \int_{\xi_1}^{x_1} \eta(\xi, x_2, \ldots, x_n)\, e^{\xi/a_1}\, d\xi$$

with a suitable $\xi_1 \in \mathbb{R}$. For ϱ sufficiently small we get in fact $\|\tilde{y}\|_\infty \leq \varepsilon$; hence $(\tilde{y}(x), z(\tilde{y}(x))) \in \operatorname{dom} h^*$ for every $x \in \Omega$ and hence $(\tilde{y}, z(\tilde{y})) \in \operatorname{dom} J_{h^*}$ by continuity of $\tilde{y}$. So in the first case the problem is solved.

(d) Now consider the case where $\Pi_y(\operatorname{dom}\ h^*) = \{y_0\}$. Since h is not affine, $\Pi_z(\operatorname{dom} h^*)$ contains at least two points. This means that (eventually with a change of coordinates) $\operatorname{dom} h^*$ contains a convex set of the form

$$\{y_0\} \times \{z_{01}\} \times \cdots \times \{z_{0r}\} \times B_{n-r},$$

with B_{n-r} an open ball with center $(z_{0,r+1}, \ldots, z_{0n})$ in a subspace of dimension $n - r \geq 1$. In the worst case $n - r = 1$, the first $n - 1$ coordinates are fixed, but z_n is free to

vary on an interval. Choosing $y \equiv y_0$, $\mu = -y_0$, and $z = z_0 + \tilde{z}$ with $\tilde{z} = (0, \dots, 0, \tilde{z}_n)$, defined by

$$\tilde{z}_n = \varrho \int_{\xi_n}^{x_n} \eta(x_1, \dots, x_{n-1}, \xi)\, d\xi,$$

we get a z having div $z = \varrho\eta$. Also $|z(x) - z_0| \le \varepsilon$ for all $x \in \Omega$ if ϱ is sufficiently small. Then $(y, z) \in \text{dom } J_{h^*}$ as required. So in both subcases, 0 is an algebraic interior point of dom f, and a supporting functional at 0 exists. Continuity of the latter follows from the lower semicontinuity of J_{h^*}. This completes the argument started in (a).

(e) Finally, consider the case where h is affine, and hence dom h^* consists of a single point (y_0, z_0). Define the function $f : C(\overline{\Omega}) \to \mathbb{R} \cup \{\infty\}$ as before. Since the value $V(P^*)$ is finite, $0 \in \text{dom } f$, and since the operators A, B have a finite dimensional range, dom f itself is contained in a finite dimensional linear subspace L of $C(\overline{\Omega})$. Linearity of A, B even gives dom$f = L$. Choose a supporting functional at $0 \in L$, and extend it to a continuous linear functional on all of $C(\overline{\Omega})$. □

Proposition 4.1 gives existence of a solution of (P^{**}) in $M(\overline{\Omega})$. We argue that under mild additional assumptions, $\bar{u}$ is in fact a function. We will even show a little more, namely, every u feasible for (P^{**}) satisfies $u \in L_\sigma(\Omega)$ for some $\sigma > 1$. Consider $u \in M(\overline{\Omega})$ with

$$\inf_{y,z,\lambda,\mu} L_D(y, z, \lambda, \mu; u) > -\infty,$$

where the infimum is over $y \in C(\bar{\Omega})$, $z \in C^1(\bar{\Omega})^n$, and $\lambda \in \mathbb{R}^N$, $\mu \in \mathbb{R}$ as before. Exploiting the form of L_D leads to three conditions:

$$\inf_{y,z} \left(J_{h^*}(y, z) + \langle u, \text{div } z - y\rangle\right) > -\infty, \tag{4.1}$$

$$\inf_{\lambda} \left(\frac{1}{2\alpha}|\lambda|^2 + \lambda \cdot c + \langle u, \text{div } B^T\lambda - A^T\lambda\rangle\right) > -\infty, \tag{4.2}$$

$$\inf_{\mu} \left(\mu - \langle u, \mu\rangle\right) > -\infty, \quad \text{i.e., } \int_\Omega du = 1. \tag{4.3}$$

As we shall see, the first condition allows for regularity considerations, while (4.2) and (4.3) will lead back to the original formulation of the constraints in (P).

First consider condition (4.1). We want to show that under suitable assumptions on h every feasible u possesses a Radon–Nikodym derivative lying in every space $L_\sigma(\Omega)$ with $1 < \sigma < \frac{n}{n-1}$. To do this we will need the following estimation for the Newton potential of a function $\varphi \in C_0^\infty(\Omega)$: Let φ be an element of $C_0^\infty(\Omega)$ and consider the corresponding Newton potential

$$v(x) = \int_\Omega \Gamma(x - s)\, \varphi(s)\, ds$$

with

$$\Gamma(x - s) = \begin{cases} \frac{1}{2\pi} \log|x - s|, & n = 2, \\ \frac{1}{n(2-n)\omega_n}|x - s|^{-(n-2)}, & n > 2, \end{cases}$$

where ω_n is the volume of the unit ball in $\mathbb{R}^n$. Then we have $v \in C^2(\overline{\Omega})$, $\Delta v = \varphi$, and

$$D_k v(x) = \int_\Omega D_k\Gamma(x - s)\, \varphi(s)\, ds \tag{4.4}$$

(cf. [9, Chapter 4]).

LEMMA 4.2. *Let* $1 < \sigma < \frac{n}{n-1}$. *Then*

$$|D_k v(x)| \le K \, \|\varphi\|_{\sigma'} \quad \textit{for all} \quad x \in \Omega \quad \textit{and for every} \quad \sigma' > n, \tag{4.5}$$

where the constant K *depends only on* σ' *and* Ω, *and* $1/\sigma + 1/\sigma' = 1$.

Proof. Using (4.4) and Hölder's inequality

$$|D_k v(x)| \le \|D_k \Gamma(x - .)\|_\sigma \, \|\varphi\|_{\sigma'}$$

provided $\|D_k \Gamma(x - .)\|_\sigma$ is finite. But for $n \ge 2$ we have

$$\int_\Omega |D_k \Gamma(x - s)|^\sigma \, ds = C_1 \int_\Omega \left(\frac{|x_k - s_k|}{|x - s|^n} \right)^\sigma \, ds \le C_2 \int_0^R r^{(n-1)(1-\sigma)} \, dr$$

with $\Omega \subset \{z \in \mathbb{R}^n : |z| \le R\}$ using n-dimensional spherical coordinates. As the last integral is finite for $(n-1)(1-\sigma) > -1$, or what is the same, $\sigma < \frac{n}{n-1}$, the lemma is proved. □

Now we want to use (4.5) and (4.1) to show that the map $\varphi \mapsto \langle u, \varphi \rangle$ is bounded on $C_0^\infty(\Omega)$ with respect to the $\|.\|_{\sigma'}$-norm, hence the Radon–Nikodym derivative of u is an element of $L_\sigma(\Omega)$, $(1/\sigma + 1/\sigma' = 1)$. To do this, we need to impose a richness condition on the domain of h^*:

$$\Pi_z(\text{dom } h^*) \text{ contains a segment.} \tag{A1}$$

As before, $\Pi_z : (y, z) \to z$ denotes the projection onto the last n coordinates.

Remark 1. Let us discuss the meaning of (A1). If $\Pi_z(\text{dom } h^*)$ does not contain a segment, dom $h^* \subset \mathbb{R} \times \{z\}$ for some $z \in \mathbb{R}^n$. This implies $h(x, y) = g(x) + y \cdot z$ for a convex function g, that is, h is linear in its second variable. We observe in a first place that z must be in the linear hull of the b_k. Therefore, in cases where we have no constraints on derivatives, $b = 0$ implies $z = 0$, leaving us with a problem without reference to derivatives. In case $b \ne 0$, the problem may be analyzed rather along classical lines as found in [7], although in general the result of Proposition 4.3 below is no longer valid. We consider objectives $h(x, \xi)$ linear in ξ as of minor importance for possible applications and do not pursue this class of objectives any further.

PROPOSITION 4.3. *Under the assumption* (A1) *every* $u \in M(\overline{\Omega})$ *feasible for* (P^{**}) *is absolutely continuous with respect to Lebesgue measure. Its Radon–Nikodym derivative lies in* $L_\sigma(\Omega)$ *whenever* $1 < \sigma < \frac{n}{n-1}$. *Furthermore, for every such* u *there exists a signed Borel vector measure* $m = m(u) \in M(\overline{\Omega})^n$ *satisfying*

$$\langle u, \operatorname{div} z \rangle = -\langle m(u), z \rangle \quad \textit{for all} \quad z \in C^1(\overline{\Omega})^n.$$

Remark 2. $m(u)$ is an extension of the distribution vector ∇u on $C(\overline{\Omega})^n$ and shall as well be denoted as ∇u. Notice however that this measure contains singular parts supported on $\partial\Omega$.

Proof. Step 1. Using a reduction argument similar to the one employed in the proof of Proposition 4.1, we may without loss assume that $\Pi_z(\text{dom} h^*)$ has nonempty interior in $\mathbb{R}^n$. The general case consists in repeating the same argument in the affine subspace generated by $\Pi_z(\text{dom} h^*)$, which by (A1) has dimension ≥ 1.

With these arrangements, assumption (A1) guarantees the existence of a ball $|z - z_0| \le \varepsilon$ and an affine function $y = y(z)$ such that $(y(z), z) \in$ dom h^* for all

$|z - z_0| \le \varepsilon$. Consider $\varphi \in C_0^\infty(\Omega)$ with $\|\varphi\|_{\sigma'} \le \frac{\varepsilon}{K}$ ($\frac{1}{\sigma} + \frac{1}{\sigma'} = 1$) for the constant K from (4.5) and let v be the corresponding Newton potential. Then using (4.5) we have $|D_k v(x)| \le \varepsilon$. Setting $z = z_0 + \nabla v$ and $y = y(z)$ we get from (4.1)

$$J_{h^*}(y(z), z) + \langle u, \varphi - y(z)\rangle > -\infty.$$

Now by construction we have $|J_{h^*}(y(z), z) + \langle u, -y(z)\rangle| \le K_1$ for some $K_1 > 0$, so we get

$$\inf_{\|\varphi\|_{\sigma'} \le \frac{\varepsilon}{K}} \langle u, \varphi\rangle > -\infty.$$

By linearity we conclude that the functional $\varphi \mapsto \langle u, \varphi\rangle$ is bounded on $(C_0^\infty(\Omega), \|.\|_{\sigma'})$ which is a dense subspace of $L_{\sigma'}(\Omega)$. For short $u \in L_\sigma(\Omega)$.

Step 2. For the second statement we have to show that for feasible u the functional $z \mapsto \langle u, \operatorname{div} z\rangle$ is bounded on $(C^1(\overline{\Omega}), \|.\|_\infty)$. This follows from (4.1) and the boundedness of $J_{h^*}(y(z), z)$ and $\langle u, -y(z)\rangle$ on the ball $\|z\|_\infty \le r$. □

For the following suppose condition (*A*1) is satisfied. In order to simplify our arguments, we continue to consider the case where $\Pi_z(\mathrm{dom} h^*)$ has nonempty interior in $\mathbb{R}^n$. Performing the same steps in the affine subspace L generated by $\mathrm{dom}(h^*)$ will settle the general case.

As a consequence of Propositions 4.1 and 4.3, and on exploiting the structure of L_D, (P^{**}) now reads

$$(P^{**}) \quad \inf_u \Big\{ \sup\big\{\langle u, y\rangle + \langle \nabla u, z\rangle - J_{h^*}(y, z) : y \in C(\overline{\Omega}), z \in C^1(\overline{\Omega})^n\big\} + \sup\Big\{-\frac{1}{2\alpha}|\lambda|^2 - \lambda \cdot c + \langle u, A^T\lambda - \operatorname{div} B^T\lambda\rangle : \lambda \in \mathbb{R}^N\Big\} : \int_\Omega u(x)\,dx = 1\Big\}.$$

To calculate the inner supremum over y and z we would like to use the following result of Rockafellar's [21] describing the conjugate of a convex integral functional J_{h^*} with respect to the dual pairing $(C(\overline{\Omega}) \times C(\overline{\Omega})^n, M(\overline{\Omega}) \times M(\overline{\Omega})^n)$.

LEMMA 4.4 (see Rockafellar [21]). *Let $\overline{\Omega}$ be a compact subset of $\mathbb{R}^n$ and suppose* $\mathrm{int}(\mathrm{dom}\ h^*) \ne \emptyset$. *Then for $\mu \in M(\overline{\Omega}) \times M(\overline{\Omega})^n$ with Lebesgue decomposition $\mu = \mu_a + \mu_s$ the conjugate of J_{h^*} equals*

$$J^*_{h^*}(\mu) = \int_\Omega h^{**}\left(\frac{d\mu_a}{dx}\right) dx + \int_\Omega \sup_{w \in \mathrm{dom}\, h^*} \left(w \cdot \frac{d\mu_s}{d\vartheta}\right) d\vartheta,$$

where μ_s is absolutely continuous with respect to the nonnegative Borel measure ϑ.

In order to apply Lemma 4.4 to (P^{**}), we first need to replace the supremum over $z \in C^1(\overline{\Omega})$ by a supremum $z \in C(\overline{\Omega})$. That this may be done without changing its value is guaranteed by the following lemma, whose proof will be given in the appendix.

LEMMA 4.5. *Let m be a measure in $M(\overline{\Omega})$, $f \in L_1(\Omega, m)$, $g \in L_1(\Omega, m)^k$ ($k \in \mathbb{N}$), and $\Phi : \mathbb{R}^k \to \mathbb{R} \cup \{\infty\}$ be a proper convex lsc function. Then for the proper convex lsc functional*

$$F(z) = \int_\Omega [\Phi(z(x))f(x) + z(x) \cdot g(x)]\, dm(x)$$

on $L_1(\Omega, m)^k$ we have

$$\inf_{z \in C^1(\overline{\Omega})^k} F(z) = \inf_{z \in L_1(\Omega, m)^k} F(z).$$

Furthermore, the values of the infima over all function spaces $\mathcal{F}$ with $C^1(\overline{\Omega})^k \subset \mathcal{F} \subset L_1(\Omega, m)^k$ agree.

Applying Lemma 4.5 to (P^{**}) with $k = n+1$, letting $z(x)$ stand for the pair $(y(x), z(x))$, $g(x) = \nabla u(x)$, and letting $\Phi(z(x))f(x)$ represent the term $h^*(y(x), z(x)) + u(x)y(x)$, we are now allowed to calculate

$$\sup_{\substack{y \in C(\overline{\Omega}) \\ z \in C(\overline{\Omega})^n}} \left(\langle u, y \rangle + \langle \nabla u, z \rangle - J_{h^*}(y, z) \right) \tag{4.6}$$

in (P^{**}), and Lemma 4.4 then shows that (4.6) equals

$$J_{h^{**}}(u, (\nabla u)_a) + \int_\Omega \sup_{z \in \Pi_z(\operatorname{dom} h^*)} \left(z \cdot \frac{d(\nabla u)_s}{d\vartheta}(x) \right) d\vartheta(x), \tag{4.7}$$

where $\nabla u = (\nabla u)_a + (\nabla u)_s$ denotes Lebesgue decomposition and $d(\nabla u)_s \ll d\vartheta$. A possible choice for ϑ is, for instance, the total variation of $d(\nabla u)_s$. For every feasible u we get in particular

$$\int_\Omega \sup_{z \in \Pi_z(\operatorname{dom} h^*)} \left(z \cdot \frac{d(\nabla u)_s}{d\vartheta}(x) \right) d\vartheta(x) = \int_\Omega \sigma_{\Pi_z(\operatorname{dom} h^*)} \left(\frac{d(\nabla u)_s}{d\vartheta}(x) \right) d\vartheta(x) < \infty. \tag{4.8}$$

Here $\sigma_{\Pi_z(\operatorname{dom} h^*)}(y)$ denotes the support function of the convex set $\Pi_z(\operatorname{dom} h^*)$ (cf. [21]).

Example 2. For the Csiszár information measures (1.4) we have $\Pi_z(\operatorname{dom} h^*) = \operatorname{dom} \phi^*$. From [21, Theorem 13.3] we deduce $\sigma_{\operatorname{dom} \phi^*} = \phi^{0+}$, the recession function of ϕ:

$$\phi^{0+}(y) = \lim_{\lambda \to \infty} \frac{1}{\lambda} (\phi(x + \lambda y) - \phi(x)) \quad \text{for an arbitrary} \quad x \in \operatorname{dom} \phi.$$

For the particular case $\phi(t) = |t|^p$, $p > 1$, we have

$$\phi^{0+}(y) = \begin{cases} \infty & \text{if } y \neq 0, \\ 0 & \text{if } y = 0, \end{cases}$$

while the case $p = 1$, $\phi(t) = |t|$ gives $\phi^{0+}(y) = |y|$. So for $p > 1$ the singular part of ∇u in (4.7) must vanish, since $\phi^{0+}((d(\nabla u)_s / d\vartheta)(x)) < \infty$ only for $(d(\nabla u)_s / d\vartheta)(x) = 0$ a.e. On the other hand, in case $p = 1$ we cannot argue that $(\nabla u)_s = 0$, so we only get $u \in BV(\Omega)$.

In general we need the assumption

$$\Pi_z(\operatorname{dom} h^*) \text{ is an affine subspace of dimension } \geq 1 \tag{A2}$$

to get $u \in W^{1,1}(\Omega)$. Notice that $(A2)$ readily implies $(A1)$. To understand the meaning of $(A2)$, consider the case where $\Pi_z(\operatorname{dom} h^*) = \mathbb{R}^n$. Then $h(x, \xi)$ is coercive in its ξ-variable. More precisely, $\Pi_z(\operatorname{dom} h^*) = \mathbb{R}^n$ implies that for every fixed x, $h(x, \xi)$ grows stronger than linearly as $|\xi| \to \infty$.

Proposition 4.6. *If assumption $(A2)$ is satisfied, every u which is feasible for (P^{**}) lies in $W^{1,1}(\Omega)$, and (P^{**}) has the form*

$$\inf_{u \in W^{1,1}(\Omega)} \left\{ I_h[u] + \frac{\alpha}{2} |Au + B\nabla u - c|^2 \, : \, \int_\Omega u(x)\, dx = 1 \right\}. \tag{P^{**}}$$

Remark 3. This means that (P^{**}) coincides with $(P)_{\rm pen}$, formulated in the Sobolev space $W^{1,1}(\Omega)$. Notice again here that this does not exclude situations where the natural space for the objective $I_h[u]$ is smaller, e.g., $\mathrm{dom} I_h \subset W^{1,2}(\Omega)$. In this case, $u \notin W^{1,2}(\Omega)$ will have $I_h[u] = \infty$ and the solution will automatically be an element of $W^{1,2}(\Omega)$.

Proof. As before, we present the argument in the case $\Pi_z(\mathrm{dom} h^*) = \mathbb{R}^n$, i.e., where the affine subspace L generated by dom h^* has dimension n. The general case is settled by repeating the argument in L.

Under these circumstances, condition $(A2)$ in tandem with $h = h^{**}$ allows reducing (4.7) to $J_h(u, \nabla u) = I_h[u]$. Indeed, with $z \in \Pi_z(\mathrm{dom}\, h^*)$ arbitrary, the supremum under the integral sign in (4.7) is $+\infty$, unless $(\nabla u)_s = 0$. Hence $(\nabla u)_a = \nabla u$, and the claim follows. Further, we can write (4.2) in the form

$$\sup_{\lambda\in\mathbb{R}^N}\left(-\frac{1}{2\alpha}|\lambda|^2 - \lambda\cdot c + \langle u, A^T\lambda - \mathrm{div}\, B^T\lambda\rangle\right)$$
$$= \sup_{\lambda\in\mathbb{R}^N}\left(\lambda\cdot(Au + B\nabla u - c) - \frac{\alpha}{2}|\lambda|^2\right) = \frac{1}{2\alpha}|Au + B\nabla u - c|^2,$$

so (P^{**}) is $(P)_{\rm pen}$ formulated in the space $W^{1,1}(\Omega)$. ∎

Propositions 4.3 and 4.6 now yield the main result for $(P)_{\rm pen}$.

THEOREM 4.7. *Under the hypothesis* $(A2)$*, the penalization model* $(P)_{\rm pen}$ *admits a solution* $\overline{u} \in W^{1,1}(\Omega)$.

5. Existence of solutions for $(P)_{\rm tol}$. Similar to $(P)_{\rm pen}$, the tolerance model $(P)_{\rm tol}$ can be written in the form

$$\inf_{u,v,e}\sup_{w,\lambda,\mu,\nu\geq 0}\tilde{L}(u, v, e; w, \lambda, \mu, \nu)$$

with

$$\tilde{L}(u, v, e; w, \lambda, \mu, \nu) = J_h(u, v) + \langle w, \nabla u - v\rangle + \lambda\cdot(Au + Bv - c - e)$$
$$+\mu\left(\int_\Omega u(x)\,dx - 1\right) + \nu(|e|^2 - \varepsilon^2).$$

Here we get the analogous results by similar reasoning so we will only cite the main theorem.

THEOREM 5.1. *If* $(A2)$ *is satisfied, the tolerance model* $(P)_{\rm tol}$ *admits a solution* $\overline{u} \in W^{1,1}(\Omega)$.

6. Regularity. In this section we show that the regularity of the solutions $\bar{u}$ of $(P)_{\rm pen}$ and $(P)_{\rm tol}$ may be improved to give $\bar{u} \in W^{1,p}(\Omega)$ for some $p > 1$ if condition $(A2)$ is strengthened. For $1 \leq \varrho \leq r$ consider the condition

$(A3)$ there exists a measurable function $y \mapsto y(z)$, $\mathbb{R}^n \to \mathbb{R}$, such that
(i) $|y(z)| \leq K(1 + |z|^\varrho)$ for every $z \in \mathbb{R}^n$, and
(ii) $J_{h^*}(y(z), z)$ is bounded on a ball $\{z \in C^1(\overline{\Omega}) : \|z\|_r \leq C\}$.

Clearly $(A3)$ implies $(A2)$ and may be understood as a coercivity condition on the integrand h. Notice that $(A3)$ is, for instance, satisfied if h^* satisfies the growth condition

(ii′) $h^*(y, z) \leq K(1 + |y|^{r/\varrho} + |z|^r)$,

which translates into a coercivity condition for h. In this case, $y(z) = |z|^\rho$ satisfies (i).

COROLLARY 6.1. *Suppose (A3) (with $1 \le \varrho \le r$) is satisfied. Then every u feasible for $(P)_{\rm pen}$ (resp., $(P)_{\rm tol}$) lies in $W^{1,p(r,\varrho)}(\Omega)$ with*

$$p(r,\varrho)\begin{cases} = \infty, & \varrho = r = 1,\\ = \frac{r}{r-1}, & \varrho = 1 < r,\\ = \frac{r}{r-1}, & \varrho > 1 \text{ and } r > n(\varrho-1),\\ < \frac{n(\varrho-1)}{n(\varrho-1)-1}, & \varrho > 1 \quad \text{and } r \le n(\varrho-1).\end{cases}$$

In particular, this is true for the solution $\overline{u}$ of $(P)_{\rm pen}$ or $(P)_{\rm tol}$.

For the proof we need the following.

LEMMA 6.2. *Given $u \in W^{1,1}(\Omega)$ with $\nabla u \in L_p(\Omega)^n$ for some $p > 1$ we have $u \in W^{1,p}(\Omega)$.*

Proof. We have to show $u \in L_p(\Omega)$. Consider the sequence

$$u_n(x) = \begin{cases} u(x), & |u(x)| \le n,\\ n, & u(x) > n,\\ -n, & u(x) < -n.\end{cases}$$

Then [28, Corollary 2.1.8] gives

$$\nabla u_n(x) = \begin{cases} \nabla u(x), & |u(x)| < n,\\ 0, & |u(x)| \ge n,\end{cases}$$

hence $u_n \in W^{1,p}(\Omega)$ for all n. We want to show that $\|u_n\|_p$ is bounded so the Fatou lemma will give the result. Following [1, Theorem 4.20] for each $\varepsilon > 0$ there exists a set $\Omega_\varepsilon \subset\subset \Omega$ such that for every $v \in W^{1,p}(\Omega)$,

$$\|v\|_p \le K\varepsilon\|\nabla v\|_p + K\|v\|_{p,\Omega_\varepsilon}$$

with $K = K(p,\Omega)$. (Here $\|v\|^p_{p,\Omega_\varepsilon} = \int_{\Omega_\varepsilon} |v(x)|^p\,dx$.) Now since $u \in L_p^{\rm loc}(\Omega)$ (cf. [11, Theorem 4.5.13]) we have $\|u\|_{p,\Omega_\varepsilon} < \infty$, and from the definition of u_n we get

$$\|u_n\|_p \le K\varepsilon\|\nabla u_n\|_p + K\|u_n\|_{p,\Omega_\varepsilon} \le K\varepsilon\|\nabla u\|_p + K\|u\|_{p,\Omega_\varepsilon} = C < \infty$$

for every n. Now, using $u_n(x) \to u(x)$ a.e., Fatou's lemma provides

$$\|u\|_p \le \liminf_{n\to\infty} \|u_n\|_p \le C,$$

hence $u \in L_p(\Omega)$. ∎

We proceed to complete the proof of the corollary.

Proof. We give the argument for $(P)_{\rm pen}$, the tolerance case being similar. Suppose $u \in W^{1,1}(\Omega)$ is feasible for $(P)_{\rm pen}$. By Proposition 4.6, it is then feasible for (P^{**}) as well, so we have

$$\inf_{z\in C^1(\overline{\Omega})^n} (J_{h^*}(y(z),z) + \langle u, y(z)\rangle + \langle \nabla u, z\rangle) > -\infty. \tag{6.1}$$

We want to construct a decreasing (possibly breaking off) sequence of exponents $r_k \ge r$ such that the term $J_{h^*}(y(z),z) + \langle u, y(z)\rangle$ is bounded on a ball $\|z\|_{r_k} \le C$, giving $\nabla u \in L_{p_k}(\Omega)$ with $p_k = \frac{r_k}{r_k-1}$ by (6.1). Lemma 6.2 will imply $u \in W^{1,p_k}(\Omega)$.

The procedure is the following. Suppose we have already constructed $r_k > r, p_k = r_k/(r_k - 1) > 1$, with $u \in W^{1,p_k}(\Omega)$ by the argument above. Then by the Sobolev embedding theorem (cf. [1, p. 97]) we have $u \in L_{s_k}(\Omega)$ with

$$s_k \begin{cases} \text{arbitrary if } \ p_k \geq n, \\ = \frac{np_k}{n-p_k} \ \text{ if } \ p_k < n. \end{cases}$$

Using Hölder's inequality and condition (i) of $(A3)$ we have

$$|\langle u, y(z)\rangle| \leq \|u\|_{s_k} \|y(z)\|_{s'_k} \leq \tilde{K} \|u\|_{s_k} \|z\|_{\varrho s'_k}$$

with $\frac{1}{s_k} + \frac{1}{s'_k} = 1$. On the other hand, condition (ii) of $(A3)$ implies that $J_{h^*}(y(z), z)$ is bounded on $\|z\|_r \leq C$. Hence we choose $r_{k+1} = \max(r, \varrho s'_k)$. By construction the sequence (r_k) is strictly decreasing unless $r_k = r$, the lowest exponent we can possibly achieve. As soon as $r_k = r$ for some index k, the process stops giving $u \in W^{1,\frac{r}{r-1}}(\Omega)$. (Notice that if $p_k \geq n$ for some k, we can always choose s_k large enough to guarantee $\varrho s'_k \leq r$, viz. $r_{k+1} = r$.)

Now we want to compute the r_k explicitly: From Theorem 4.7 we know $u \in W^{1,1}(\Omega)$, i.e., $p_0 = 1$, giving $s_1 = \frac{n}{n-1}$ and $r_1 = \max(r, \varrho n)$. So we can actually stop after the first step if $r \geq \varrho n$. Otherwise we get $p_1 = \frac{\varrho n}{\varrho n - 1}$,

$$\begin{cases} s_1 = \frac{n\varrho}{n\varrho - \varrho - 1} \ \text{ and } \ r_2 = \max(r, \frac{n\varrho^2}{1+\varrho}) & \text{for } \ \frac{n\varrho}{n\varrho - 1} < n, \\ r_2 = r & \text{if } \ \frac{n\varrho}{n\varrho - 1} \geq n. \end{cases}$$

Proceeding like this we get

$$r_k = \begin{cases} \frac{n}{k} \to 0 & \text{if } \ \varrho = 1, \\ n\frac{\varrho - 1}{1 - \varrho^{-(k+1)}} \searrow n(\varrho - 1) & \text{if } \ \varrho > 1, \end{cases}$$

so the process will break off, giving $u \in W^{1,\frac{r}{r-1}}(\Omega)$, unless $\varrho > 1$ and $r \leq n(\varrho - 1)$. In the latter case we still get $u \in W^{1,p}(\Omega)$ for every $p < \frac{n(\varrho-1)}{n(\varrho-1)-1}$. □

Example 3. For the Csiszár information measures with $\phi(t) = |t|^p$ discussed in the preparatory section 1, we may choose $y(z) = K|z|^{p'}$ $(\frac{1}{p} + \frac{1}{p'} = 1)$, so $r := \varrho := p'$ will do, and the corollary gives $u \in W^{1,\tilde{p}}(\Omega)$ with

$$\tilde{p} \begin{cases} = p, & p > n, \\ < \frac{n}{n-p+1} = 1 + \frac{p-1}{n-p+1}, & p \leq n. \end{cases}$$

In particular, for p close to 1 we have $\tilde{p} = 1 + \epsilon(p)$ and $u \in W^{1,1+\varepsilon(p)}(\Omega)$ with $\varepsilon(p) := \frac{p-1}{n-p+1} \downarrow 0$ for $p \downarrow 1$.

Appendix. We still have to prove Lemma 4.5.

LEMMA 4.5. *Let m be a measure in $M(\overline{\Omega})$, $f \in L_1(\Omega, m)$, $g \in L_1(\Omega, m)^k$ $(k \in \mathbb{N})$, and $\Phi : \mathbb{R}^k \to \mathbb{R} \cup \{\infty\}$ be a proper convex lsc function. Then for the proper convex lsc functional*

$$F(z) = \int_\Omega [\Phi(z(x))f(x) + z(x) \cdot g(x)]\, dm(x)$$

on $L_1(\Omega, m)^k$ we have

$$\inf_{z \in C^1(\overline{\Omega})^k} F(z) = \inf_{z \in L_1(\Omega,m)^k} F(z).$$

Furthermore, the values of the infima over all function spaces $\mathcal{F}$ with $C^1(\overline{\Omega})^k \subset \mathcal{F} \subset L_1(\Omega, m)^k$ agree.

Proof. We give the argument in the case where dom Φ has nonempty interior in $\mathbb{R}^k$, the general case being reducible to the former.

It is sufficient to show that for any $z \in L_1(\Omega, \mu)^m$ with $F(z) < \infty$ and any $\varepsilon_0 > 0$ we can find a $y \in C^1(\overline{\Omega})^m$ such that $F(y) \leq F(z) + \varepsilon_0$. The construction of such a y will be divided into three steps. First we will prove the existence of a function $\tilde{z} \in L_1(\Omega, \mu)^m$ with values only in int(dom Φ) having $F(\tilde{z}) \leq F(z) + \frac{\varepsilon_0}{3}$. Then we will modify $\tilde{z}$ to get a function $\tilde{z}_{n_0}$ which maps Ω into a compact subset of int(dom Φ), again having $F(\tilde{z}_{n_0}) \leq F(\tilde{z}) + \frac{\varepsilon_0}{3}$. For the last step we will use the Lipschitz continuity of Φ on compact subsets of int(dom Φ) to find a suitable measure ν such that the approximation of $\tilde{z}_{n_0}$ with respect to ν by $C^1(\overline{\Omega})^m$-functions will also approach $F(\tilde{z}_{n_0})$. This will finally prove the existence of a $y \in C^1(\overline{\Omega})^m$ with $F(y) \leq F(\tilde{z}_{n_0}) + \frac{\varepsilon_0}{3}$, giving $F(y) \leq F(z) + \varepsilon_0$ altogether.

Step 1. Consider $z \in L_1(\Omega, \mu)^m$ with $F(z) < \infty$. Then we have $z(x) \in \operatorname{dom} \Phi$ a.e. for every representative of z. In particular, we can choose a measurable representative also denoted by z having $z(x) \in \operatorname{dom} \Phi$ for all $x \in \Omega$.

For fixed $\varepsilon > 0, \delta > 0$ we define the set-valued mapping

$$\Gamma(x) = \{\zeta \in \text{int}(\text{dom } \Phi) : |\zeta - z(x)| \leq \varepsilon,\ \Phi(\zeta) \leq \Phi(z(x)) + \delta\} \text{ for all } x \in \Omega.$$

We want to show that Γ admits a measurable selector using the Kuratovski and Ryll-Nardczevski measurable selection theorem [14]: Suppose Γ is a measurable set-valued mapping with nonempty closed images. Then there exists a measurable $\tilde{z} : \Omega \to \mathbb{R}^m$ having $\tilde{z}(x) \in \Gamma(x)$ for every $x \in \Omega$.

As will be seen, for sufficiently small ε, δ, this selector satisfies $F(\tilde{z}) \leq F(z) + \frac{\varepsilon_0}{3}$.

We have to verify three properties of Γ:

$\Gamma(x)$ *is closed for every* $x \in \Omega$: Fix $x \in \Omega$ and suppose (ζ_n) is a sequence in $\Gamma(x)$ converging to some $\zeta \in \text{int}(\text{dom } \Phi)$. Then we have $|\zeta - z(x)| = \lim_{n\to\infty} |\zeta_n - z(x)| \leq \varepsilon$ and $\Phi(\zeta) \leq \liminf_{n\to\infty} \Phi(\zeta_n) \leq \Phi(z(x)) + \delta$, so $\zeta \in \Gamma(x)$ and $\Gamma(x)$ is closed in int(dom Φ).

$\Gamma(x)$ *is nonempty for all* $x \in \Omega$: Since Φ is proper convex and lsc, epi Φ is a closed convex subset of $\mathbb{R}^m \times \mathbb{R}$ with

$$\operatorname{epi} \Phi = \overline{\operatorname{int}(\operatorname{epi} \Phi)}$$

(cf. [21, p. 46]). Hence any point $(z(x), \Phi(z(x))) \in \operatorname{epi} \Phi$ can be approximated by a sequence $(\zeta_n, \Phi(\zeta_n) + \delta_n) \in \text{int}(\text{epi } \Phi)$; that means $\zeta_n \in \text{int}(\text{dom } \Phi)$ and $\delta_n > 0$. But then we must have $\zeta_n \in \Gamma(x)$ for n sufficiently large, so $\Gamma(x)$ is nonempty.

Γ *is measurable:* We have to show that for each measurable $M \subset \mathbb{R}^m$ the preimage $\Gamma^{-1}(M)$ is a measurable subset of Ω. Here, without loss of generality, we can assume $M \subset \text{int}(\text{dom } \Phi)$. We get

$$\begin{aligned}
\Gamma^{-1}(M) &= \big\{x \in \Omega : \exists\, \zeta \in M :\ z(x) \in B(\zeta, \varepsilon), \Phi(z(x)) \geq \Phi(\zeta) - \delta\big\} \\
&= \big\{x \in \Omega : \exists\, y \in B(0, \varepsilon),\ \exists\, \zeta \in M :\ z(x) = y + \zeta,\ \Phi(z(x)) \geq \Phi(\zeta) - \delta\big\} \\
&= \big\{x \in \Omega : \exists\, y \in B(0, \varepsilon) \colon (z(x), \Phi(z(x))) \in \operatorname{epi}(\Phi - \delta) \cap (M \times \mathbb{R}) + \{(y, 0)\}\big\} \\
&= \big\{x \in \Omega :\ (z(x), \Phi(z(x))) \in \operatorname{epi}(\Phi - \delta) \cap (M \times \mathbb{R}) + B(0, \varepsilon) \times \{0\}\big\} \\
&= (z, \Phi(z))^{-1} \left(\operatorname{epi}(\Phi - \delta) \cap (M \times \mathbb{R}) + B(0, \varepsilon) \times \{0\}\right).
\end{aligned}$$

Since epi $(\Phi - \delta)$ is closed and therefore measurable, $\Gamma^{-1}(M)$ is now the preimage of a measurable set under the measurable map $x \mapsto (z(x), \Phi(z(x)))$; hence $\Gamma^{-1}(M)$ is measurable.

Now we can apply the selection theorem of Kuratovski and Ryll-Nardczevski to get a measurable function $\tilde{z}$ with $\tilde{z}(x) \in \Gamma(x)$ for all $x \in \Omega$. By the definition of Γ we have $\tilde{z} \in L_1(\Omega, \mu)^m$ and

$$F(\tilde{z}) \leq F(z) + \delta \int_\Omega |f(x)|\, d\mu(x) + \varepsilon \int_\Omega |g(x)|\, d\mu(x),$$

so for sufficiently small δ and ε, $F(\tilde{z}) \leq F(z) + \frac{\varepsilon_0}{3}$, and the first step is proven.

Step 2. Without loss of generality, assume $0 \in \text{int}(\text{dom } \Phi)$ and $\Phi(0) = 0$. Now choose an increasing sequence (D_n) of compact subsets of $\mathbb{R}^m$ having $0 \in D_1$ and

$$D_n \uparrow \text{int}(\text{dom } \Phi),\ D_n \subseteq \{\zeta \in \mathbb{R}^m : |\Phi(\zeta)| \leq n\},\ D_n \subset \text{int}\,(D_{n+1}).$$

Defining $\Omega_n = \tilde{z}^{-1}(D_n) = \{x \in \Omega : \tilde{z}(x) \in D_n\}$ and letting $\tilde{z}_n = \chi_{\Omega_n} \cdot \tilde{z}$, we have $\tilde{z}_n(x) \to \tilde{z}(x)$ and $\Phi(\tilde{z}_n(x)) \to \Phi(\tilde{z}(x))$ pointwise, since $\Omega_n \uparrow \Omega$. Now

$$|F(\tilde{z}_n)| \leq \int_\Omega |\Phi(\tilde{z}(x))f(x) + \tilde{z}(x) \cdot g(x)|\, d\mu(x) \quad \text{for all} \quad n \in \mathbb{N}$$

and dominated convergence implies $F(\tilde{z}_n) \to F(\tilde{z})$, so we can choose some $n_0 \in \mathbb{N}$ with $F(\tilde{z}_{n_0}) \leq F(\tilde{z}) + \frac{\varepsilon_0}{3}$.

Step 3. Now we have $\tilde{z}_{n_0}(x) \in D_{n_0}$ for all $x \in \Omega$, so if we want to approach $\tilde{z}_{n_0}$ by smooth functions we can restrict ourselves to functions with values in D_{n_0+1} since D_{n_0} is a compact subset of $\text{int}\,(D_{n_0+1})$ having a positive distance from its boundary. But for each $y \in C^1(\overline{\Omega})^m$ with values in D_{n_0+1} we get

$$|F(\tilde{z}_{n_0}) - F(y)| \leq \int_\Omega |\tilde{z}_{n_0}(x) - y(x)|\ (L_{n_0+1}|f(x)| + |g(x)|)\ d\mu(x),$$

where L_{n_0+1} denotes the Lipschitz constant of Φ on D_{n_0+1}. So we have to approximate $\tilde{z}_{n_0}$ with respect to the measure $d\nu = (L_{n_0+1}|f| + |g|)\, d\mu$. Choosing $y \in C^1(\overline{\Omega})^m$ with $\|\tilde{z}_{n_0} - y\|_{L_1(\Omega,\nu)^m} \leq \frac{\varepsilon_0}{3}$ (notice $\tilde{z}_{n_0} \in L_1(\Omega, \nu)^m$, hence such a y exists), we finally get

$$F(y) \leq F(\tilde{z}_{n_0}) + \frac{\varepsilon_0}{3} \leq F(\tilde{z}) + 2\frac{\varepsilon_0}{3} \leq F(z) + \varepsilon_0,$$

and the proof is complete. ∎

REFERENCES

[1] R.A. Adams, *Sobolev Spaces*, Academic Press, New York, NY, 1978.

[2] G. Aubert and L. Vese, *A variational method in image recovery*, SIAM J. Numer. Anal., 34 (1997), pp. 1948–1979.

[3] J.M. Borwein and A.S. Lewis, *Partially-finite programming in L_1 and the existence of maximum entropy estimates*, SIAM J. Optim., 3 (1993), pp. 248–267.

[4] J.M. Borwein, A.S. Lewis, and D. Noll, *Maximum entropy reconstruction using derivative information* I: *Fisher information and convex duality*, Math. Oper. Res., 21 (1996), pp. 442–468.

[5] T. Chan and L. Vese, *Variational Image Restoration and Segmentation Models and Approximations*, Technical Report CAM 97-47, University of California, Los Angeles, 1997.

[6] D. Dobson and F. Santosa, *Recovery of blocky images from noisy and blurred data*, SIAM J. Appl. Math., 56 (1996), pp. 1181–1198.

[7] I. Ekeland and R. Temam, *Convex Analysis and Variational Problems*, North-Holland Publishing Company, Amsterdam, 1976.
[8] S. Geman and D. Geman, *Stochastic relaxation, Gibbs distribution and the Bayesian restoration of images*, IEEE Trans. Pattern Anal. Machine Intell., 6 (1984), pp. 721–741.
[9] D. Gilbarg and N.S. Trudinger, *Elliptic Partial Differential Equations of Second Order*, Springer-Verlag, New York, 1977.
[10] U. Hermann, *Rekonstruktionsprobleme bei unvollständiger Information: Funktionale vom erweiterten Entropietyp*, Doctoral thesis, Mathematisches Institut A, Universität Stuttgart, Stuttgart, 1997, Shaker-Verlag, Aachen.
[11] L. Hörmander, *The Analysis of Linear Partial Differential Operators* I, Springer-Verlag, New York, 1990.
[12] K. Ito and K. Kunisch, *An active set strategy based on the augmented Lagrangian formulation for image restoration*, RAIRO Modél. Math. Anal. Numér, to appear.
[13] K.E. Casas, K. Kunisch, and C. Pola, *Regularization by Functions of Bounded Variation and Applications to Image Enhancement*, Technical Report 1996.5, Dpto. Mat. Est. y Comp., Universidad de Cantabria, Spain, 1996.
[14] K. Kuratovski and C. Ryll-Nardczevski, *A general theorem on selectors*, Bull. Acad. Pol. Sci. Sér. Sci. Math. Astron. Phys., 13 (1965), pp. 397–403.
[15] Y. Li and F. Santosa, *An affine scaling algorithm for minimizing total variation in image enhancement*, Technical Report CTC94TR201, Cornell Theory Center, Cornell University, Ithaca, NY, 1994.
[16] A.K. Louis, *Inverse und schlecht gestellte Probleme*, Teubner-Verlag, Leipzig, 1989.
[17] J.-P. Morel and S. Solimini, *Variational Methods in Image Segmentation*, Progr. Nonlinear Differential Equations Appl. 14, Birkhäuser-Verlag, Basel, Switzerland, 1995.
[18] D. Noll, *Restoration of degraded images with maximum entropy*, J. Global Optim., 10 (1997), pp. 91–103.
[19] D. Noll, *Reconstruction with noisy data—an approach via eigenvalue optimization*, SIAM J. Optim., 8 (1998), pp. 82–104.
[20] D. Noll, *Variational methods in image restoration*, in Recent Advances in Optimization, Lecture Notes in Econom. and Math. Systems 452, P. Gritzmann, R. Horst, E. Sachs, and R. Tichatschke, eds., Springer-Verlag, Berlin, 1997, pp. 229–245.
[21] R.T. Rockafellar, *Convex Analysis*, Princeton University Press, Princeton, NJ, 1970.
[22] R.T. Rockafellar, *Integrals which are convex functionals*, I and II, Pacific J. Math., 24 (1968), pp. 525–539, and 39 (1971), pp. 439–469.
[23] L. Rudin and S. Osher, *Total variation based image restoration with free local constraints*, Proc. IEEE ICIP, I (1994), pp. 31–35.
[24] S. Osher and L. Rudin, *Feature-oriented image enhancement using shock filters*, SIAM J. Numer. Anal., 27 (1990), pp. 919–940.
[25] L. Rudin, S. Osher, and E. Fatemi, *Nonlinear total variation based noise removal algorithms*, Phys. D, 60 (1992), pp. 259–268.
[26] L. Vese, *A study in the BV space of a denoising-deblurring variational problem*, to appear.
[27] L. Vese and T. Chan, *Reduced non-convex functional approximations for image restoration and segmentation*, Technical Report CAM 97-56, University of California, Los Angeles, 1997.
[28] W.P. Ziemer, *Weakly Differentiable Functions*, Springer-Verlag, New York, 1989.

SIAM J. CONTROL OPTIM.
Vol. 38, No. 4, pp. 1241–1264

PARAMETERIZED LMIs IN CONTROL THEORY*

PIERRE APKARIAN† AND HOANG DUONG TUAN‡

Abstract. A wide variety of problems in control system theory fall within the class of parameterized linear matrix inequalities (LMIs), that is, LMIs whose coefficients are functions of a parameter confined to a compact set. Such problems, though convex, involve an infinite set of LMI constraints and hence are inherently difficult to solve numerically.

This paper investigates relaxations of parameterized LMI problems into standard LMI problems using techniques relying on directional convexity concepts. An in-depth discussion of the impact of the proposed techniques in quadratic programming, Lyapunov-based stability and performance analysis, μ analysis, and linear parameter-varying control is provided. Illustrative examples are given to demonstrate the usefulness and practicality of the approach.

Key words. linear matrix inequalities, robust semidefinite programming, directional convexity, robustness analysis, parametric uncertainty, linear parameter-varying control

AMS subject classifications. 93D05, 93D09, 93D10, 93D25, 90C34

PII. S036301299732612X

1. Introduction.

LMI techniques are now well-rooted as a unifying framework for formulating and solving problems in control theory with a remarkable degree of simplicity. The main thrust of these techniques is that certain complicated control problems can be solved very efficiently. Specifically, the interior-point methods for semidefinite programming have worst-case polynomial complexity with respect to the problem size. From a practical viewpoint, extensive experience shows that interior-point methods solve problems in roughly less than a hundred iterations, independent of the problem size. Each elementary iteration reduces to solving a least-square problem which incurs the main computational overhead. Recent and thorough studies of interior-point techniques for semidefinite programming are, among others, Jarre [23], Vandenberghe and Boyd [43], Rendl, Vanderbei, and Wolkowicz [34], and the master book by Nesterov and Nemirovski [28].

Basically, the simple feasibility problem of semidefinite programming consists in seeking a solution to the LMI

$$F_0 + z_1 F_1 + \cdots + z_r F_r < 0\,, \tag{1.1}$$

where the F_i's are given real symmetric matrices and the z_i's are the sought decision variables. A significantly more complicated generalization of problem (1.1) is the feasibility problem

$$F_0(\theta) + z_1(\theta) F_1(\theta) + \cdots + z_r(\theta) F_r(\theta) < 0, \tag{1.2}$$

where $\theta := [\theta_1, \ldots, \theta_N]^T$ is an additional parameter allowed to take any value in a compact set H of $\mathbf{R}^N$, typically a polytope. In contrast to problem (1.1) the problem

*Received by the editors August 18, 1997; accepted for publication (in revised form) October 29, 1998; published electronically April 20, 2000.

http://www.siam.org/journals/sicon/38-4/32612.html

†ONERA-CERT, Control System Dept., 2 av. Edouard Belin, 31055 Toulouse, France (apkarian@cert.fr).

‡Department of Control and Information, Toyota Technological Institute, Hisekata 2-12-1, Tempaku, Nagoya 468-8511, Japan (tuan@toyota.ti.ac.jp).

data, $F_i(\theta)$, are now symmetric matrix-valued functions of θ, and we are seeking (arbitrary) functions of θ, $z_i(\theta)$ such that the LMI constraints (1.2) hold for any admissible value of θ. The complexity of problem (1.2) is twofold:

1. It is infinite-dimensional since the $z_i(.)$'s are sought in the infinite-dimensional space of functions of θ.

2. This is an infinitely constrained LMI problem for which each constraint corresponds to a given point in the range of θ.

A common and practical approach to overcome the difficulties arising from dimensionality is to select a finite basis of functions for the z_i's and reconsider the problem over the resulting spanned finite-dimensional space. In such a case, problem (1.2) simplifies to an LMI problem of the form

$$F_0(\theta) + z_1 F_1(\theta) + \cdots + z_r F_r(\theta) < 0, \qquad \forall \theta \in H, \tag{1.3}$$

where $z_1, \ldots, z_r$ are conventional scalar decision variables as in (1.1) and H is a compact set. Such problems are referred to as *robust semidefinite programming* problems in [4] and are designated here as *parameterized LMI* (PLMI) problems to stress the connections with the LMI control theory literature.

For reasons raised above, PLMI feasibility problems still have high complexity and are even known to be NP-hard [4]. The aim of this paper is to develop systematic relaxation techniques to turn, potentially conservatively, this problem into a standard LMI problem. A fruitful technique for turning PLMI problems into conventional LMI problems is the well-known **S**-procedure [45, 14]. With this approach, scaling or multipliers are utilized to eliminate the LMI parameter-dependence. The price to pay is the insuperable conservatism of the resulting conditions and also the extra computational effort, often prohibiting, introduced by the multiplier variables. This paper exploits competitive techniques invoking directional convexity concepts to derive a finite set of LMI conditions. Generally speaking, the approach requires significantly less variables than **S**-procedure techniques whilst producing more LMI constraints. Since the flop cost of interior-point techniques is roughly linear with respect to the size of the LMI constraint but polynomial with respect to the number of decision variables, the proposed techniques offer a valuable alternative to **S**-procedure techniques. It is, however, difficult to draw definitive conclusions at this stage since the respective performance of each technique is probably problem-dependent. As demonstrated in the body of the paper, the techniques therein also offer possibilities for handling polytopic representations, that is when the parameter θ designates polytopic coordinates, $\sum_{i=1}^{N} \theta_i = 1$, $\theta_i \geq 0$. We also briefly discuss relaxations of linear objective minimization problems subject to PLMI constraints and PLMI problems subject to algebraic constraints on the parameter θ.

The scope of applications of PLMIs is quite large and goes far beyond the area of robust control theory. In [4, 5], Ben-Tal and Nemirovski lay the foundations of *robust convex programming* and investigate its theoretical *tractability* in conjunction with the analysis of some generic uncertain convex programs. The same stream of ideas are applied to a truss topology design problem in [6]. In [29], the authors provide a thorough study of the regularity properties of solutions to PLMIs using the **S**-procedure and discuss its implications for a variety of topics: linear programming, polynomial interpolation, integer programming, and so forth. Our contribution is in line with that of [29] or what is called "approximate robust counterpart" of an uncertain semidefinite programming problem in [4]. The general instance of the problems is essentially intractable and we are constructing relaxed forms, generally conservative,

that are directly amenable to the use of interior-point methods. Note also that alternative techniques to those considered here are developed in [40] using either convex approximations or difference convex (d.c.) representations.

This work is mostly control theory oriented, and special attention is paid to the following topics:

(i) *Quadratic programming.* It is shown that some neither convex nor concave quadratic programming problems can be converted into boolean programming problems. The results so introduced constitute the core of the subsequent derivations and have a direct impact for relaxing PLMI problems.

(ii) *Lyapunov-based stability and performance analysis.* A rich catalog of Lyapunov-based stability and performance criteria for uncertain systems can be handled via PLMIs, thus providing generalizations of the single quadratic Lyapunov function approach.

(iii) *μ-analysis.* PLMIs have direct applications in the μ-analysis context or robust non-singularity analysis and can be utilized to refine the computation of upper-bounds.

(iv) *Linear parameter-varying (LPV) control synthesis.* PLMIs and the concepts developed here are also central in LPV control synthesis to overcome the difficulties arising from gridding phases and reduce the computational efforts.

The paper is structured as follows. Section 3 discusses a variety of directional convexity concepts and their implications in functional optimization. These results are then extended to PLMI problems in section 4. Important robust and LPV control issues mentioned above are investigated in section 5. Numerical examples illustrating the techniques and tools are given in section 6.

2. Preliminaries. The following definitions and notations are used throughout the paper.

$\mathbf{R}$ and $\mathbf{C}$ denote the sets of real and complex numbers, respectively. M^T is the transpose of the matrix M, and M^* denotes its complex-conjugate transpose. The notation $\operatorname{Tr} M$ stands for the trace of M. For Hermitian or symmetric matrices, $M > N$ means that $M - N$ is positive definite and $M \geq N$ means that $M - N$ is positive semidefinite.

Let S be a convex subset of $\mathbf{R}^n$. A function $f : S \to \mathbf{R}$ is quasi-convex if and only if for all u, v in S and α in [0, 1],

$$f(\alpha u + (1-\alpha)v) \leq \max\{f(u), f(v)\} .$$

Strict quasi-convexity is obtained when the inequality is strict for all $0 < \alpha < 1$. This notion is weaker than convexity, which requires

$$f(\alpha u + (1-\alpha)v) \leq \alpha f(u) + (1-\alpha) f(v) .$$

The relative interior, the closure, and the relative boundary of S are denoted as $\operatorname{ri} S$, $\operatorname{cl} S$, and $\operatorname{rbd} S$, respectively. We then have $\operatorname{rbd} S = \operatorname{cl} S \setminus \operatorname{ri} S$.

A polytope Π in $\mathbf{R}^n$ is defined as the compact set

$$\Pi := \left\{ \sum_{i=1}^{L} \alpha_i v_i : \sum_{i=1}^{L} \alpha_i = 1,\ \alpha_i \geq 0,\ v_i \in \mathbf{R}^n \right\} .$$

Equivalently, it is also the convex hull of the set $V = \{v_1, \ldots, v_L\}$, denoted $\operatorname{co} V$. The notation $\operatorname{vert} \Pi$ designates the set of vertices of Π, $\operatorname{vert} \Pi := V$. The affine hull, $\operatorname{aff} S$,

of a set S is defined as the set of all affine combinations of elements of S, i.e.,

$$\operatorname{aff} S := \left\{ \sum_{i=1}^{k} \alpha_i s_i : \ s_i \in S, \quad \sum_{i=1}^{k} \alpha_i = 1 \right\} .$$

The direction space associated to aff S is defined as aff $S - s_0$, where s_0 is any point of aff S. The notation $\#S$ stands for the number of elements in a set S.

3. Extreme point results. This section introduces some useful tools that permit us to convert the maximization of a function over a polytope Π into the combinatorial problem of maximizing f over $\operatorname{vert} \Pi$. We begin with a general result which is the core of the subsequent derivations.

THEOREM 3.1 (Central result). *Consider a polytope* Π *and assume that for any* x *in* Π*, there exists a direction* d *in the direction space of* aff Π *such that* f *is quasi-convex on the line segment*

$$L_d(x) := \{ z \in \Pi : \ z = x + \lambda d, \ \lambda \in \mathbf{R} \} .$$

Then, f *has a maximum over* Π *in* $\operatorname{rbd} \Pi$.

Proof. Assume f has a maximum $\hat{x}$ in $\operatorname{ri} \Pi$. Consider a line segment $L_d(\hat{x})$ where f is quasi-convex. From this property, we infer that f has a maximum point in $\operatorname{rbd} \Pi \cap L_d(\hat{x})$, and therefore

$$f(\hat{x}) \leq f(\bar{x}),$$

for some $\bar{x}$ in $\operatorname{rbd} \Pi \cap L_d(\hat{x})$. ∎

By virtue of Theorem 3.1, the search of a maximum point is reduced to exploring the relative boundary of Π. This result is analogous to the well-known *maximum principle* for analytic functions of complex variables. Although this constitutes an appealing result which might find applications, it is still hardly tractable for our particular purpose. A stronger result is obtained by forcing the directions d to be parallel to the edges of the polytope. The corollary below clarifies this fact.

COROLLARY 3.2 (Multi-quasi-convexity). *Consider a polytope* Π *and the directions* $d_1, \ldots, d_q$ *determined by the edges of* Π*. Assume that for any* x *in* Π*, the function* f *is quasi-convex on the line segments* $L_{d_i}(x)$ *for* $i = 1, \ldots, q$*. Then,* f *has a maximum over* Π *at a vertex of* Π.

Proof. Immediate by application of Theorem 3.1 to Π and to the (polytopic) faces and edges of Π. ∎

An obvious consequence of Theorem 3.1 is the following.

COROLLARY 3.3. *Under the hypotheses of Corollary* 3.2*, the following conditions are equivalent:*

(i) $f(x) < 0 \quad \forall x \in \Pi$.
(ii) $f(x) < 0 \quad \forall x \in \operatorname{vert} \Pi$.

As claimed previously, the maximization problem in Corollary 3.2 and the sign verification problem in Corollary 3.3 are turned into simpler combinatorial problems of lower complexity. This is a consequence of the multi-quasi-convexity property defined in Corollary 3.2. Note that the term *multi-quasi-convex* emphasizes the fact that f is separately quasi-convex along parallels to the edges of the polytope. This property is attached to the function f but is also intimately related to the particular geometry of the polytope.

Quasi-convexity is a less stringent requirement than usual convexity, the counterpart being the difficulty of its verification even for differentiable functions. Alternative

conditions that are more easily amenable to numerical computation are derived by replacing quasi-convexity with convexity in Theorem 3.1 and Corollaries 3.2 and 3.3. For twice continuously differentiable functions, Corollary 3.2 then becomes as follows.

COROLLARY 3.4 (Multiconvexity). *With the definitions in Corollary* 3.2, *f has a maximum over Π in* $\operatorname{vert}\Pi$ *whenever it holds that*

$$\frac{\partial^2 f(x+\lambda d_i)}{\partial \lambda^2} \geq 0 \quad \forall x \in \Pi, \quad i = 1, \ldots, q\,. \tag{3.1}$$

Affine functions are trivially multi-quasi-convex functions, so any of the above results is applicable. It is instructive to consider the case in which f is a quadratic function and Π is a hyperrectangle.

COROLLARY 3.5 (Quadratic functions). *Consider a quadratic function, $f(x) = x^T Q x + c^T x + a$, and assume Π is a hyperrectangle with edges paralleling the axes of coordinates, that is, $x = [x_1, \ldots, x_n]^T$ with*

$$\alpha_i \leq x_i \leq \beta_i, \quad i = 1, \ldots, n\,.$$

Assume further that

$$Q_{ii} \geq 0, \quad i = 1, \ldots, n. \tag{3.2}$$

Then, f has a maximum over Π in $\operatorname{vert}\Pi$.

Proof. From Corollary 3.4, the conditions (3.2) express multiconvexity of the quadratic function. □

Clearly, the conditions (3.2) are less demanding than (global) convexity which requires $Q \geq 0$. When such conditions hold, the maximization of f over the polytope Π reduces to a boolean programming problem [35], which is much simpler (though possibly costly) than the maximization of a general f. One possible advantage is that some costly but practically useful concave minimization techniques, such as simplicial and conical partitioning (branch and bound) techniques like those of Tuy and Thach, might be used to find a global optimal solution. The reader is referred to the book of Tuy [41] for a thorough treatment.

4. Relaxation of PLMIs. This section presents some applications of these results to PLMIs whose coefficients are dependent on a parameter evolving in a polytopic set. To emphasize the fact that these parameters might be interpreted as uncertainties or scheduled variables of robust control or LPV control problems, the free variable x is denoted θ or α, hereafter.

Before proceeding further, it is instructive to have in mind the following important facts from [4]. Consider the "robust counterpart" (parameterized convex program in our terminology) of a general uncertain convex program:

$$\begin{array}{c} \text{minimize } c^T z\,, \\ \text{subject to } F(z,\theta) \in K, \forall \theta \in H, \end{array} \tag{4.1}$$

where K is a closed convex cone, H is a generalized ellipsoidal set including as instances standard ellipsoids but also ellipsoidal cylinders and polyhedras, $F(z,\theta)$ is K-concave with respect to z. A key additional assumption is that $F(z,\theta)$ must be K-concave with respect to θ. With these assumptions in place, Ben-Tal and Nemirovski established the following:

(i) The robust counterpart of an uncertain linear program is a conic quadratic program; thus it is perfectly tractable.

(ii) The robust counterpart of an uncertain quadratically constrained convex quadratic program is a semidefinite program, hence tractable, but is NP-hard for intersections of ellipsoidal uncertainty sets.

(iii) The robust counterpart of an uncertain semidefinite program is generally NP-hard even for a single ellipsoidal uncertainty set.

The problems examined in what follows fall within the latter class, so they are generally NP-hard. They also generally fail to satisfy the K-concavity in θ, mentioned above. By virtue of its inherent complexity, one must, as a last resort, use relaxation techniques to end up with tractable "approximate" programs. Again with reference to [4], we take advantage of some directional K-concavity instead of complete K-concavity in the uncertain parameter to derive such relaxations.

4.1. PLMIs with quadratic parameter dependence. We consider PLMIs in the class

$$\mathcal{L}(z,\alpha) := M_0(z) + \sum_{i=1}^{L} \alpha_i M_i(z) + \sum_{i,j=1}^{L} \alpha_i \alpha_j M_{ij}(z) < 0, \tag{4.2}$$

where z stands for the decision variable and $M_0(.)$, $M_i(.)$, and $M_{ij}(.)$ are real symmetric matrix-valued and linear functions of z. In addition, it is supposed that the parameter $\alpha = [\alpha_1, \ldots, \alpha_L]^T$ evolves in the simplex

$$\Gamma := \left\{ \alpha : \sum_{i=1}^{L} \alpha_i = 1,\ \alpha_i \geq 0 \right\}. \tag{4.3}$$

Note that the problem presented in (4.2) involves infinitely many LMIs associated with each value of the parameter α and is known to be intractable [4]. By enforcing some constraints of geometric nature on the functional dependence in α, it is however possible to reduce, potentially conservatively, the problem to solving a finite number of LMIs. This is established in the next proposition.

PROPOSITION 4.1. *The infinite set of LMIs* (4.2) *is feasible for some z whenever the finite set of LMIs*

$$M_0(z) + M_k(z) + M_{kk}(z) < 0, \tag{4.4}$$

$$M_{ii}(z) + M_{jj}(z) - (M_{ij}(z) + M_{ji}(z)) \geq 0, \tag{4.5}$$

where $1 \leq k \leq L$ and $1 \leq i < j \leq L$, is feasible for some z.

Proof. Note first that the conditions (4.2) are equivalent to $x^T \mathcal{L}(z,\alpha) x < 0\ \forall\ x \neq 0$. For fixed $x \neq 0$, consider $x^T \mathcal{L}(z,\alpha) x$ as function of α. By virtue of Corollary 3.4, it is negative whenever it is multiconvex along lines paralleling the edges of Γ and furthermore is negative over $\operatorname{vert} \Gamma$. The remainder of the proof follows from the fact that $\operatorname{vert} \Gamma$ is composed of the canonical basis of $\mathbf{R}^L$, and the directions of the edges of Γ are determined by vectors with all but two zero coordinates, the nonzero coordinates having opposite signs:

$$d_1 := [1, -1, 0, \ldots, 0],$$
$$d_2 := [1, 0, -1, 0, \ldots, 0],\ \ldots.$$

Repeating the reasoning for all $x \neq 0$ yields the condition (4.4) and (4.5), as desired. ∎

Remarks. By strengthening the conditions in (4.4), one can slightly relax the multiconvexity requirement in (4.5). As an example, the solutions (z, Z_i) to the LMI feasibility problem

$$M_0(z) + \sum_{i=1}^{L} \alpha_i M_i(z) + \sum_{i,j=1}^{L} \alpha_i \alpha_j M_{ij}(z) < -\sum_{i=1}^{L} \alpha_i^2 Z_i \qquad \forall \alpha \in \Gamma,$$
$$Z_i \geq 0, \quad i = 1, \ldots, L,$$

give solutions z to the feasibility problem (4.2). Arguing as in proposition 4.1, associated sufficient solvability conditions are easily obtained as

$$M_0(z) + M_k(z) + M_{kk}(z) < -Z_k, \tag{4.6}$$
$$M_{ii}(z) + M_{jj}(z) - (M_{ij}(z) + M_{ji}(z)) \geq -(Z_i + Z_j), \tag{4.7}$$
$$Z_k \geq 0\,, \tag{4.8}$$

where $1 \leq k \leq L$ and $1 \leq i < j \leq L$. Due to the strict nature of (4.6), the nonstrict inequalities in (4.7) and (4.8) can be changed into strict inequalities without any loss of generality. In the strict form, such problems are readily solved using interior-point semidefinite programming techniques such as those in [8, 42, 27]. Note also that the Z_i's can be chosen as general symmetric matrices whose size is that of the LMI condition (4.2). Less costly characterizations are obtained by using diagonal or scalar matrices instead, such as

$$Z_i = \operatorname{diag} \lambda_i \quad \text{or} \quad Z_i = \lambda_i I\,.$$

When Γ is a hyperrectangle and the LMIs (4.2) are expressed in terms of the Cartesian coordinates of α (as opposed to polytopic ones), the main result in [18] is recovered as a special case. Assume $\theta := [\theta_1, \ldots, \theta_N]^T$ ranges over a hyperrectangle, denoted H, that is,

$$\underline{\theta}_i \leq \theta_i \leq \bar{\theta}_i\,. \tag{4.9}$$

Then

$$\mathcal{L}(z, \theta) := M_0(z) + \sum_{i=1}^{N} \theta_i M_i(z) + \sum_{i,j=1}^{N} \theta_i \theta_j M_{ij}(z) < 0 \quad \forall \theta \in H$$

whenever

$$\begin{array}{lcl} \mathcal{L}(z, \theta) < 0, & \theta & \in \quad \text{vert } H, \\ M_{ii}(z) \geq 0, & i & = \quad 1, \ldots, N. \end{array}$$

As before, one can relax the multiconvexity requirement above by replacing these conditions with

$$\mathcal{L}(z, \theta) < -\sum_{i=1}^{N} \theta_i^2 \lambda_i I, \quad \theta \in \text{vert } H \tag{4.10}$$
$$M_{ii}(z) \geq -\lambda_i I, \quad i = 1, \ldots, N. \tag{4.11}$$

More generally, any nonpositive matrix-valued function of θ is a good candidate for the right-hand side of (4.10). More complicated polynomial functions lead naturally to more costly characterizations. From our practical experience, a reasonable compromise between computational efficiency and tightness of the test is obtained with nonhomogeneous functions of the form $-(\lambda_0 + \sum_{i=1}^{N} \theta_i^2 \lambda_i)\, I$.

4.2. Gridding techniques. The techniques developed in sections 3 and 4 provide sufficient and computationally simple conditions for checking the sign of a function or the feasibility of a PLMI problem. These conditions may introduce conservatism though it turns out to be small from our practical experience. A different technique which is guaranteed to provide a nonconservative answer but is potentially optimistic and generally computationally intensive, is to use a fine gridding of the parameter range and solve a finite set of LMIs corresponding to each point on the grid. Denoting the grid as G, the PLMI problem (4.2) is then replaced with the finite set of LMIs

$$M_0(z) + \sum_{i=1}^{L} \alpha_i M_i(z) + \sum_{i,j=1}^{L} \alpha_i \alpha_j M_{ij}(z) < 0, \quad \alpha \in G\,.$$

Such a technique is currently used in stability analysis and LPV control. It is, however, limited to problems of reasonable size, say less than three parameters. There is also the risk of missing a critical value of the parameter, hence leading to overly optimistic answers. With the approaches presented earlier, these difficulties are inherently ruled out. These techniques can be mixed with gridding approaches, hence offering alternative possibilities. Indeed, instead of gridding the entire parameter range, there is only a need to grid a surface of lower dimension whenever the function is quasiconvex or convex along some direction. This is an immediate consequence of Theorem 3.1. A simple illustration of this fact is given below. For the sake of simplicity, we restrict the discussion to 2 parameters θ_1 and θ_2 evolving in the normalized square

$$|\theta_1| \le 1, \qquad |\theta_2| \le 1\,, \tag{4.12}$$

and we consider the PLMI problem

$$M_0(z) + \sum_{i=1}^{2} \theta_i M_i(z) + \sum_{i,j=1}^{2} \theta_i \theta_j M_{ij}(z) < 0\,. \tag{4.13}$$

A potential technique for checking the feasibility of this problem consists first of enforcing convexity in the direction of θ_1. This is equivalent to the LMI constraint

$$M_{11}(z) \ge 0\,. \tag{4.14}$$

Thanks to this condition, it is then enough to grid the line segments

$$\theta_1 = \pm 1 \quad \text{and} \quad |\theta_2| \le 1$$

to check the feasibility of (4.13).

Finally, let us note that the approaches presented in the previous subsections are also very useful for developing a global optimization algorithm solving PLMIs. Indeed, the main difficulty in global optimization is "the curse of dimensionality," i.e., the size of the space where the global search is performed. Thus exploiting convexity properties such as directional convexity is very important for developing an efficient global optimization algorithm (see, e.g., [24, 41]) since it allows us to drastically simplify the problem by limiting the global search to a restricted region of the feasible domain. For instance, with condition (4.14), it is sufficient to perform a global search for (4.13) just on the line segment $|\theta_2| \le 1$ instead of on the square (4.12) in R^2.

4.3. PLMIs with polynomial parameter dependence. In this section, we are considering polynomially θ-dependent PLMIs of the form

$$\mathcal{L}(\theta, z) := \sum_{\nu \in J} \theta^{[\nu]} M_\nu(z) < 0\,, \tag{4.15}$$

where the terms $M_\nu(z)$ denote symmetric matrix-valued functions of the decision variable z that are linear in z. The notation $[\nu]$ is the vector of partial degrees $[\nu] = [\nu_1, \ldots, \nu_N]$ associated with the lexicographically ordered term

$$\theta^{[\nu]} = \theta_1^{\nu_1} \theta_2^{\nu_2} \ldots \theta_N^{\nu_N}\,,$$

with the convention $\theta^{[0]} = 1$. It is assumed that θ ranges over a hyperrectangle H as in (4.9). J is a set of N-tuples of partial degrees describing the polynomial expansion (4.15). Again exploiting Corollary 3.4, it is possible to reduce (conservatively) this problem to a finitely constrained LMI problem. The symbols d_k and d designate the partial and total degrees in the matrix polynomial expansion.

LEMMA 4.2. *Consider the PLMI* (4.15)*, where θ ranges over a hyperrectangle. Then the LMI conditions*

$$\mathcal{L}(\theta, z) < 0 \quad \forall \theta \in H \tag{4.16}$$

hold for some z, whenever the finite family of LMI conditions

$$\mathcal{L}(\theta, z) < 0 \quad \forall \theta \in \operatorname{vert} H, \tag{4.17}$$

$$(-1)^m \frac{\partial^{2m}}{\partial \theta_{l_1}^2 \ldots \partial \theta_{l_m}^2} \mathcal{L}(\theta, z) \le 0, \quad \forall \theta \in \operatorname{vert} H\,, \tag{4.18}$$

where

$$1 \le l_1 \le l_2 \le \cdots \le l_m \le N, \quad 1 \le m \le \frac{d}{2},$$

$$2\#\{l_j = k :\ j \in \{1, \ldots, m\}\} \le d_k, \quad k = 1, 2, \ldots, N,$$

are feasible for some z.

Proof. The proof is obtained by a repeated use of Corollary 3.4. ∎

As an example, consider the PLMI feasibility problem

$$\mathcal{L}(z, \theta) := M_0(z) + \theta_1^2 \theta_2 M_{112}(z) + \theta_2^3 M_{222}(z) \ < \ 0, \qquad |\theta_i| \le 1\,.$$

Replacing this problem with, for instance,

$$\widetilde{\mathcal{L}}(z, \theta) := M_0(z) + \theta_1^2 \theta_2 M_{112}(z) + \theta_2^3 M_{222}(z) + \lambda_0 I + \lambda_1 \theta_1^2 I + \lambda_2 \theta_2^2 I \ < \ 0, \qquad |\theta_i| \le 1,$$

and using Lemma 4.2 yields the LMI conditions:

$$\widetilde{\mathcal{L}}(z, \theta) < 0 \quad \forall \theta \in \operatorname{vert} H$$

$$-\frac{\lambda_1}{2} I \le M_{112}(z) \le \frac{\lambda_1}{2} I, \qquad -\frac{\lambda_2}{3} I \le M_{222}(z) \le \frac{\lambda_2}{3} I, \quad \lambda_1 \ge 0, \quad \lambda_2 \ge 0\,.$$

It is of interest to note that when PLMIs also involve full-matrix parameters Δ_j, it is more appropriate for computational reasons to take advantage of a combined use of directional convexity concepts and of the **S**-procedure to formulate a feasibility test.

4.4. Algebraically constrained PLMI problems. PLMI problems with algebraic constraints are described as

$$\mathcal{L}(z,\theta) < 0 \quad \forall \theta \in H \tag{4.19}$$

subject to

$$g_1(\theta) = 0, \ldots, g_q(\theta) = 0\,, \tag{4.20}$$

where $g_1, \ldots, g_q$ are polynomials in θ. Note that for consistency, the algebraic surface (4.20) should have a nonvoid intersection with the hypercube H.

It is readily verified that solutions to the unconstrained PLMI problem

$$\mathcal{L}(z,\theta) < \sum_{i=1}^{q} g_i(\theta)^2 \lambda_i I \quad \forall \theta \in H, \tag{4.21}$$

$$\lambda_i \geq 0, \quad i = 1, \ldots, q, \tag{4.22}$$

also solve (4.19)–(4.20). Recast as the sufficient conditions (4.21)–(4.22), the hard problem (4.19)–(4.20) can be handled with the technical machinery developed in section 4 and is therefore amenable to a conventional LMI problem. Once again, there is some practically useful flexibility for selecting the right-hand side of the first inequality in (4.21).

4.5. Linear objective minimization under PLMI constraints. The directional convexity concepts introduced previously are applicable with minor changes to linear objective minimization problems subject to PLMI constraints. This means problems of the form

$$\begin{array}{c} \text{minimize } c^T z \\ \text{subject to } \mathcal{L}(z,\theta) < 0, \quad \theta \in H\,, \end{array} \tag{4.23}$$

where c is a given vector and the inequalities constitute a PLMI constraint. It is also possible to handle min-max problems of the form

$$\begin{array}{c} \text{minimize } \max\limits_{\theta \in H} c(\theta)^T z \\ \text{subject to } \mathcal{L}(z,\theta) < 0, \quad \theta \in H\,, \end{array} \tag{4.24}$$

using standard manipulations [4, 29]. Defining

$$\widetilde{z} = \begin{bmatrix} z \\ \lambda \end{bmatrix}, \ \widetilde{c} = \begin{bmatrix} 0 \\ 1 \end{bmatrix}, \ \widetilde{\mathcal{L}}(\widetilde{z},\theta) = \begin{bmatrix} \mathcal{L}(z,\theta) & 0 \\ 0 & c(\theta)^T z - \lambda \end{bmatrix},$$

problem (4.24) is equivalently formulated as

$$\begin{array}{c} \text{minimize } \widetilde{c}^T \widetilde{z} \\ \text{subject to } \widetilde{\mathcal{L}}(\widetilde{z},\theta) < 0, \quad \theta \in H\,, \end{array} \tag{4.25}$$

which has a form similar to problem (4.23).

In this form, provided that the parameter dependence is polynomial, such problems are easily converted into standard LMI problems using directional convexity concepts. This is left to the reader. Finally, we note that since these concepts amount to shrinking the z-feasible set, the optimal value of the relaxed LMI optimization problem is an upper bound for problems (4.23) and (4.24).

5. Applications in control theory. The techniques and tools presented in sections 3 and 4 enjoy a wide scope of applications. They are useful for the analysis of both the stability and the performance of uncertain systems. Potentially, all Lyapunov-based stability and performance measures can be handled with the proposed techniques which are more general and less conservative than single quadratic function approaches [7, 3]. They also have implications in the context of μ analysis where some upper bounds can be refined into less conservative upper bounds. Another important domain of application concerns LPV control techniques. For brevity, we only report a few of these applications.

5.1. Robust stability. We consider the linear uncertain system

$$\dot{x} = A(\alpha)x, \qquad A(\alpha) := \alpha_1 A_1 + \cdots + \alpha_L A_L\,, \tag{5.1}$$

where α is a fixed uncertain parameter evolving in the simplex (4.3). It follows that the uncertain matrix $A(\alpha)$ ranges over a matrix polytope

$$A(\alpha) \in \text{co}\,\{A_1, \ldots, A_L\}\,.$$

We are seeking a quadratic parameter-dependent Lyapunov function with similar structure,

$$V(x,\alpha) := x^T\,(\,\alpha_1 X_1 + \cdots + \alpha_L X_L\,)\,x,$$

establishing stability of the uncertain system for all admissible dynamics. If we make explicit the Lyapunov conditions for stability

$$V(x,\alpha) > 0, \quad \frac{d}{dt}V(x,\alpha) < 0 \qquad \forall x \neq 0\,,$$

we obtain

$$\begin{aligned} \alpha_1 X_1 + \cdots + \alpha_L X_L &> 0, \\ A(\alpha)^T(\alpha_1 X_1 + \cdots + \alpha_L X_L) + (\alpha_1 X_1 + \cdots + \alpha_L X_L)A(\alpha) &< 0\,, \end{aligned}$$

which constitutes a PLMI problem. Thus, Proposition 4.1 can be used to convert the problem into a finite number of LMI feasibility conditions. The following sufficient test for robust stability is derived.

PROPOSITION 5.1. *Assume one of the A_i's is stable. Then, the uncertain system* (5.1) *is stable whenever there exist symmetric matrices $X_1, \ldots, X_L$ and scalars $\lambda_1, \ldots, \lambda_L$ such that the following LMI conditions hold for $k = 1, \ldots, L$ and $1 \le i < j \le L$:*

$$\begin{aligned} A_k^T X_k + X_k A_k &< -\lambda_k I, \\ A_i^T X_i + X_i A_i + A_j^T X_j + X_j A_j - (A_i^T X_j + X_j A_i + A_j^T X_i + X_i A_j) &\ge -(\lambda_i + \lambda_j) I, \\ \lambda_k &\ge 0\,. \end{aligned}$$

In such case, the Lyapunov function $V(x,\alpha)$ establishes stability of the uncertain system (5.1).

Proof. The above conditions ensure that $\frac{d}{dt}V(x,\alpha) < 0$ for all admissible values of the parameter. Moreover, $V(x,\alpha)$ is a candidate Lyapunov function since at least one of the A_i's is stable, and since $\alpha_1 X_1 + \cdots + \alpha_L X_L$ cannot be singular, we infer that $\alpha_1 X_1 + \cdots + \alpha_L X_L > 0$ for all α in the simplex (4.3). ∎

5.2. Robust performance. As claimed earlier, the proposed techniques are potentially applicable to any Lyapunov-based performance measure. We illustrate this claim with the H_2 performance criterion. See [9] for a Lyapunov characterization of the H_2 norm.

Consider the uncertain system

$$\begin{array}{rcl} \dot{x} & = & A(\alpha)x + B(\alpha)w, \\ z & = & C(\alpha)x, \end{array} \tag{5.2}$$

where

$$\begin{bmatrix} A(\alpha) & B(\alpha) \\ C(\alpha) & 0 \end{bmatrix} \in \text{co} \left\{ \begin{bmatrix} A_1 & B_1 \\ C_1 & 0 \end{bmatrix}, \ldots, \begin{bmatrix} A_L & B_L \\ C_L & 0 \end{bmatrix} \right\}.$$

By virtue of Proposition 4.1, and paralleling the argument in Proposition 5.1, we deduce that the H_2 norm from w to z of the uncertain system (5.2) is bounded by ν for all values of α in the simplex (4.3) whenever there exist symmetric matrices $X_1, \ldots, X_L$, Q and scalars $\lambda_1, \ldots, \lambda_L$ such that the following LMI conditions hold:

$$\begin{array}{rcl} U_k(X_k) & < & -\lambda_k I, \\ V_i(X_i) + V_j(X_j) - (V_i(X_j) + V_j(X_i)) & \geq & -(\lambda_i + \lambda_j) I, \\ \lambda_k \geq 0, \quad \begin{bmatrix} P_k & C_k^T \\ C_k & Q \end{bmatrix} > 0, \quad \text{Tr}\, Q & < & \nu \end{array}$$

for $k = 1, \ldots, L$ and $1 \leq i < j \leq L$, with the definitions

$$U_i(X_j) := \begin{bmatrix} A_i^T X_j + X_j A_i & X_j B_i \\ B_i^T X_j & -I \end{bmatrix}, \quad V_i(X_j) := \begin{bmatrix} A_i^T X_j + X_j A_i & X_j B_i \\ B_i^T X_j & 0 \end{bmatrix}.$$

Extensions and reformulations for time-varying uncertain parameters, pole clustering in LMI regions, H_∞ and passivity constraints, and many others are straightforward. When the dependence on the parameter is polynomial, the same line of attack is still valid with the assistance of Lemma 4.2.

5.3. μ analysis. The structured singular value (SSV) or μ is an important linear algebra tool to study a class of matrix perturbation problems [11, 12, 36]. Since many robust stability/performance problems can be recast as one of computing μ with respect to an appropriate block-diagonal structure, it is also particularly useful in control theory and practice. The computation of μ involves an optimization problem which is not convex and is known to be NP-complete [32], so that it is difficult to compute μ exactly. Fortunately, it is possible to compute lower and upper bounds for μ with reasonable computational effort [13, 46]. This is the approach considered in this section.

The computation of μ can be formulated as computing the smallest norm perturbation for which the matrix $I - \Delta M$ becomes singular, where M denotes the plant's transfer function at some given frequency and Δ stands for uncertainties which are generally assumed to have a specific block-diagonal structure. In this section, we assume without loss of generality that uncertainties are real, $\Delta_{ij} \in \mathbf{R}$, and range over a polytope

$$\Delta \in \text{co}\, \{\Delta_1, \ldots, \Delta_L\}.$$

Extensions to mixed real/complex uncertainties are readily derived.

Our goal is to determine sufficient conditions for which $I - \Delta M$ remains nonsingular for all admissible uncertainties. Our approach is inspired by the work in [15, 26] and goes as follows. A necessary and sufficient condition for the nonsingularity of $(I - \Delta M)$ is the existence of a parameter-dependent matrix $F(\Delta)$, such that

$$F(\Delta)(I - \Delta M) + (I - \Delta M)^* F(\Delta)^* < 0\,. \tag{5.3}$$

The awkward condition (5.3) is simplified by restricting the search of $F(\Delta)$ matrices to those having the form

$$F(\Delta) := \sum_{i=1}^{L} \alpha_i F_i\,,$$

where the α_i's are the coordinates of Δ in the convex decomposition

$$\Delta := \sum_{i=1}^{L} \alpha_i \Delta_i\,.$$

With these restrictions, it is not difficult to see that inequality (5.3) takes a form similar to (4.2), that is, a PLMI feasibility problem. Therefore, by a direct application of Proposition 4.1 or its refined version (4.6)–(4.8), sufficient conditions for the nonsingularity of $I - \Delta M$ express as the existence of suitably dimensioned complex matrices F_1, $F_2, \ldots$, F_L such that

$$\begin{array}{rcl} F_k + F_k^* - (F_k \Delta_k M + (F_k \Delta_k M)^*) & < & -\lambda_k I, \\ (F_i \Delta_i M + F_j \Delta_j M) - (F_i \Delta_j M + F_j \Delta_i M) + (\star)^* & \leq & (\lambda_i + \lambda_j) I, \\ \lambda_k & \geq & 0\,. \end{array}$$

This new upper bound for μ reduces to the upper bound proposed by Fu and Barabanov in [15] when $F = F_1 = \cdots = F_L$; thus it is less conservative but also more costly. It is also less conservative than the more classical upper bound in [13], as is easily proved by choosing $F = F_1 = \cdots = F_L = -D - jM^*G$. A similar approach, though somewhat more conservative, has been proposed by Chen and Sugie in [10].

5.4. Linear parameter-varying control. In this section, we more thoroughly investigate how the concepts and tools introduced can be utilized in the context of LPV control. For clarity, we recall the general statement of the problem.

We are considering an LPV plant with state-space realization

$$\begin{array}{rcl} \dot{x} & = & A(\theta)x + B_1(\theta)w + B_2(\theta)u, \\ z & = & C_1(\theta)x + D_{11}(\theta)w + D_{12}(\theta)u, \\ y & = & C_2(\theta)x + D_{21}(\theta)w\,, \end{array} \tag{5.4}$$

where

$$A \in \mathbf{R}^{n \times n}, \quad D_{12} \in \mathbf{R}^{p_1 \times m_2}, \qquad \text{and} \quad D_{21} \in \mathbf{R}^{p_2 \times m_1}$$

define the problem dimension. It is assumed that

(A1) the state-space data $A(\theta)$, $B_1(\theta), \ldots$ are bounded continuous functions of θ,

(A2) the time-varying parameter $\theta(t) := [\theta_1(t), \ldots, \theta_N(t)]^T$ and its rate of variation $\dot{\theta}(t)$, defined at all times and continuous, evolve in hyperrectangles H and H_d, that is,

$$\theta_i(t) \in [\underline{\theta}_i, \bar{\theta}_i] \quad \forall t \geq 0\,, \tag{5.5}$$

$$\dot{\theta}_i(t) \in [\underline{\nu}_i, \bar{\nu}_i] \quad \forall t \geq 0\,. \tag{5.6}$$

The assumptions (A1) and (A2) are general. They secure existence and uniqueness of the solutions to (5.4) for given initial conditions and also specify the parameter trajectories under consideration.

With these assumptions in place, the general LPV control problem with guaranteed L_2-gain performance consists of finding a dynamic LPV controller with state-space equations

$$\begin{array}{rcl} \dot{x}_K & = & A_K(\theta,\dot{\theta})x_K + B_K(\theta,\dot{\theta})y, \\ u & = & C_K(\theta,\dot{\theta})x_K + D_K(\theta,\dot{\theta})y, \end{array} \tag{5.7}$$

which ensures internal stability and a guaranteed L_2-gain bound γ for the closed-loop operator (5.4)–(5.7) from the disturbance signal w to the error signal z, that is,

$$\int_0^T z^T z\, d\tau \le \gamma^2 \int_0^T w^T w\, d\tau \quad \forall T \ge 0$$

for all admissible parameter trajectories $\theta(t)$.

Sufficient solvability conditions for this problem can be derived using a suitable extension of the bounded real lemma [44] and by confining the search of (Lyapunov) variables to some finite-dimensional subspace of functions of θ. The next theorem provides such a set of conditions for the general LPV control problem. An alternative approach, based on polytopic covering techniques, is proposed by Yu and Sideris in [47]. For technical reasons that are clarified in the proof, we also assume that

(A3) the matrices $[\, B_2^T(\theta) \quad D_{12}^T(\theta)\,]$, $[\, C_2(\theta) \quad D_{21}(\theta)\,]$ have full row-rank over H.

Note that the dependence of data and variables on θ, or $\dot{\theta}$, is generally dropped for simplicity.

THEOREM 5.2. *With the assumptions* (A1)–(A3) *in force, the following conditions are equivalent:*

(i) *The bounded real lemma conditions with L_2-gain performance level γ hold for some quadratic Lyapunov function*

$$V(x,x_K,\theta) := \begin{bmatrix} x \\ x_K \end{bmatrix}^T P(\theta) \begin{bmatrix} x \\ x_K \end{bmatrix},$$

where $P(\theta)$ is continuously differentiable, and for some LPV controller (5.7).

(ii) *There exist continuously differentiable parameter-dependent symmetric matrices $X(\theta)$ and $Y(\theta)$ such that the following PLMI problem is feasible:*

$$\left[\begin{array}{c|c} \mathcal{N}_X & 0 \\ \hline 0 & I \end{array}\right]^T \left[\begin{array}{cc|c} \dot{X}+XA+A^TX & XB_1 & C_1^T \\ B_1^TX & -\gamma I & D_{11}^T \\ \hline C_1 & D_{11} & -\gamma I \end{array}\right] \left[\begin{array}{c|c} \mathcal{N}_X & 0 \\ \hline 0 & I \end{array}\right] < 0, \tag{5.8}$$

$$\left[\begin{array}{c|c} \mathcal{N}_Y & 0 \\ \hline 0 & I \end{array}\right]^T \left[\begin{array}{cc|c} -\dot{Y}+YA^T+AY & YC_1^T & B_1 \\ C_1Y & -\gamma I & D_{11} \\ \hline B_1^T & D_{11}^T & -\gamma I \end{array}\right] \left[\begin{array}{c|c} \mathcal{N}_Y & 0 \\ \hline 0 & I \end{array}\right] < 0, \tag{5.9}$$

$$\begin{bmatrix} X & I \\ I & Y \end{bmatrix} > 0 \tag{5.10}$$

for all $(\theta,\dot{\theta})$ on $H \times H_d$ and where $\mathcal{N}_X$ and $\mathcal{N}_Y$ designate any bases of the nullspaces of $[\, C_2 \quad D_{21}\,]$ and $[\, B_2^T \quad D_{12}^T\,]$, respectively.

(iii) *There exist continuously differentiable parameter-dependent symmetric matrices* $X(\theta)$ *and* $Y(\theta)$ *and a scalar* σ *solving the PLMI problem:*

$$\begin{bmatrix} \dot{X}+XA+A^TX & XB_1 & C_1^T \\ B_1^TX & -\gamma I & D_{11}^T \\ C_1 & D_{11} & -\gamma I \end{bmatrix} - \sigma \begin{bmatrix} C_2^T \\ D_{21}^T \\ 0 \end{bmatrix} \begin{bmatrix} C_2 & D_{21} & 0 \end{bmatrix} < 0, \tag{5.11}$$

$$\begin{bmatrix} -\dot{Y}+YA^T+AY & YC_1^T & B_1 \\ C_1Y & -\gamma I & D_{11} \\ B_1^T & D_{11}^T & -\gamma I \end{bmatrix} - \sigma \begin{bmatrix} B_2 \\ D_{12} \\ 0 \end{bmatrix} \begin{bmatrix} B_2^T & D_{12}^T & 0 \end{bmatrix} < 0\,, \tag{5.12}$$

$$\begin{bmatrix} X & I \\ I & Y \end{bmatrix} > 0 \tag{5.13}$$

for all $(\theta, \dot{\theta})$ *on* $H \times H_d$.

Proof. See Appendix A. □

Equipped with Theorem 5.2, it is relatively straightforward to show how multiconvexity concepts can be used to reduce complexity in LPV control problems with polynomial parameter-dependence.

For simplicity of the presentation, it is first assumed the the state-space data in (5.4) and the Lyapunov variables are affine functions of the parameter θ, that is,

$$A(\theta) := A_0 + \sum_{i=1}^{N} \theta_i A_i, \quad B_1(\theta) := B_{10} + \sum_{i=1}^{N} \theta_i B_{1i} \ \ \ldots. \tag{A4}$$

THEOREM 5.3. *With the assumptions* (A1)–(A4) *above, there exists an LPV controller* (5.7) *solution to the LPV control problem with guaranteed* L_2*-gain performance with level* γ *whenever there exist symmetric matrices* X_0, $X_1, \ldots, X_N$ *and* Y_0, $Y_1, \ldots,$ Y_N *and scalars* λ_0, $\lambda_1, \ldots \lambda_N$, μ_0, $\mu_1, \ldots \mu_N$, *and* σ *such that*

$$\begin{bmatrix} \dot{X}+XA+A^TX & XB_1 & C_1^T \\ B_1^TX & -\gamma I & D_{11}^T \\ C_1 & D_{11} & -\gamma I \end{bmatrix} - \sigma \begin{bmatrix} C_2^T \\ D_{21}^T \\ 0 \end{bmatrix} \begin{bmatrix} C_2 & D_{21} & 0 \end{bmatrix} < -\left(\lambda_0 + \sum_{i=1}^{N} \theta_i^2 \lambda_i\right) I, \tag{5.14}$$

$$\begin{bmatrix} -\dot{Y}+YA^T+AY & YC_1^T & B_1 \\ C_1Y & -\gamma I & D_{11} \\ B_1^T & D_{11}^T & -\gamma I \end{bmatrix} - \sigma \begin{bmatrix} B_2 \\ D_{12} \\ 0 \end{bmatrix} \begin{bmatrix} B_2^T & D_{12}^T & 0 \end{bmatrix} < -\left(\mu_0 + \sum_{i=1}^{N} \theta_i^2 \mu_i\right) I, \tag{5.15}$$

$$\begin{bmatrix} X & I \\ I & Y \end{bmatrix} > 0 \tag{5.16}$$

for $(\theta, \dot{\theta}) \in \operatorname{vert} H \times \operatorname{vert} H_d$ *and*

$$\begin{bmatrix} X_iA_i + A_i^TX_i & X_iB_{1i} \\ B_{1i}^TX_i & 0 \end{bmatrix} - \sigma \begin{bmatrix} C_{2i}^TC_{2i} & C_{2i}^TD_{21i} \\ D_{21i}^TC_{2i} & D_{21i}^TD_{21i} \end{bmatrix} \geq -\lambda_i I, \tag{5.17}$$

$$\begin{bmatrix} Y_iA_i^T + A_iY_i & Y_iC_{1i}^T \\ C_{1i}Y_i & 0 \end{bmatrix} - \sigma \begin{bmatrix} B_{2i}B_{2i}^T & B_{2i}D_{12i}^T \\ D_{12i}B_{2i}^T & D_{12i}D_{12i}^T \end{bmatrix} \geq -\mu_i I\,, \tag{5.18}$$

and

$$\lambda_0 \geq 0, \quad \lambda_i \geq 0, \quad \mu_0 \geq 0, \quad \mu_i \geq 0 \tag{5.19}$$

for $i = 1, \ldots, N$ with the notations

$$X := X_0 + \sum_{i=1}^{N} \theta_i X_i, \qquad Y := Y_0 + \sum_{i=1}^{N} \theta_i Y_i \,.$$

Proof. The proof is a direct consequence of Theorem 5.2 combined with an application of Proposition 4.1 or Lemma 4.2 to the particular case where data and variables are affine with respect to θ. ∎

Remarks. The conditions in Theorem 5.3 constitute a standard semidefinite programming problem. The linear objective γ should be minimized subject to a finite number of LMI constraints, and a number of softwares are available for this purpose. The characterization is easily modified to encompass any polynomial parameter dependence for both the state-space data and the variables $X(\theta)$ and $Y(\theta)$ by direct application of Lemma 4.2. The multiconvexity requirements in (5.17) and (5.18) can be relaxed using the simple techniques in section 4. When either the multiconvexity approach is too conservative or brute force gridding of the parameter range is too costly (more than two parameters), it might be appropriate to enforce multiconvexity along some direction and to grid a surface of lower dimension. See the examples in section 6 for illustrations.

5.4.1. LPV controller construction. The PLMI conditions (ii) and (iii) in Theorem 5.2 are equivalent and provide lossless solvability conditions for problem (i). The characterization in Theorem 5.3 may be conservative but give tractable conditions for solving the same problem. Clearly, when any of the latter problems is feasible, the state-space data (5.7) of an LPV controller solving the problem can be constructed for any pair $(\theta, \dot\theta)$ in $H \times H_d$ from any solutions $X(\theta)$, $Y(\theta)$, and σ by the very same algebraic formulae. For completeness, we provide the following sequential scheme:

(i) Compute D_K solution to

$$\overline{\sigma}(D_{11} + D_{12} D_K D_{21}) < \gamma, \tag{5.20}$$

and set $D_{c\ell} := D_{11} + D_{12} D_K D_{21}$.

(ii) Compute $\widehat{\boldsymbol{B}}_K$ and $\widehat{\boldsymbol{C}}_K$ solutions to the linear matrix equations

$$\begin{bmatrix} 0 & D_{21} & 0 \\ D_{21}^T & -\gamma I & D_{c\ell}^T \\ 0 & D_{c\ell} & -\gamma I \end{bmatrix} \begin{bmatrix} \widehat{\boldsymbol{B}}_K^T \\ \star \end{bmatrix} = - \begin{bmatrix} C_2 \\ B_1^T X \\ C_1 + D_{12} D_K C_2 \end{bmatrix}, \tag{5.21}$$

$$\begin{bmatrix} 0 & D_{12}^T & 0 \\ D_{12} & -\gamma I & D_{c\ell} \\ 0 & D_{c\ell}^T & -\gamma I \end{bmatrix} \begin{bmatrix} \widehat{\boldsymbol{C}}_K \\ \star \end{bmatrix} = - \begin{bmatrix} B_2^T \\ C_1 Y \\ (B_1 + B_2 D_K D_{21})^T \end{bmatrix}. \tag{5.22}$$

(iii) Compute

$$\begin{aligned} \widehat{\boldsymbol{A}}_K \; &= - (A + B_2 D_K C_2)^T \\ &+ \begin{bmatrix} X B_1 + \widehat{\boldsymbol{B}}_K D_{21} & (C_1 + D_{12} D_K C_2)^T \end{bmatrix} \begin{bmatrix} -\gamma I & D_{c\ell}^T \\ D_{c\ell} & -\gamma I \end{bmatrix}^{-1} \begin{bmatrix} (B_1 + B_2 D_K D_{21})^T \\ C_1 Y + D_{12} \widehat{\boldsymbol{C}}_K \end{bmatrix}. \end{aligned} \tag{5.23}$$

(iv) Solve for N, M, the factorization problem

$$I - XY = N M^T \,.$$

(v) Finally, compute A_K, B_K, and C_K with the help of

$$A_K = N^{-1}(X\dot{Y} + N\dot{M}^T + \widehat{\boldsymbol{A}}_K - X(A - B_2 D_K C_2)Y - \widehat{\boldsymbol{B}}_K C_2 Y - X B_2 \widehat{\boldsymbol{C}}_K) M^{-T}, \tag{5.24}$$

$$B_K = N^{-1}(\widehat{\boldsymbol{B}}_K - X B_2 D_K), \tag{5.25}$$

$$C_K = (\widehat{\boldsymbol{C}}_K - D_K C_2 Y) M^{-T}. \tag{5.26}$$

The reader might consult [17, 16, 22, 37, 1] for details on this construction.

6. Numerical examples. In this section, the concepts and tools developed above are illustrated by some numerical examples. All LMI-related computations were performed on an ULTRA 1 SUN station using the LMI control toolbox [19].

6.1. Stability analysis. We consider the following example from [39]. The A matrix of the uncertain system is given in the form

$$A(\theta_1, \theta_2) = \begin{bmatrix} -2+\theta_1 & 0 & -1+\theta_1 \\ 0 & -3+\theta_2 & 0 \\ -1+\theta_1 & -1+\theta_2 & -4+\theta_1 \end{bmatrix}.$$

We are seeking the maximum rectangle in the (θ_1, θ_2) space for which stability is guaranteed. In this context, Proposition 5.1 is directly applicable to the polytope of extreme values of the parameters θ_1 and θ_2. The uncertain system is found stable for all values of θ_1 and θ_2 in the rectangle

$$-1e6 \le \theta_1 \le 1.7499, \qquad -1e6 \le \theta_2 \le 2.99\,.$$

This result is consistent with the true domain of stability ($\theta_1 < 1.75$, $\theta_2 < 3$) and is markedly superior to existing results [39].

6.2. LPV synthesis example. The following example provides an illustration of the proposed LPV control synthesis techniques. The discussion emphasizes the complexity and cost associated with various LPV synthesis strategies. The problem setup comes from [33]. It has been slightly complicated to incorporate two time-varying parameters for illustration purposes while retaining the same design specifications.

The LPV model of the longitudinal dynamics of the missile is given as

$$\begin{bmatrix} \dot{\alpha} \\ \dot{q} \end{bmatrix} = \begin{bmatrix} -0.89 & 1 \\ -142.6 & \end{bmatrix} \begin{bmatrix} \alpha \\ q \end{bmatrix} + \begin{bmatrix} 0 & -0.89 \\ 178.25 & 0 \end{bmatrix} \begin{bmatrix} w_{\theta_1} \\ w_{\theta_2} \end{bmatrix} + \begin{bmatrix} -0.119 \\ -130.8 \end{bmatrix} \delta_{\text{fin}},$$

$$\begin{bmatrix} w_{\theta_1} \\ w_{\theta_2} \end{bmatrix} = \begin{bmatrix} \theta_1 & 0 \\ 0 & \theta_2 \end{bmatrix} \begin{bmatrix} -1 & 0 \\ 1 & 0 \end{bmatrix} \begin{bmatrix} \alpha \\ q \end{bmatrix},$$

$$\begin{bmatrix} \eta_z \\ q \end{bmatrix} = \begin{bmatrix} -1.52 & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} \alpha \\ q \end{bmatrix},$$

where α, q, η_z, and δ_{fin} denote the angle of attack, the pitch rate, the vertical accelerometer measurement, and the fin deflection, respectively; and θ_1, θ_2 are two time-varying parameters, measured in real time, resulting from changes in missile aerodynamic conditions (angle of attack from 0 up to 20 degrees). The synthesis structure used in this problem is depicted in Figure 6.1.

The problem specifications are as follows:

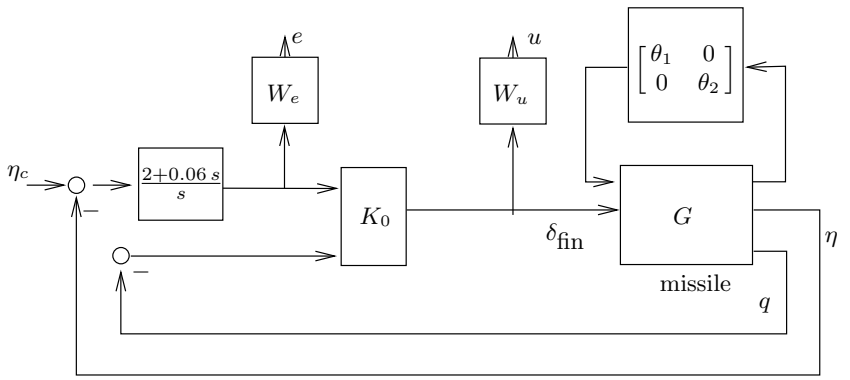


FIG. 6.1. *Synthesis structure.*

(i) A settling time of 0.2 second with minimal overshoot and zero steady-state error for the vertical acceleration η_z in response to a step command η_c.

(ii) The controller must achieve an adequate high-frequency roll-off for noise attenuation and to withstand neglected dynamics and flexible modes. Magnitude constraints of 2 are also imposed to the control signal δ_{fin}.

Moreover, those specifications must be met for all parameter values:

$$|\theta_1| \le 1, \qquad |\theta_2| \le 1\,.$$

An integrator has been introduced on the acceleration channel to ensure zero steady-state error. It turns out that the resulting LPV controller K is obtained as the composition of the operators K_0 and

$$\begin{bmatrix} \frac{2+0.06\,s}{s} & 0 \\ 0 & 1 \end{bmatrix}.$$

The weighting functions W_e and W_u were chosen to be

$$W_e = 0.8, \qquad W_u = \frac{0.001s^3 + 0.03s^2 + 0.3s + 1}{\text{1e-5}s^3 + \text{3e-2}s^2 + 30s + 10000}\,.$$

The design synthesis consists of the computation of a parameter-dependent controller $K_0(\theta_1, \theta_2)$ such that all specifications above are met. For simplicity of the discussion, we assume that the LPV model can be considered as a parameterized family of linear time-varying models. Similar conclusions can be drawn with time-varying parameters with bounded rates of variation. The synthesis problem is attacked via three different strategies with increasing conservatism and decreasing computational effort:

(i) The full gridding approach makes use of a 6×6 point gridding of the parameter range of (θ_1, θ_2).

(ii) The mixed strategy uses a grid in the θ_2 direction and enforces multiconvexity along the θ_1 direction.

(iii) The multiconvexity approach enforces multiconvexity in both directions θ_1 and θ_2.

Results and numerical features of each technique are collected in Table 6.1.

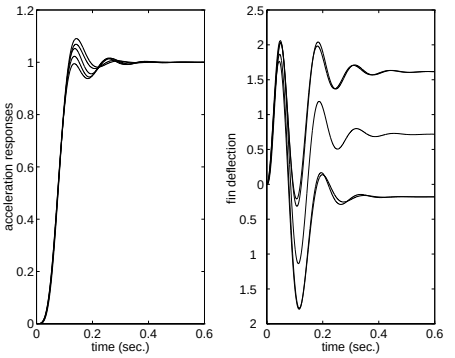


FIG. 6.2. *Time domain responses—LPV controller* 1 *full-gridding technique.*

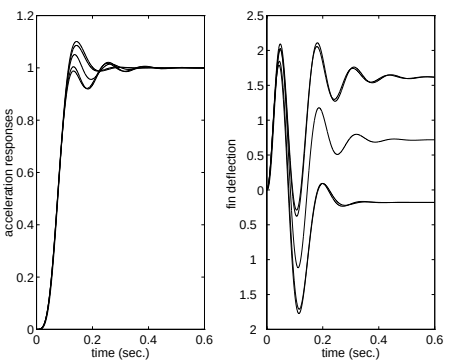


FIG. 6.3. *Time domain responses—LPV controller* 2 *mixed technique.*

TABLE 6.1
Numerical comparisons of LPV synthesis techniques.

	# of gridding points	# of LMIs	cputime	Achieved perf. level
Full gridding	36	108	17 min. 24 sec.	0.1265
Mixed strategy	12	30	6 min. 20 sec.	0.1282
Multiconvexity	0	16	3 min. 12 sec.	0.1293

It is instructive to see that all techniques provide about the same performance level. This indicates that there is no significant growth of conservatism when using multiconvexity concepts to reduce or eradicate the gridding points. This is confirmed by the time-domain simulations in Figures 6.2–6.4, which correspond, for each derived LPV controller, to parameter values at the vertices and the center of the (θ_1, θ_2) range. Performance specs, as well as the roll-off property of the controllers, have been found to be satisfactory for each technique. In spite of the consistency in the results, it must be born in mind that the multiconvex synthesis is the only one to provide theoretical stability and performance guarantees at any operating condition of the parameter range. The full-gridding technique gives similar guarantees solely at the grid points, and the achieved performance γ is necessarily a lower bound of the actual performance. The situation is slightly more embarrassing for the mixed strategy since the performance level is overestimated in the direction of θ_2 and underestimated in the

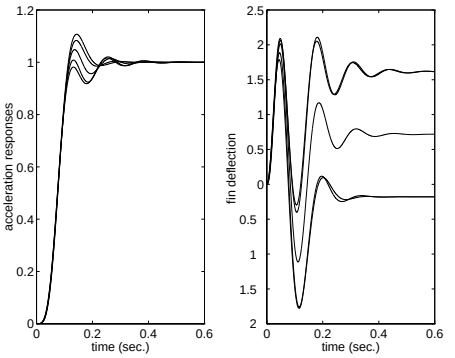


FIG. 6.4. *Time domain responses—LPV controller* 3 *multiconvexity technique.*

direction of θ_1. So we cannot decide whether the result is conservative or optimistic on the whole parameter range. Nevertheless, the approach is of practical interest for computational reasons. It appears clearly in Table 6.1 that LPV syntheses exploiting either partial or complete multiconvexity are significantly cheaper than full-gridding techniques. This difference is likely to be even more dramatic for problems involving more than two parameters for which full gridding is practically prohibited. This is a direct consequence of the exponential growth of the number of LMIs in the full-gridding approach. Note that we do not account for scalar dimensional constraints of the type λ_i, $\mu_i \geq 0$ in Table 6.1, as they negligibly affect the overall computational time.

Any of the LPV synthesis techniques considered in this section turn out to be less conservative than linear fractional transformation (LFT) gain-scheduling techniques [30, 2, 20, 38] which disregard the parameter variation rates. About 10 percent degradation of the performance level has been observed in this simple application. Finally, the techniques behave as theoretically expected and provide valuable LPV synthesis alternatives.

7. Conclusion. A general framework for relaxing PLMI problems into conventional LMI problems has been introduced. The techniques are simple and exploit or enforce directional convexity properties of PLMI problems. A nonexhaustive list of implications of the proposed techniques in control theory have been examined with a particular focus on LPV synthesis, a most important emerging technique in recent years.

This work raises some open questions, some of which might be beyond reach, but also suggests some directions for future research:

(i) For affine PLMI problems, directional convexity concepts are less conservative than the **S**-procedure but a theoretical comparison is still lacking in the general case. From the viewpoint of computational efforts, one can hardly draw definitive conclusions but the proposed approach is better exploited by using primal SDP interior-point techniques since it involves less decision variables than the **S**-procedure. We note experimentally that the multiconvexity approach is more efficient for problems with significantly more states than parameters, which is a common situation in control applications.

(ii) An unsolved issue is the following: Is it possible to exploit directional quasi-convexity instead of directional convexity for some classes of PLMI problems?

(iii) Other topics not examined in this paper and for which the proposed techniques might prove useful are robust least-squares and robust interpolation and approximation. The relaxation of some intractable generic robust convex programs is also of interest.

Appendix A. Proof of Theorem 5.2. Following [44], the LPV control problem with guaranteed L_2-gain performance γ is solvable whenever one can find an LPV controller such that a suitable extension of the bounded real lemma is satisfied with a quadratic parameter-dependent Lyapunov function, continuously differentiable with respect to θ. This is nothing else than statement (i) of the theorem.

In turn, the latter conditions are equivalent to the existence of continuously parameter-dependent symmetric matrices $X(\theta)$ and $Y(\theta)$ such that LMI conditions (5.8)–(5.10) hold for all $(\theta, \dot{\theta})$ on $H \times H_d$. This assertion is a slight extension of the main result in [44, 17, 22]. Assume a (closed-loop) Lyapunov function establishing L_2-gain performance is

$$V(x, x_K, \theta) := \begin{bmatrix} x \\ x_K \end{bmatrix}^T P(\theta) \begin{bmatrix} x \\ x_K \end{bmatrix},$$

where

$$P := \begin{bmatrix} X & N \\ N^T & \star \end{bmatrix} \quad \text{and} \quad P^{-1} := \begin{bmatrix} Y & M \\ M^T & \star \end{bmatrix}.$$

It trivially holds that

$$I - XY = NM^T.$$

Then, defining

$$\Pi_Y := \begin{bmatrix} Y & I \\ M^T & 0 \end{bmatrix}, \qquad \Pi_X := \begin{bmatrix} I & X \\ 0 & N^T \end{bmatrix}$$

yields the identities $P\Pi_Y = \Pi_X$, and

$$\Pi_Y^T \frac{d}{dt} P \Pi_Y = \begin{bmatrix} -\dot{Y} & -(X\dot{Y} + N\dot{M}^T)^T \\ -(X\dot{Y} + N\dot{M}^T) & \dot{X} \end{bmatrix},$$

which is the only additional term with respect to the customary H_∞ control problem in [17]. This establishes statement (ii).

To prove assertion (iii), we first note that the rates of variation $\dot{\theta}_i$ are involved linearly in (5.8) and (5.9), and thus it suffices to assess feasibility of these LMIs over $H \times \operatorname{vert} H_d$. By virtue of Finsler's lemma [31, 22], the LMIs (5.8)–(5.9) with $(\theta, \dot{\theta})$ ranging over $H \times \operatorname{vert} H_d$ are feasible if and only if there exists a function $\sigma(.)$ of θ such that

$$\begin{bmatrix} \dot{X} + XA + A^T X & XB_1 & C_1^T \\ B_1^T X & -\gamma I & D_{11}^T \\ C_1 & D_{11} & -\gamma I \end{bmatrix} - \sigma(\theta) \begin{bmatrix} C_2^T \\ D_{21}^T \\ 0 \end{bmatrix} \begin{bmatrix} C_2 & D_{21} & 0 \end{bmatrix} < 0, \tag{A.1}$$

$$\begin{bmatrix} -\dot{Y} + YA^T + AY & YC_1^T & B_1 \\ C_1 Y & -\gamma I & D_{11} \\ B_1^T & D_{11}^T & -\gamma I \end{bmatrix} - \sigma(\theta) \begin{bmatrix} B_2 \\ D_{12} \\ 0 \end{bmatrix} \begin{bmatrix} B_2^T & D_{12}^T & 0 \end{bmatrix} < 0. \tag{A.2}$$

Therefore, we end up with an infinite set of LMIs whose members are all of the form

$$\Psi(\theta) - \sigma(\theta) R(\theta) R(\theta)^T < 0, \qquad \theta \in H\,. \tag{A.3}$$

Denote $R_\perp$ a continuous basis of the nullspace of R^T. This is always possible [21, 25] by virtue of assumption (A3). It follows that $[\, R(\theta) \quad R_\perp(\theta)\,]$ is a continuous invertible matrix over H. With the *proviso* that the LMIs (A.3) are feasible, it is not difficult to see using Schur complements that admissible σ are described as

$$\sigma(\theta) > l(\theta) := \overline{\lambda}\left\{(R^T\Psi R - (R^T\Psi R_\perp)(R_\perp^T\Psi R_\perp)^{-1}(R^T\Psi R_\perp)^T)(R^T R R^T R)^{-1}\right\}, \quad \theta \in H, \tag{A.4}$$

where $\overline{\lambda}(.)$ stands for the maximum eigenvalue of a matrix. From the continuity of both the plant's state-space data (assumption (A1)) and the variables $X(\theta)$ and $Y(\theta)$, we deduce that $l(\theta)$ in (A.4) is also continuous with respect to θ. Now since H is a compact set, the choice

$$\sigma(\theta) := \sigma > \sup_{\theta \in H} l(\theta)$$

is again a valid choice for σ. It follows that the LMIs (A.1) and (A.2) are feasible if and only if this is so for a *constant* sufficiently large σ. This completes the proof of the theorem. □

REFERENCES

[1] P. Apkarian and R. J. Adams, *Advanced gain-scheduling techniques for uncertain systems*, IEEE Trans. Control System Tech., 6 (1997), pp. 21–32.

[2] P. Apkarian and P. Gahinet, *A convex characterization of gain-scheduled H_∞ controllers*, IEEE Trans. Automat. Control, 40 (1995), pp. 853–864, p. 1681.

[3] D. Arzelier, J. Bernussou, and J. M. Garcia, *About quadratic stabilizability of generalized linear systems*, in Proceedings of the IEEE Conference on Decision and Control, Brighton, UK, 1991, pp. 914–920.

[4] A. Ben-Tal and A. Nemirovski, *Robust convex optimization*, Math. Oper. Res., 23 (1998), pp. 769–805.

[5] A. Ben-Tal and A. Nemirovski, *Robust solutions to uncertain linear problems*, Oper. Res. Lett., 25 (1999), pp. 1–13.

[6] A. Ben-Tal and A. Nemirovski, *Robust Truss topology design via semidefinite programming*, SIAM J. Optim., 7 (1997), pp. 991–1016.

[7] J. Bernussou, P. L. D. Peres, and J. C. Geromel, *A linear programming oriented procedure for quadratic stabilization of uncertain systems*, Systems Control Lett., 13 (1989), pp. 65–72.

[8] S. Boyd and L. ElGhaoui, *Method of centers for minimizing generalized eigenvalues*, Linear Algebra Appl., 188 (1992), pp. 63–111.

[9] S. Boyd, L. ElGhaoui, E. Feron, and V. Balakrishnan, *Linear Matrix Inequalities in Systems and Control Theory*, SIAM Stud. Appl. Math., 15, SIAM, Philadelphia, 1994.

[10] G. Chen and T. Sugie, *New upper bound of the real μ on the parameter dependent multiplier*, Proceedings of the IEEE Conference on Decision and Control, 1996, pp. 1293–1294.

[11] J. Doyle, *Analysis of feedback systems with structured uncertainties*, Proc. IEE-D, 129 (1982), pp. 242–250.

[12] J. Doyle, J. E. Wall, and G. Stein, *Performance and robustness analysis for structured uncertainties*, Proceedings of the IEEE Conference on Decision and Control, 1982, pp. 629–636.

[13] M. K. H. Fan, A. L. Tits, and J. C. Doyle, *Robustness in the presence of mixed parametric uncertainty and unmodeled dynamics*, IEEE Trans. Automat. Control, AC-36 (1991), pp. 25–38.

[14] A. L. Fradkov and V. A. Yakubovich, *The $\mathcal{S}$-procedure and duality relations in nonconvex problems of quadratic programming*, Vestn. Lening. Univ. Math., 6 (1979), pp. 101–109. In Russian, 1973.

[15] M. Fu and N. E. Barabanov, *Improved upper bounds for the mixed structured singular value*, IEEE Trans. Automat. Control, 42 (1997), pp. 1447–1452.

[16] P. Gahinet, *Explicit controller formulas for LMI-based H_∞ synthesis*, in Proceedings of the American Control Conference, Baltimore, MD, IFAC, 1994, pp. 2396–2400.

[17] P. Gahinet and P. Apkarian, *A linear matrix inequality approach to H_∞ control*, Internat. J. Robust Nonlinear Control, 4 (1994), pp. 421–448.

[18] P. Gahinet, P. Apkarian, and M. Chilali, *Parameter-dependent Lyapunov functions for real parametric uncertainty*, IEEE Trans. Automat. Control, 41 (1996), pp. 436–442.

[19] P. Gahinet, A. Nemirovski, A. J. Laub, and M. Chilali, *LMI Control Toolbox*, The MathWorks Inc., Natick, MA, 1995.

[20] A. Helmersson, *Methods for Robust Gain-Scheduling*, Ph.D. thesis, Linkoping University, Sweden, 1995.

[21] A. Isidori, *Nonlinear Control Systems: An Introduction*, 2nd ed., Springer-Verlag, Berlin, 1989.

[22] T. Iwasaki and R. E. Skelton, *All controllers for the general H_∞ control problem: LMI existence conditions and state space formulas*, Automatica J. IFAC, 30 (1994), pp. 1307–1317.

[23] F. Jarre, *An interior-point method for minimizing the maximum eigenvalue of a linear combination of matrices*, SIAM J. Control Optim., 31 (1993), pp. 1360–1377.

[24] H. Konno, P. Thach, and H. Tuy, *Optimization on Low Rank Nonconvex Structures*, Kluwer Academic Publishers, Norwell, MA, 1997.

[25] W. Lu and J. C. Doyle, *H_∞ Control of nonlinear systems: A convex characterization*, IEEE Trans. Automat. Control, 40 (1995), pp. 1668–1674.

[26] G. Meinsma, Y. Shrivastava, and M. Fu, *A dual formulation of mixed μ and the losslessness of (D, G)-scaling*, in Proceedings of the IEEE Conference on Decision and Control, Kobe, Japan, 1996, pp. 1287–1292.

[27] A. Nemirovski and P. Gahinet, *The projective method for solving linear matrix inequalities*, Math. Programming, 77 (1997), pp. 163–190.

[28] Y. E. Nesterov and A. S. Nemirovski, *Interior Point Polynomial Methods in Convex Programming: Theory and Applications*, SIAM Stud. Appl. Math. 13, SIAM, Philadelphia, 1994.

[29] F. Oustry, L. Elghaoui, and H. Lebret, *Robust solutions to uncertain semidefinite programs*, SIAM J. Optim., 9 (1998), pp. 33–52.

[30] A. Packard, *Gain scheduling via linear fractional transformations*, Systems Control Lett., 22 (1994), pp. 79–92.

[31] I. R. Petersen and C. V. Hollot, *A Riccati equation approach to the stabilization of uncertain linear systems*, Automatica J. IFAC, 22 (1986), pp. 397–411.

[32] S. Poljak and J. Rohn, *Checking robust nonsingularity is NP-complete*, Math. Control Signals Systems, 6 (1994), pp. 1–9.

[33] R. T. Reichert, *Robust autopilot design using μ-synthesis*, in Proceedings of the American Control Conference, San Diego, CA, IFAC, 1990, pp. 2368–2373.

[34] F. Rendl, R. Vanderbei, and H. Wolkowicz, *A primal-dual interior-point method for the max-min eigenvalue problem*, Tech. report CORR 93-30, University of Waterloo, Dept. of Combinatorics and Optimization, Waterloo, Ontario, Canada, 1993.

[35] I. Rosenberg, $0-1$ *optimization and nonlinear programming*, Rev. Francaise Automat. Informat. Recherche Operationelle, 6 (1972), pp. 95–97.

[36] M. G. Safonov and J. Doyle, *Minimizing conservativeness of robust singular values*, in Multivariable Control, S. G. Tzafestas, ed., Reidel, Dordrecht, Boston, London, 1984, pp. 197–207.

[37] C. Scherer, *Mixed H_2/H_∞ control*, in Trends in Control: A European Perspective, Special Contribution to the European Control Conference 95, Springer-Verlag, Berlin, 1995.

[38] G. Scorletti and L. E. Ghaoui, *Improved linear matrix inequality conditions for gain-scheduling*, in Proceedings of the IEEE Conference on Decision and Control, New Orleans, LA, 1995, pp. 3626–3631.

[39] D. D. Siljak, *Parameter space methods for robust control design: A guided tour*, IEEE Trans. Automat. Control, 34 (1989), pp. 674–688.

[40] H. D. Tuan and P. Apkarian, *Relaxations of parameterized* LMIs *with control applications*, Internat. J. Robust Nonlinear Control, 9 (1999), pp. 59–84.

[41] H. Tuy, *Convex Analysis and Global Optimization*, Kluwer Academic Publishers, Norwell, MA,

1998.

[42] L. Vandenberghe and S. Boyd, *Primal-dual potential reduction method for problems involving matrix inequalities*, Math. Programming Ser. B, 69 (1995), pp. 205–236.

[43] L. Vandenberghe and S. Boyd, *Semidefinite programming*, SIAM Rev., 38 (1996), pp. 49–95.

[44] F. Wu, X. Yang, A. Packard, and G. Becker, *Induced L_2-norm control for LPV system with bounded parameter variations rates*, Internat. J. Robust Nonlinear Control, 6 (1996), pp. 983–998.

[45] V. A. Yakubovich, *The $\mathcal{S}$-procedure in non-linear control theory*, Vestnik Leningrad Univ. Math., 4 (1977), pp. 73–93. In Russian, 1971.

[46] P. M. Young, M. P. Newlin, and J. C. Doyle, *μ analysis with real parametric uncertainty*, in Proceedings of the IEEE Conference on Decision and Control, vol. 2, Brighton, UK, 1991, pp. 1251–1256.

[47] J. Yu and A. Sideris, *H_∞ control with parametric Lyapunov functions*, Systems Control Lett., 30 (1997), pp. 57–69.

SIAM J. CONTROL OPTIM.
Vol. 38, No. 4, pp. 1265–1287

PIERI HOMOTOPIES FOR PROBLEMS IN ENUMERATIVE GEOMETRY APPLIED TO POLE PLACEMENT IN LINEAR SYSTEMS CONTROL*

BIRKETT HUBER† AND JAN VERSCHELDE‡

Abstract. Huber, Sottile, and Sturmfels [*J. Symbolic Comput.*, 26 (1998), pp. 767–788] proposed Pieri homotopies to enumerate all p-planes in $\mathbb{C}^{m+p}$ that meet n given $(m+1-k_i)$-planes in general position, with $k_1+k_2+\cdots+k_n=mp$ as a condition to have a finite number of solution p-planes. Pieri homotopies turn the deformation arguments of classical Schubert calculus into effective numerical methods by expressing the deformations algebraically and applying numerical path-following techniques. We describe the Pieri homotopy algorithm in terms of a poset of simpler problems. This approach is more intuitive and more suitable for computer implementation than the original chain-oriented description and provides also a self-contained proof of correctness. We extend the Pieri homotopies to the quantum Schubert calculus problem of enumerating all polynomial maps of degree q into the Grassmannian of p-planes in $\mathbb{C}^{m+p}$ that meet $mp+q(m+p)$ given m-planes in general position sampled at $mp+q(m+p)$ interpolation points. Our approach mirrors existing counting methods for this problem and yields a numerical implementation for the dynamic pole placement problem in the control of linear systems.

Key words. cheater's homotopy, combinatorial root count, continuation methods, control theory, dynamic pole placement problem, enumerative geometry, Grassmannian, linear system, localization pattern, numerical Schubert calculus, Pieri homotopy, polynomial system, poset, quantum Schubert calculus

AMS subject classifications. 14N10, 14M15, 65H10, 68Q40, 93B27, 93B55

PII. S036301299935657X

1. Introduction.

A general method to solve geometric problems proceeds by moving the input data to a special position, solving the special configuration, and then deforming the solutions of the special problem into those of the original configuration. As long as the number of solutions remains finite, this number is constant during deformation. This principle of "conservation of number" was developed by Pieri [27], Schubert [34], and Zeuthen [50].

A classical example in enumerative geometry as explained in [18] consists of finding the two lines in projective 3-space that meet four given lines. More generally, we want to enumerate all p-planes in $\mathbb{C}^{m+p}$ that meet mp given m-planes. In the late nineteenth century, Schubert [35] established recursive and explicit formulas for the *number* of p-planes meeting a set of mp m-planes in general position. This problem was also treated by Pieri in [27]. In the early 1980s, Brockett and Byrnes [5] showed that these p-planes correspond to feedback laws which control a machine whose evolution is governed by a linear system of first-order differential equations. This problem of control theory [6], [10], [17], [48] is known as the (static) pole placement problem.

*Received by the editors May 20, 1999; accepted for publication (in revised form) October 21, 1999; published electronically May 2, 2000. This work was completed while the authors were postdoctorate at the Mathematical Sciences Research Institute, 1000 Centennial Drive, Berkeley, CA 94720-5070.
http://www.siam.org/journals/sicon/38-4/35657.html

†Wolfram Research, Inc., 100 Trade Center Drive, Champaign, IL 61820-7237 (birkh@wolfram.com).

‡Department of Mathematics, Statistics, and Computer Science, University of Illinois at Chicago, 851 S. Morgan St., Chicago, IL (jan.verschelde@na-net.ornl.gov, http://www.mth.msu.edu/~jan).
The research of this author was supported in part by the NSF under grant DMS-9804846 while at the Department of Mathematics, Michigan State University, East Lansing, MI 48824-1027.

Based on his geometric proof of Pieri's formula [27], Sottile introduced intrinsic deformations in [38] to solve general intersection conditions given by n planes in general position of dimension $m+1-k_i$ for $i=1,2,\ldots,n$. The condition that $k_1+k_2+\cdots+k_n=mp$ guarantees a finite number of p-planes meeting those n given planes. In [16], Huber, Sottile, and Sturmfels reinterpreted these deformations as computational procedures. The resulting algorithm was called the Pieri homotopy algorithm.

Our understanding arose from a first implementation of the Pieri homotopy algorithm and yielded the notion of what we call a "localization pattern." We feel that this is the key concept in our derivation of the algorithm. It is the vehicle through which the combinatorial techniques for enumerating solutions are translated back into synthetic geometry and realized. Since our aim is to understand practical complexity issues, we excuse ourselves from the explicit use of Schubert varieties (as in [15]) in our description. Nevertheless we obtain a self-contained proof of correctness, independent of [16]. The backbone of our approach is the poset of localization patterns used to count the roots combinatorially [43]. We use this poset to set up the homotopies and to control the flow of data between them. Our current description of the algorithm in [16] is both simpler than the original and also more suitable for computer implementation.

A novel feature of this paper is the development of Pieri homotopies for the dynamic pole placement problem. The connection between this problem and the Schubert calculus was established by Rosenthal in [31] and further developed in [28], [29], [46], and [47]. We arrived at this generalization thanks to a hint by Sottile (explicitly pronounced in [40]) by reverse engineering the root counting procedure described in [30]. Given $mp+q(m+p)$ general m-planes in $\mathbb{C}^{m+p}$ sampled at $mp+q(m+p)$ interpolation points, we can compute all polynomial maps of degree q producing p-planes that meet those given m-planes at the prescribed interpolation points.

Other root counts for these enumerative problems are given in [4]. In [42], quantum Gröbner and SAGBI homotopies are derived. The adjective "quantum" refers to the important connections between this problem of enumerative geometry and physics [7], although as mentioned in [4], one could also speak of "modular" Schubert calculus because of the "modulo $(m+p)$" calculations. Another related problem is the eigenvalue completion problem [13]. We mention [3] as a recent alternative approach to the dynamic pole placement problem.

Homotopies for enumerative geometry are available in a separate module of the publicly available software PHCpack [44]. The Pieri homotopy algorithm consists of a combinatorial root count, the symbolic setup of the homotopies, and the tracing of the solution curves. This last stage accounts for most of the computational work and is achieved using numerical continuation methods. See [1], [22] for introductions to and [2], [19] for recent surveys of numerical continuation. Computational experience suggests that the Pieri homotopies perform better than the Gröbner and SAGBI homotopies described in [16]. An explanation for this might be that the Pieri homotopies are more finely tuned to the geometry of the intersection problem than these other, extrinsic methods. In addition we point out that those Gröbner and SAGBI homotopies are restricted to the hypersurface $(k_i=1)$ intersection conditions.

In our study of the Pieri homotopy algorithm, we first specialized the general description of [16] into the simpler case of hypersurface intersection conditions. Our presentation below reflects this progression from the special to the general. In the next section we show how to find the two lines in projective 3-space that meet four

given lines. This example is then generalized to hypersurface and general intersection conditions in the third and fourth sections. In the fifth section we describe our quantum Pieri homotopies and their use in solving the dynamic pole placement problem. The key step is the extension of our notion of localization patterns to degree q-maps of p-planes. The development of numerical algorithms for the pole placement problem is formulated as an open problem in [33]. At the end of this paper we list computational experiences to illustrate the efficiency of our solution method.

2. The geometry of the Pieri deformations. We can illustrate the geometry of the Pieri deformations on the simplest example of determining the two lines meeting four given lines in projective 3-space.

Recall that projective 3-space is the set of 1-dimensional subspaces of $\mathbb{C}^4$. We write $\mathbb{C}^4$ as the span $\langle \mathbf{e}_1, \mathbf{e}_2, \mathbf{e}_3, \mathbf{e}_4 \rangle$ of the standard basis vectors $\mathbf{e}_1 = (1,0,0,0)$, $\mathbf{e}_2 = (0,1,0,0)$, $\mathbf{e}_3 = (0,0,1,0)$, and $\mathbf{e}_4 = (0,0,0,1)$. Denote the four given lines by L_1, L_2, L_3, and L_4. Without loss of generality we assume that L_3 is spanned by the first two basis vectors, $L_3 = \langle \mathbf{e}_1, \mathbf{e}_2 \rangle$, and L_4 by the last two basis vectors, $L_4 = \langle \mathbf{e}_3, \mathbf{e}_4 \rangle$. In general, the other two lines L_1 and L_2 do not meet each other. This general situation is at the right of Figure 2.1. Here we are visualizing the real positive orthant of projective 3-space as the interior of the tetrahedron spanned by $\mathbf{e}_1$, $\mathbf{e}_2$, $\mathbf{e}_3$, and $\mathbf{e}_4$.

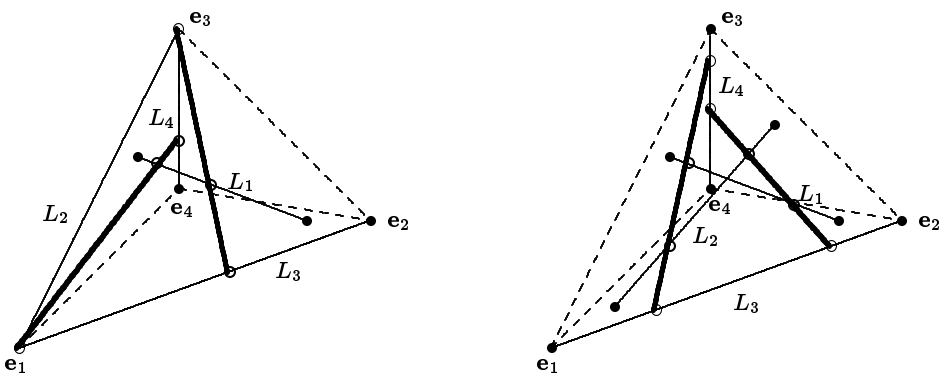


FIG. 2.1. *In $\mathbb{P}^3$ two thick lines meet four given lines L_1, L_2, L_3, and L_4 in a point. At the left we see a special configuration and the general configuration is at the right.*

The special position of L_3 and L_4 allows the following special choice of coordinates for lines X which meet them:

$$X = \langle \mathbf{x}_1, \mathbf{x}_2 \rangle = \begin{bmatrix} x_{11} & 0 \\ x_{21} & 0 \\ 0 & x_{32} \\ 0 & x_{42} \end{bmatrix}. \tag{2.1}$$

Here the two columns of X, its first and second generators, are chosen to be its points of intersection with L_3 and L_4, respectively. Lines expressed in these coordinates already satisfy the intersection conditions imposed by L_3 and L_4. Since the columns of X may be scaled independently we have exactly the two degrees of freedom needed to meet the two remaining intersection conditions imposed by L_1 and L_2.

If we take $L_2 = \langle \mathbf{e}_1, \mathbf{e}_3 \rangle$, as at the left of Figure 2.1, we solve the problem as follows. The lines passing through both L_2 and L_3 form the plane $\langle \mathbf{e}_1, \mathbf{e}_2, \mathbf{e}_3 \rangle$ which

intersects L_4 at $\mathbf{e}_3$. Similarly we see that lines through L_2 and L_4 form the plane $\langle \mathbf{e}_1, \mathbf{e}_3, \mathbf{e}_4 \rangle$ which must intersect L_3 at $\mathbf{e}_1$. Thus we see that any line other than L_2 which meets L_2, L_3, and L_4 must contain either $\mathbf{e}_1$ or $\mathbf{e}_3$. In the first case, consider the set of lines passing through $\mathbf{e}_1$ and intersecting L_4. These lines form a plane $\langle \mathbf{e}_1, \mathbf{e}_3, \mathbf{e}_4 \rangle$ which must intersect L_1 in a unique point $\mathbf{a}$ and thus the line $\langle \mathbf{e}_1, \mathbf{a} \rangle$ is the unique line through $\mathbf{e}_1$ intersecting L_1, L_2, L_3, and L_4. In the second case, the set of lines through $\mathbf{e}_3$ which intersect L_3 form a plane which intersects L_1 in a unique point $\mathbf{a}'$. So we find that the line $\langle \mathbf{e}_3, \mathbf{a}' \rangle$ is the unique line passing through $\mathbf{e}_3$ and intersecting L_1, L_2, L_3, and L_4.

An algebraic way to look at this argument goes as follows. Any line which meets both L_3 and L_4 must be expressible in the form (2.1). In order for X to meet $L_2 = \langle \mathbf{e}_1, \mathbf{e}_3 \rangle$ we must be able to find nontrivial linear combinations such that $\lambda_1 \mathbf{x}_1 + \lambda_2 \mathbf{x}_2 = \gamma_1 \mathbf{e}_1 + \gamma_2 \mathbf{e}_3$. That is, the system

$$\begin{bmatrix} 1 & x_{11} & 0 & 0 \\ 0 & x_{21} & 0 & 0 \\ 0 & 0 & 1 & x_{32} \\ 0 & 0 & 0 & x_{42} \end{bmatrix} \begin{bmatrix} -\gamma_1 \\ \lambda_1 \\ -\gamma_2 \\ \lambda_2 \end{bmatrix} = 0 \tag{2.2}$$

must have a nontrivial solution. This can clearly happen only when $x_{21} = 0$ or $x_{42} = 0$. The first case then describes the plane of lines containing $\mathbf{e}_1$ and meeting L_4 and gives rise to the solution $\langle \mathbf{e}_1, \mathbf{a} \rangle$ as above, while $x_{42} = 0$ yields the solution $\langle \mathbf{e}_3, \mathbf{a}' \rangle$.

Schubert's principle [34], [50] of "conservation of number" suggests that the number of solutions to the general problem will be the same as the number of the solutions in the special case we just looked at, i.e., 2. In order to adjust this argument to allow us to *find* the solutions as well as count them we must consider the process of deforming L_2 to special position.

The special position we use for L_2 is given by $S = \langle \mathbf{e}_1, \mathbf{e}_3 \rangle$, so our deformed problem is the following: find all X such that $\det(X|L_1) = 0$, $\det(X|S) = 0$, $\det(X|L_3) = 0$, and $\det(X|L_4) = 0$. Note that the last two equations are already satisfied when X is chosen as in (2.1). Specializing the second condition can be achieved by moving the line S into L_2 taking a convex combination $(1-t)S + tL_2$ of the generating matrices S and L_2 with t varying from 0 to 1. Fixing the first condition and enforcing nontrivial intersection with the moving line yields the homotopy

$$H(X,t) = \begin{cases} \det(X|(1-t)S + tL_2) = 0, & \text{where} \quad X = \langle \mathbf{x}_1, \mathbf{x}_2 \rangle, \\ \det(X|L_1) = 0, & \mathbf{x}_1 \in L_3, \mathbf{x}_2 \in L_4. \end{cases} \tag{2.3}$$

Numerical continuation methods trace the solution paths starting at the two solutions $X = \langle \mathbf{e}_1, \mathbf{a} \rangle$ and $X = \langle \mathbf{e}_3, \mathbf{b} \rangle$ calculated above for $t = 0$ to the solutions of the original problem at $t = 1$.

The straightforward application of this geometric principle of conservation of number may fail if our special position is too special in the sense that it is a singular point in the parameter space of configurations. This is a very subtle condition that is a property of the special position as a member of the set of problem instances and can not be determined by considering the geometry of the special position itself. Therefore we need to examine the algebraic formulation of the deformation near $t = 0$. As local coordinates we choose $x_{11} = 1$ and $x_{32} = 1$. Then $\det(X|S) = 0 \Leftrightarrow x_{21} x_{42} = 0$. So, at $t = 0$, as either $x_{21} = 0$ or $x_{42} = 0$, the structure of the matrix of partial derivatives

of the system $H(X, t = 0)$ is

$$\begin{bmatrix} x_{42} & x_{21} \\ C_{13}x_{42} - C_{14} & C_{13}x_{21} - C_{23} \end{bmatrix}, \tag{2.4}$$

with C_{ij} the 2-by-2 minor by selecting rows i and j from the matrix representation of L_4. For L_1 in general position, the two start solutions are regular and admit well-defined deformations. For general L_2, this regularity is maintained up to $t = 1$.

3. Pieri-type deformations for hypersurface conditions. In this section we extend the approach of the last section to produce an algorithm for computing all p-planes in $\mathbb{C}^{m+p}$ that intersect mp given m-planes L_i in general position. In order to have a uniform description of the algorithm, we will skip the step of specializing the last two m-planes, and the last step of using linear algebra to intersect with the first m-plane. Instead we will describe the algorithm solely in terms of the continuation step which involved deforming L_2 to S. Later we shall see that adding these linear algebra steps back in amounts to truncating the algorithm as we describe it and does yield a significant computational savings.

We represent p-planes in $\mathbb{C}^{m+p}$ by $(m + p) \times p$-matrices whose columns form a set of generators. Solutions to our problem are represented by variable $(m + p) \times p$-matrices. We move the input data into special positions such that the solution planes have matrix representations with zero coordinates at specific positions. To specify those positions, we need the following definition.

DEFINITION 3.1. *A* localization pattern *is an element of* $\{0, \star\}^{(m+p)\times p}$ *such that all stars in a column are contiguous and the row indices in which the bottommost and topmost stars occur strictly increase as functions of the column number. These row indices are called the* top *and* bottom pivots*, respectively. A p-plane* fits a localization pattern *if it can be represented by a matrix of generators with zero entries everywhere the localization pattern prescribes them.*

Consider, for example, the case where $p = 2, m = 4$ and we are looking for all 2-planes which meet eight 4-planes. We can take the last two planes as $L_7 = \langle \mathbf{e}_1, \mathbf{e}_2, \mathbf{e}_3, \mathbf{e}_4 \rangle$ and $L_8 = \langle \mathbf{e}_3, \mathbf{e}_4, \mathbf{e}_5, \mathbf{e}_6 \rangle$. For any 2-plane that meets both these planes nontrivially we may choose its generators such that one generator lies in L_7 and the other in L_8. Such a 2-plane is represented by a variable matrix

$$\begin{bmatrix} x_{11} & 0 \\ x_{21} & 0 \\ x_{31} & x_{32} \\ x_{41} & x_{42} \\ 0 & x_{52} \\ 0 & x_{62} \end{bmatrix} \text{ which fits the pattern } \begin{bmatrix} \star & 0 \\ \star & 0 \\ \star & \star \\ \star & \star \\ 0 & \star \\ 0 & \star \end{bmatrix}. \tag{3.1}$$

Planes which satisfy this localization pattern already meet L_7 and L_8. Equivalently, any specialization of the variables x_{ij} in the pattern will result in a p-plane which has nontrivial intersection with L_7 and L_8. Thus our problem is to find values of these variables such that the plane represented by (3.1) meets the remaining 4-planes. We achieve this by specializing the input planes as follows.

DEFINITION 3.2. *Given a localization pattern* X*,* the special m-plane for the top (bottom) pivots of X *is the* m*-plane* S_X *spanned by the standard basis vectors not indexed by top (bottom) pivots.*

It turns out that p-planes, which fit a localization pattern X and meet its special m-plane S_X, must fit a set of localization patterns easily derived from X in the following way.

DEFINITION 3.3. *A* top (bottom) child *of a localization pattern X is a localization pattern obtained by turning a topmost (bottommost) star of X to a zero.*

For the bottom localization pattern in Figure 3.1 having bottom pivots [3 5], the special 4-plane is $\langle \mathbf{e}_1, \mathbf{e}_2, \mathbf{e}_4, \mathbf{e}_6 \rangle$. Lemma 3.4 states that all 2-planes meeting this special 4-plane and fitting the bottom pattern must fit one of its two children.

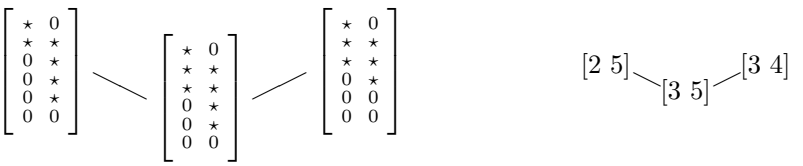


FIG. 3.1. *A localization pattern with bottom pivots* [3 5] *and its two children* [2 5] *and* [3 4]. *At the right we see the bottom pivots used as a shorthand notation.*

LEMMA 3.4 (hypersurface Pieri rule). *Let S_X be the special m-plane for top (bottom) pivots of the localization pattern X. Every p-plane that fits X and meets S_X also fits one of the (at most p) children of X. Conversely every p-plane that fits a child of X fits X and meets S_X.*

Proof. A p-plane that fits X meets S_X if and only if, for all specific values for the variables corresponding to the stars in the pattern X, we can find a linear combination of the columns of the p-plane which lies in S_X. This means that the intersection of the p-plane with S_X can be represented as a linear combination of the columns of S_X. Finding this linear combination is equivalent to finding a nontrivial element of the kernel of $[X|S_X]$. Since the m columns of S_X have pivots distinct from those of the p columns of X, the matrix $[X|S_X]$ is an $(m+p) \times (m+p)$-matrix with $m+p$ distinct pivot rows. So we can rearrange $[X|S_X]$ into a triangular matrix, which is singular precisely when one of the diagonal elements coming from X is specialized to zero. ∎

The repeated application of Lemma 3.4 builds up a poset diagram which allows us to count the solution planes as shown in Figure 3.2. The poset starts at the trivial localization pattern X_0 whose ith column has its top pivot in row i and its bottom pivot in row $m+i$. While there are p-planes that do not fit X_0, the set of p-planes that do fit X_0 includes those for which all Plücker coordinates are nonzero. This is a dense open subset of the Grassmannian, and with the general position assumption on our input planes, it will suffice to count the number of solution planes which fit X_0. This start pattern, X_0 has exactly $mp+p$ stars, and our original list of conditions contains exactly mp given m-planes. Initially we apply Theorem 3.5 starting at $X = X_0$ and $n = mp$ in the following.

THEOREM 3.5. *Consider a localization pattern X with $p+n$ stars. For $n = 0$, X counts for one solution. For $n > 0$, the number of p-planes fitting X and meeting n general m-planes equals the sum of the number of solution planes fitting the children of X and meeting $n-1$ general m-planes.*

Theorem 3.5 is proven by induction and follows from Lemma 3.7 establishing the correctness of the Pieri homotopy. Applying the Pieri homotopy algorithm following the root counting procedure through the poset yields an effective algorithm for actually finding the solutions.

DEFINITION 3.6. *Let X be a localization pattern and S_X the special m-plane*

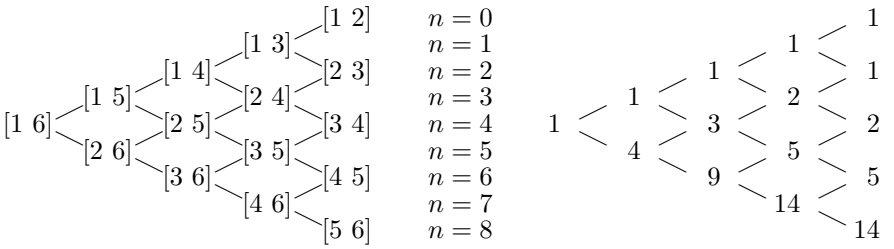


FIG. 3.2. *Combinatorial root count for $p = 2$, $m = 4$. The brackets at the left are the bottom pivots. The top pivots remain* [1 2]. *The trivial localization pattern X_0 corresponds to* [5 6]. *The roots are counted at the right, starting at the top and adding up the numbers at the leaves while moving down to the root of the poset yielding* 14 *solutions.*

for top (bottom) pivots of X. Suppose that $n-1$ general intersection conditions are satisfied for children Y of X, i.e., $\det(Y|L_i) = 0$ *for $i = 1, 2, \ldots, n-1$. To satisfy the nth intersection condition, the* Pieri homotopy $H(X,t) = \mathbf{0}$ *is*

$$H(X,t) = \begin{cases} \det(X|(1-t)S_X + tL_n) = 0, & \\ \det(X|L_i) = 0, \quad i = 1, 2, \ldots, n-1, & t \in [0,1]. \end{cases} \tag{3.2}$$

To illustrate the Pieri homotopy, we consider the situation as in going down to [3 5], with the three 2-planes that fit the left child [2 5] and two 2-planes that fit the right child [3 4]. Locally the situation is pictured in Figure 3.1, with the global root counting procedure in Figure 3.2. At $t = 0$, $n = 4$, and we have already folded in the first four intersection conditions and are moving the special 4-plane $\langle \mathbf{e}_1, \mathbf{e}_2, \mathbf{e}_4, \mathbf{e}_6 \rangle$ to the fifth input plane. In general we have the situation as in the following Lemma.

LEMMA 3.7. *Consider a localization pattern X with $p+n$ stars and n complex m-planes L_i in general position. Suppose we are given all p-planes that meet L_i, $i = 1, 2, \ldots, n-1$, and fit one of the children of X. Then the Pieri homotopy defines regular paths of p-planes that start at the given p-planes and end at those p-planes which meet all n general m-planes L_i while fitting X.*

Proof. Our working space is a product of projective spaces, with as many spaces as there are columns in X. To describe this multiprojective space more precisely, we count the stars in X. For X defined by p-tuples of top and bottom pivots, respectively, $\alpha, \beta \in \mathbb{N}^p$, we embed the p-planes into $\mathbb{P}^{d_1} \times \mathbb{P}^{d_2} \times \cdots \times \mathbb{P}^{d_p}$ with $d_i = \beta_i - \alpha_i$, $i = 1, 2, \ldots, p$. To fix an affine coordinate chart we set one coordinate to 1 in every column by scaling the corresponding generator of the p-plane.

Essentially we are applying a multihomogeneous homotopy [23, 24] in a general situation. By the assumption of Lemma 3.7, all p-planes at $t = 0$ are regular solutions, as the m-planes L_i are in general position. This general position is maintained for all $t \in [0,1]$, ensuring the smoothness of the solution paths.

We still have to show that we will find all p-planes. Suppose that at $t = 1$ there are more solutions than the number of paths we started with. Going backward with those additional solutions from $t = 1$ to $t = 0$ we move either to a singular solution or to a solution at infinity. But at $t = 0$, all solutions are regular. By Lemma 3.4, all solutions have been found and since the localization patterns admit any affine chart, there are no solutions at infinity for $t = 0$. ∎

We remark that the root count in Figure 3.2 obtained by changing only bottom pivots is just one possibility. One could alternatingly change bottom and top pivots

in building up the levels of the poset. As was done in section 2, linear algebra can be applied to intersect with the first input plane. Another, more significant optimization includes the special position of the last two input planes. Choosing a basis so that the last two m-planes are already in special position $L_{mp-1} = \langle \mathbf{e}_1, \ldots, \mathbf{e}_m \rangle$ and $L_{mp} = \langle \mathbf{e}_{p+1}, \ldots, \mathbf{e}_{m+p} \rangle$ amounts to enforcing a different starting pattern for the poset. This has two advantages. The first is that it really forces any pattern meeting all original planes to meet the starting localization pattern regardless of the genericity of the remaining input planes. The second is that it saves the work of doing the two final and most costly (because of highest dimension) continuation steps.

In contrast to the approach of [16], our description treats the hypersurface Pieri homotopy algorithm separately from the case of general intersection conditions. This allows us to simplify the definition of special m-planes. However, when all input m-planes are spanned by real matrices, we need the more general choice of special m-planes as defined in [16] to enforce complex arithmetic and avoid quadratic turning points along the solution paths. Also for the general intersection conditions we need the definition of [16] for special m-planes. So in Definition 3.2 we replace each of the standard basis vectors $\mathbf{e}_i$ by vectors $\mathbf{b}_i$ that have random complex entries in rows i and higher (lower) for bottom (top) pivots i. We assume this adaptation of Definition 3.2 for the rest of the paper.

4. General intersection conditions. In the general case we are looking for p-planes in $\mathbb{C}^{m+p}$ which nontrivially meet a list of generic subspaces $L_1, L_2, \ldots, L_n$, where each L_i has dimension $m+1-k_i$. It turns out each of these conditions removes k_i degrees of freedom, and thus we will expect a finite number of solutions when $k_1 + k_2 + \cdots + k_n = mp$.

In specializing the ith $(m+1-k_i)$-plane L_i we move it to the intersection of k_i special m-planes. This deformation of the solution p-planes while moving L_i from general to special position proceeds in k_i steps, each time specializing some of the coordinates of our current representation for the solution planes to zero. In modeling this deformation we need to be aware of the left and right sides of our localization patterns when working, respectively, with bottom and top pivots.

DEFINITION 4.1. *For the localization patterns X^1 and X^2 we write $X^1 \supset_{c_1} X^2$ to indicate that X^2 is the child of X^1 obtained by turning the pivot in column c_1 of X^1 to zero. We say that the* chain of length j

$$C = X^1 \supset_{c_1} X^2 \supset_{c_2} \cdots \supset_{c_{j-1}} X^j \tag{4.1}$$

is bottom-left *if $c_i \leq c_{i-1}$, $i = 2, \ldots, j$, for bottom pivots c_i, or* top-right *if $c_i \geq c_{i-1}$, $i = 2, \ldots, j$, for top pivots c_i. If C is bottom-left (resp., top-right), we will call any child X^{j+1} of X^j a* bottom-left child *(resp.,* top-right child*) of C if the chain $X^1 \supset_{c_1} \cdots \supset_{c_{j-1}} X^j \supset_{c_j} X^{j+1}$ is also bottom-left (resp., top-right).*

In other words, in a bottom-left (top-right) chain, we never take away a bottom (top) pivot in a column to the right (left) of one we have already taken. Note that when $C = X^1$ is a chain of length 1, we have not taken any pivots away yet and all children of X^1 are bottom-left (top-right) children of C.

Notice that the pivots of $X^1 \supset_{c_1} X^2$ are the same except in column c_1 and that the new pivot of X^2 cannot have been a pivot of X^1 (otherwise X^2 would not be a localization pattern). Thus $S_{X^1} \cap S_{X^2}$ has dimension $m-1$. Similarly we can see that the intersection $S_{X^1} \cap S_{X^2}$ with S_{X^3} drops the dimension by 1 whenever X^3 is a child of X^2 obtained by deleting a pivot in a column with index less than or equal to c_1. This justifies the dimension of the following definition.

DEFINITION 4.2. *For a bottom-left (top-right) chain* $C = X^1 \supset_{c_1} X^2 \supset_{c_2} \cdots \supset_{c_{j-1}} X^j$ *and an index* $1 \le k \le j$ *we define a* special $(m+1-k)$-plane S_C^k *to be* $S_C^k = \cap_{i=1}^k S_{X^i}$, *where the* S_{X^i} *are the special* m*-planes associated to the* X^i *in* C.

From now on we assume that we work with bottom pivots.

LEMMA 4.3 (general Pieri rule). *Let* C *be a bottom-left chain of length* j *ending at* X^j.

(1) *All* p*-planes fitting* X^j *and meeting* S_C^j *must fit at least one of* X^j*'s children.*
(2) *A child* X^{j+1} *meets* S_C^j *if and only if* X^{j+1} *is a bottom-left child of* C, *i.e., if and only if* $X^j \supset_{c_j} X^{j+1}$, *where* $c_{j+1} \le c_j$.

Proof. Any p-plane which fits X^j and meets S_C^j must also meet S_{X^j} and thus must be a child of X^j by Lemma 3.4. This proves the first claim.

Recall from the proof of Lemma 3.4 that X^{j+1} meets S_C^j if and only if, given any specification of the variables of X^j for which the pivot elements are nonzero, it is possible to find a linear combination of the columns which lies in S_C^j. Clearly any such linear combination must involve the decremented column c_{j+1} and no columns after it, since all pivots of vectors in S_C^j must be nonpivots of X^j.

If $c_{j+1} \le c_j$, then the first $c_j - 1$ columns of X^{j+1} have remained unchanged all the way through the chain, and the last has only had successive pivots removed. Thus S_C^j must contain a basis vector with a pivot row in each nonpivot of the first c_{j+1} columns of X^j. So if we take these basis vectors along with the first c_{j+1} columns of X^j we get a set of r_j column vectors with r_j distinct pivots all in rows r_j. Since X^{j+1} is obtained from X^j by deleting the pivot of column c_{j+1} we find that the (c_{j+1})th column must be in the span of the remaining columns, and this linear relation can be rewritten to express a vector of S_C^j as a linear combination of columns of X^{j+1}.

On the other hand, if $c_{j+1} > c_j$, then at least one nonpivot of these first c_{j+1} columns of X^j must have been a pivot previously and S_C^j must have no vector with a pivot in this row. In other words we are looking for a solution to a subsystem of the previous system where we have removed some number of columns (aside from that corresponding to column c_{j+1} of X^{j+1}). Looking at just the pivot rows appearing we get either a square triangular system with nonzeros on the diagonal (if the pivot of column c_{j+1} did not appear) or a nonsquare system with nonzero pivots in each column. In either case no nontrivial solution is possible. □

THEOREM 4.4. *Let* X^1 *be a localization pattern with* $p + \sum_{i=1}^{n-1} k_i$ *stars where the* k_i *are the codimension conditions specified in the problem (i.e.,* $\dim(L_i) = m+1-k_i$). *Further suppose that* $C = X^1 \supset_{c_1} X^2 \supset_{c_2} \cdots \supset_{c_{j-1}} X^j$ $(1 \le j \le k_n)$ *is a bottom-left chain starting at* X_1. *For* $1 \le k \le j$ *let* U_C^k *be a generic* $(m+1-k_n)$*-dimensional subspace of* $S_C^k = \cap_{i=1}^k S_{X^i}$. *Then the number of* p*-planes meeting* U_C^{j-1} *as well as* $L_1, \ldots, L_{n-1}$ *and fitting* X^j *is equal to the sum of the numbers of* p*-planes meeting* $U_C^j, L_1, \ldots, L_{n-1}$ *and meeting* Y *as* Y *runs over all bottom-left children of* C.

As with the hypersurface case, Theorem 4.4 is proved by induction and follows from the correctness of a homotopy. Before describing this homotopy we will show that the theorem provides a root count procedure generalizing that of the last section. Combining the homotopy with this counting procedure will yield an effective technique for producing all solutions. It is a little easier to understand the implications of the theorem for root counting if we restate it without reference to the intermediate steps where our input planes are partially specialized.

COROLLARY 4.5. *Suppose* X *is a localization pattern with* $p + \sum_{i=1}^{n} k_i$ *stars where the* k_i *are codimension conditions specified in the problem (i.e.,* $\dim(L_i) = m+1-k_i$). *Then the number of* p*-planes fitting* X *and meeting* $L_1, \ldots, L_n$ *is equal to the total*

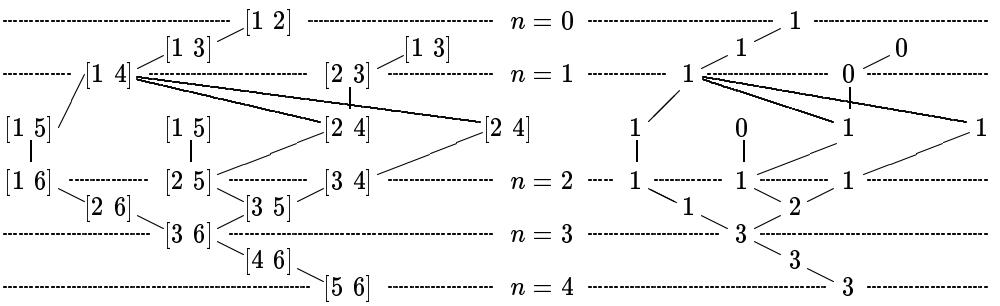


FIG. 4.1. *Bottom pivots for $p = 2$, $m = 4$, $k_i = 2$, $i = 1, 2, 3, 4$. Only those arcs and nodes are drawn that connect localization patterns reachable by bottom-left chains. The corresponding root count is at the right.*

number of p-planes meeting $L_1, \ldots, L_{n-1}$ and fitting at least one descendant Y of X accessible as a bottom-left child of a bottom-left chain of length k_n starting at X.

Proof. To see that the corollary follows from Theorem 4.4, consider applying the theorem k_n times starting from the hypothesis of Corollary 4.5. Note that the first application of the theorem leaves us with the problem of counting the number of p-planes meeting $L_1, \ldots, L_{n-1}$, and also meeting an $m + 1 - k_n$ plane lying in S_X and fitting a child of X. Each of these numbers can now be counted by applying the theorem to the new case with $n = n - 1$, $j = 2$, and $C = C^1 \supset C^2$. This process continues until all our chains have length k_n and the subspace $U_C^{j+1} = S_C^{k_n}$ in the conclusion of Theorem 4.4 is a special $m + 1 - k_n$ space and that Lemma 4.3 implies that any p-plane fitting a bottom-left child of C will meet S_C^j. ∎

Figure 4.1 illustrates the combinatorial counting method arising from Theorem 4.4 just as Figure 3.2 did for the hypersurface case. Starting at the root of the poset we apply the theorem to the case $n = 4$, $j = 1$, $C = [5\ 6]$ to see that the number of 2-planes which fit the pattern [5 6] and meet 4 general 3-planes L_1, L_2, L_3, L_4 is equal to the number of p-planes which fit [4 6], meet the 3 general 3-planes L_1, L_2, L_3, and meet a general 3-plane lying inside the special 4-plane $S_{[5\ 6]}$. A second application of the theorem in the case $n = 3$, $j = 2$, $C = [5\ 6] \supset_1 [4\ 6]$ now tells us that this number is equal to the number of p-planes fitting the pattern [3 6] and meeting the 3 general 3-planes L_1, L_2, L_3. Note that by Lemma 4.3 we do not get any solutions fitting the pattern [4 5] (so we omitted [4 5] in Figure 4.1). Continuing, with $n = 3$, $j = 1$, $C = [3\ 6]$, we find that the number of 2-planes fitting the pattern [3 6] and meeting L_1, L_2, L_3 is equal to the sum of the numbers fitting the two children of [3 6] and meeting L_1, L_2, and a general 3-plane lying inside of $S_{[3\ 6]}$. Applying the theorem with $n = 3$, $j = 2$, to the chains $[3\ 6] \supset_1 [2\ 6]$ and $[3\ 6] \supset_2 [3\ 5]$ we find that the total number of 2-planes fitting the pattern defined by [3 6] and meeting L_1, L_2, and L_3 is equal to the sum of the numbers of 2-planes meeting L_1 and L_2 and fitting any of the localization patterns defined by [1 6], [2 5], and [3 4].

It is important to note that during the intermediate steps when the chains C have length $j > 1$, the number of solutions at a node depends on both the localization pattern and the chain. Thus we can only share nodes corresponding to the $j = 1$ case, i.e., those nodes with no partially specified input planes. In Figure 4.1 these cases (with $j = 1$) are indicated by dotted horizontal lines. Also note that some chains may not make it k_n steps and may simply stop and fail to yield any solutions. Both these points are illustrated by considering the localization pattern [1 5] in Figure 4.1.

Applying Theorem 4.4 with $n = 2$, $j = 2$, on the chains [1 6] $\supset_2$ [1 5] yields an equivalence with the numbers of solutions fitting [1 4] and meeting L_1, while applying the theorem to with $n = 2$, $j = 2$, and the chain [2 5] $\supset_1$ [1 5] yields no solutions at all since this chain has no bottom-left children. We are now ready to define the homotopy which provides our effective proof of Theorem 4.4.

DEFINITION 4.6. *Let X, n, j, and C be as in Theorem* 4.4*, and suppose that the subspaces U_C^j and U_C^{j-1} are represented by $(m+p)\times(m+1-k_n)$-dimensional matrices made up of random linear combinations of the S_C^j and S_C^{j-1} (except that $U_C^0 = L_n$). We define the* Pieri homotopy *$H_C^j(X,t) = \mathbf{0}$ to be*

$$H_C^j(X,t) = \begin{cases} \text{all maximal minors of } [X|(1-t)U_C^j + tU_C^{j-1}], & \\ \text{all maximal minors of } [X|L_i] = 0, & t \in [0,1]. \\ \qquad\qquad\qquad\qquad i = 1,2,\ldots,n-1, & \end{cases} \tag{4.2}$$

Note that while this homotopy may well have more equations than unknowns it will turn out to define a nonsingular path of solutions through each starting point. Here we will prove that these equations define a homotopy which can be used to find all solutions. We will defer a discussion of techniques for actually following these paths numerically to section six.

LEMMA 4.7. *Supposing that the solutions of $H_C^j(X,0) = \mathbf{0}$ are all nonsingular as X varies over all bottom-left extensions of C then the Pieri homotopy $H_C^j(X,t) = \mathbf{0}$ defines regular paths through these solutions which lead to all solutions of $H_C^j(X,1)$.*

Proof. As in the proof of Theorem 3.5, our working space is the product $\mathbb{P}^{d_1}\times\mathbb{P}^{d_2}\times\cdots\times\mathbb{P}^{d_p}$ of projective spaces where d_j is the number of stars in each column of X, and we are essentially applying a multihomogeneous homotopy [23, 24]. By assumption all p-plane solutions at $t = 0$ are regular, and since U^j defines an $(m+1-k_n)$-dimensional subspace of S_C^{j-1} we see that U^{j-1}, which defines a generic $(m+1-k_n)$-dimensional subspace of S_C^{j-1}, must also define an intersection problem with a solution set consisting of a finite number of regular p-planes. This general position is maintained for all $t \in [0,1]$, ensuring the smoothness of the solution paths.

It remains to show that we will find all p-planes. This argument is exactly analogous to the proof of Theorem 3.5 except that we rely on Lemma 4.3 to show that all solutions found must fit some bottom-left child of C when we go backwards. □

Theorem 4.4 follows from the lemma by induction. We start at the tops of the counting poset where the localization patterns all have exactly p stars and no subspace conditions to meet and hence define unique solutions. The hypotheses of the homotopy lemma are then met and this forces the conditions to be met all the way down.

5. Control of linear systems and quantum Schubert calculus. This section is organized into three parts. We first rephrase the dynamic pole placement problem into a problem of enumerative geometry [31]. Thereafter we derive the critical dimension [47] for this problem to have a finite number of solutions. Lastly we state the root-counting theorem [29] and prove it inductively from the correctness of the Pieri homotopies.

Suppose we want to control a plant with n internal states $\mathbf{x} \in \mathbb{R}^n$ that takes m-inputs $\mathbf{u} \in \mathbb{R}^m$ and produces p-outputs $\mathbf{y} \in \mathbb{R}^p$ with a dynamic compensator that has q internal states $\mathbf{z} \in \mathbb{R}^q$. We picture this situation schematically in Figure 5.1.

The evolution in time of the plant is described by a system of first-order differential

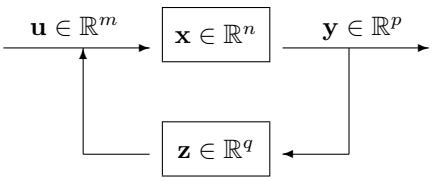


FIG. 5.1. *Control of an m-input and p-output plant by a qth-order dynamic compensator.*

equations:

$$\begin{cases} \dot{\mathbf{x}} = A\mathbf{x} + B\mathbf{u} \\ \mathbf{y} = C\mathbf{x} \end{cases} \quad \begin{array}{l} \text{with } \mathbf{x} \in \mathbb{R}^n, \mathbf{u} \in \mathbb{R}^m, \mathbf{y} \in \mathbb{R}^p, \\ \text{and } A \in \mathbb{R}^{n \times n}, B \in \mathbb{R}^{n \times m}, C \in \mathbb{R}^{p \times n}. \end{array} \tag{5.1}$$

The dynamic compensator obeys a qth-order differential equation, described by the system

$$\begin{cases} \dot{\mathbf{z}} = F\mathbf{z} + G\mathbf{y} \\ \mathbf{u} = H\mathbf{z} + K\mathbf{y} \end{cases} \quad \text{with} \quad \mathbf{z} \in \mathbb{R}^q \quad \text{and} \quad \begin{array}{l} F \in \mathbb{R}^{q \times q}, G \in \mathbb{R}^{q \times p}, \\ H \in \mathbb{R}^{m \times q}, K \in \mathbb{R}^{m \times p}. \end{array} \tag{5.2}$$

After elimination of $\mathbf{u}$ and $\mathbf{y}$, concatenation of (5.1) and (5.2) yields the following closed-loop system:

$$\begin{bmatrix} \dot{\mathbf{x}} \\ \dot{\mathbf{z}} \end{bmatrix} = \begin{bmatrix} A + BKC & BH \\ GC & F \end{bmatrix} \begin{bmatrix} \mathbf{x} \\ \mathbf{z} \end{bmatrix}. \tag{5.3}$$

The behavior of this closed-loop system is determined by the $n+q$ eigenvalues of the matrix in (5.3). For a plant given by the matrix triplet (A, B, C) and $n+q$ eigenvalues, the dynamic pole placement problem asks for the matrix quadruples (F, G, H, K) which determine the dynamic compensators that yield closed-loop systems with a specific set of eigenvalues.

We can formulate the dynamic pole placement problem as a geometric problem. In rewriting the characteristic equation of (5.3) the subscripts in I denote the dimension of the identity matrix. There is only some hindsight involved in deriving (5.6). The rest of the equivalences are deduced by elementary row and column operations.

$$\det\left(s \begin{bmatrix} I_n & 0 \\ 0 & I_q \end{bmatrix} - \begin{bmatrix} A + BKC & BH \\ GC & F \end{bmatrix} \right) = 0, \tag{5.4}$$

$$\Leftrightarrow \det \begin{bmatrix} sI_n - A - BKC & -BH \\ -GC & sI_q - F \end{bmatrix} = 0, \tag{5.5}$$

$$\Leftrightarrow \det \begin{bmatrix} sI_n - A - BKC & -BH & BK & -B \\ -GC & sI_q - F & G & 0 \\ 0 & 0 & I_p & 0 \\ 0 & 0 & 0 & I_m \end{bmatrix} = 0, \tag{5.6}$$

$$\Leftrightarrow \det \begin{bmatrix} sI_n - A & 0 & 0 & -B \\ 0 & sI_q - F & G & 0 \\ C & 0 & I_p & 0 \\ 0 & -H & K & I_m \end{bmatrix} = 0, \tag{5.7}$$

$$\Leftrightarrow \det \begin{bmatrix} I_n & 0 & 0 & -(sI_n - A)^{-1}B \\ 0 & I_q & (sI_q - F)^{-1}G & 0 \\ C & 0 & I_p & 0 \\ 0 & -H & K & I_m \end{bmatrix} = 0, \tag{5.8}$$

$$\Leftrightarrow \det \begin{bmatrix} I_n & 0 & 0 & -(sI_n - A)^{-1}B \\ 0 & I_q & (sI_q - F)^{-1}G & 0 \\ 0 & 0 & I_p & C(sI_n - A)^{-1}B \\ 0 & 0 & H(sI_q - F)^{-1}G + K & I_m \end{bmatrix} = 0. \tag{5.9}$$

The last $m+p$ rows and columns of (5.9) represent the intersection of m-planes defined by the given triplet (A, B, C) with p-planes determined by the unknown quadruple (F, G, H, K). As these p-planes depend on the variable s introduced by the term $(sI_q - F)^{-1}$, they have maximal minors of degree q and we call them degree q-maps. Thus the dynamic pole placement problem is equivalent to the computation of all degree q-maps into the Grassmannian of p-planes that meet n given m-planes at prescribed s-values. It is easy to see that for each specification of an eigenvalue λ_i the condition that the characteristic polynomial (5.9) vanish at $s = \lambda_i$ enforces one polynomial condition on the set of degree q-maps.

Note that when $q = 0$ we are solving the static pole placement problem ($\mathbf{u} = K\mathbf{y}$ and $F, G, H = 0$) and are looking for maps of degree 0 (i.e., constant maps) which meet a specific set of given m-planes. In this case the characteristic equation (5.9) has degree n and we can find solution planes whenever n is less than the dimension mp of the space of p-planes in $(m + p)$-dimensional space. At the critical dimension $n = mp$, we expect the number of such solutions to be finite. In other words when $q = 0$ and $n = mp$ this analysis transforms the static pole placement problem into the problem of finding all p-planes meeting a set of mp given m-planes as dealt with in section 3. For general q we will be able to use similar methods to solve it, but will need the following lemma which specifies the critical dimension and allows us to produce localization patterns for our maps.

LEMMA 5.1. *Let $q = dp + r$ with $d, r \in \mathbb{N}$ and $r < p$. The set of $(m + p) \times p$ degree q-maps is of dimension $mp+q(m+p)$ and has a dense open subset which can be represented by matrices with entries of maximum degree d in all but the last r columns and degree $d + 1$ in the remaining r columns.*

Proof. Given a polynomial matrix we will define the degree of a column to be the maximum degree of any of its entries. Also note that two $(m + p) \times p$ polynomial matrices define the same map if there is an invertible $p \times p$-matrix $U(s)$ which takes one into the other by right multiplication. Lemma 3.1 from [47] tells us that for any $(m + p) \times p$ degree q-map $X'(s)$, there is a unique choice of column degrees $d_1 \leq \cdots \leq d_p$ (with $q = d_1 + \cdots + d_p$) such that the $X'(s)$ is equivalent to a matrix $X(s)$ with these column degrees. In other words the sets $U_{d_1,\ldots,d_p}$ of degree q-maps which can be represented by $(m + p) \times p$-matrices with column degrees $d_1, \ldots, d_p$ disjointly cover the set of degree q-maps.

In order to find the dimension of $U_{d_1,\ldots,d_p}$ we will want to find a canonical form for a general member. By a general member of $U_{d_1,\ldots,d_p}$ we mean a matrix $X(s)$ whose entries are polynomials of appropriate degrees with no algebraic relations among the coefficients. If we write $a^k_{i,j}$ for the coefficient of s^k in the (i, j)th entry of $X(s)$, we can express our normal form by requiring $a^0_{i,i} = 1$ for $i = 1, \ldots, p$, and if for all $i < j$ we have $a^k_{i,j} = 0$ if $k \leq d_j - d_i$ and $a^0_{j,i} = 0$ if $d_j = d_i$. This normal form is illustrated in Figure 5.2 and can be achieved by the following algorithm:

$$\begin{bmatrix} 1 & 0 & 0 & 0+0s & 0+0s+0s^2 \\ 0 & 1 & 0 & 0+0s & 0+0s+0s^2 \\ 0 & 0 & 1 & 0+0s & 0+0s+0s^2 \\ \star & \star & \star & 1+\star s & 0+0s+\star s^2 \\ \star & \star & \star & \star+\star s & 1+\star s+\star s^2 \\ \star & \star & \star & \star+\star s & \star+\star s+\star s^2 \\ \star & \star & \star & \star+\star s & \star+\star s+\star s^2 \end{bmatrix} \qquad \begin{bmatrix} 1 & 0 & 0+0s & 0+0s & 0+0s \\ 0 & 1 & 0+0s & 0+0s & 0+0s \\ \star & \star & 1+\star s & 0+\star s & 0+\star s \\ \star & \star & 0+\star s & 1+\star s & 0+\star s \\ \star & \star & 0+\star s & 0+\star s & 1+\star s \\ \star & \star & \star+\star s & \star+\star s & \star+\star s \\ \star & \star & \star+\star s & \star+\star s & \star+\star s \end{bmatrix}$$

$U_{0,0,0,1,2}$ $U_{0,0,1,1,1}$

FIG. 5.2. *Standard forms for $m = 2$, $p = 5$, and $q = 3$ with 28 and 31 stars, respectively.*

for $i = 1, 2, \ldots, p$ do
  $X_i := \frac{1}{a^0_{i,i}} X_i$;
  for $j = i+1, i+2, \ldots, p$ do
    if $d_i = d_j$, then $X_i := X_i - \frac{a^0_{j,i}}{a^0_{j,j}} X_j$; *zero out* $a^0_{j,i}$;
    for $k = 0, 1, \ldots, d_j - d_i$ do $X_j := X_j - \frac{a^k_{i,j}}{a^0_{i,i}} s^k X_i$; *zero out* $a^k_{i,j}$.

The genericity of $X(s)$ ensures that the replacements in this algorithm each introduce no zeros besides the desired one in the coefficient $a^k_{i,j}$ so that the algorithm is well defined and will terminate having introduced exactly the zeros required for the canonical form. It is easy to check that each of the replacements appearing in the algorithm is equivalent to the application of an invertible transformation. Thus the result defines the same map as the input. Also note that replacing any column of a matrix in this form by a nontrivial linear combination of the other columns yields a matrix which is no longer in this form. Thus any two maps in this form are equivalent if and only if they have the same specifications for the remaining variables, i.e., the result provides an affine chart defined on a Zariski dense open subset of $U_{d_1,\ldots,d_p}$.

Since each column of degree d_i polynomials has $(d_i+1)(m+p)$ coefficients, the total number of parameters we started with was $(p+q)(m+p) = mp+q(m+p)+p^2$. Finally, the algorithm specifies the p diagonal constant coefficients to 1 and introduces at least two zeros for each ordered pair $i < j$. In fact the number of zeros introduced is 2 if $d_i = d_j$ or $1 + d_j - d_i$ otherwise. Rewriting this last number as $2 + (d_j - d_i - 1)$ and adding up over all $\frac{p(p-1)}{2}$ ordered pairs $1 \le i < j \le p$ we find that the number of parameters specified is $p^2 + \sum_{i<j} \max(d_j - d_i - 1, 0)$. The conclusion of the lemma now follows since the choice of column degrees specified is the only way to choose the d_i's so that they add up to q and no two differ from each other by more than 1. ∎

The normal form used above was easy to define for general $U_{d_1,\ldots,d_p}$ and was used in [47] where they use the $U_{d_1,\ldots,d_p}$ as cells in a cellation of the space of degree q-maps. For our purposes we need only work in the largest of these cells $U_{d,\ldots,d,d+1,\ldots,d+1}$. Since this cell has dimension $mp+q(m+p)$ and each root of the degree $n+q$ characteristic equation (5.9) enforces one polynomial condition on the space of degree q-maps, we will expect no solutions whenever n is greater than the critical degree $mp+q(m+p)-q$ and a finite number of solutions when $n = mp+q(m+p)-q$. In both cases we expect no solutions of a generic dynamic pole placement problem to lie on these lower-dimensional pieces. In order to work on $U_{d,\ldots,d,d+1\ldots,d+1}$, however, we will want to use a different normal form analogous to the form used in section 3. To describe it we start by identifying $(m+p)$-vectors of degree d polynomials with their $(m+p)(d+1)$-element

coefficient vectors as follows:

$$\begin{bmatrix} a_1^0 + a_1^1 s + \cdots + a_1^d s^d \\ a_2^0 + a_2^1 s + \cdots + a_2^d s^d \\ \vdots \\ a_{m+p}^0 + a_{m+p}^1 s + \cdots + a_{m+p}^d s^d \end{bmatrix} \Leftrightarrow \begin{bmatrix} a_1^0 \\ a_2^0 \\ \vdots \\ a_{m+p}^0 \\ a_1^1 \\ \vdots \\ a_{m+p}^1 \\ \vdots \\ a_1^d \\ \vdots \\ a_{m+p}^d \end{bmatrix}; \tag{5.10}$$

i.e., the coefficient a_j^k of s^k in polynomial entry in row j in the representation on the left is stored in the $((k+1)(m+p)+j)$th position in the representation on the right.

DEFINITION 5.2. *Let $q = dp + r$ with $d, r \in \mathbb{N}$ and $r < p$. A* localization pattern *for $(m+p) \times p$-maps of degree q is given by a table of p columns of zeros and stars, where the first $p-r$ columns have dimension $(d+1)(m+p)$ and the remaining columns have dimensions $(d+2)$. As in the degree 0 case, all stars within a column should be contiguous and the row indices in which the bottommost and topmost stars occur strictly increase as a function of the column index. These row indices are called the* top *and* bottom *pivots, respectively. It is further required that no two top (bottom) pivots differ by $m+p$ or more. We say that a map* fits *a localization pattern if it is equivalent to a map given by a matrix of polynomials with the column degrees specified by the pattern and zero coefficients everywhere the pattern requires them.*

The dimension of the space of degree q-maps which meet a localization pattern X with n stars is $n-p$. To see this note the requirement that no pivots differ by $m+p$ places ensures that no multiple of any column by a power of s has a bottommost star in a pivot row of any other column. This implies that any linear combination involving more than one column has the lowest bottommost pivots and highest topmost pivots of the columns used. Thus only by scaling each column can we change the entries of a generic map fitting a localization pattern without producing a map which no longer fits that pattern. On the other hand we may freely scale each column to set one nonzero entry to 1. This also tells us that patterns with two equal pivots or pivots that differ by $m+p$ or more will be fit by a space of dimension lower than $n-p$.

Given appropriate generality we can use invertible column operations to transform the normal form for $U_{d,\ldots,d,d+1,\ldots,d+1}$ into a localization pattern as follows. First we allow each column to be arbitrarily scaled and thus turn all ones into stars. Next, we use column operations with constant scalars on the first block of $p-r$ dimension $(d+1)(m+p)$ columns to move all the zeros below the diagonal to the bottom of the columns. We then use multiples of these columns by s to introduce $p-r$ rows of zeros to the tops of the block or r $(d+2)(m+p)$-dimensional columns. Finally, we can use scalar column operations on this block to move the below diagonal elements of the normal form to the bottoms of the columns. This process is illustrated in Figure 5.3, and by counting diagonally in each block we see that the resulting localization pattern has top pivots $[1, 2, \ldots, p]$ and bottom pivots $[d(m+p)+m+r+1, \ldots, d(m+p)+m+p, (d+1)(m+p)+m+1, \ldots, (d+1)(m+p)+m+1+r]$. We thus see that

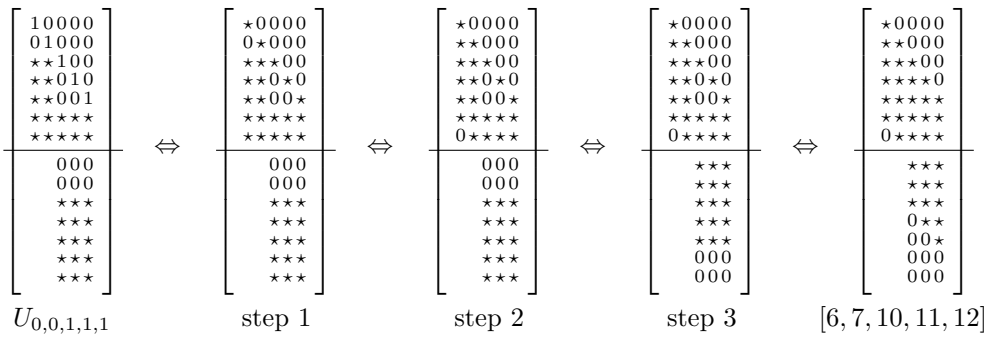


FIG. 5.3. *Transforming standard form for $m = 2$, $p = 5$, and $q = 3$ to top level localization pattern.*

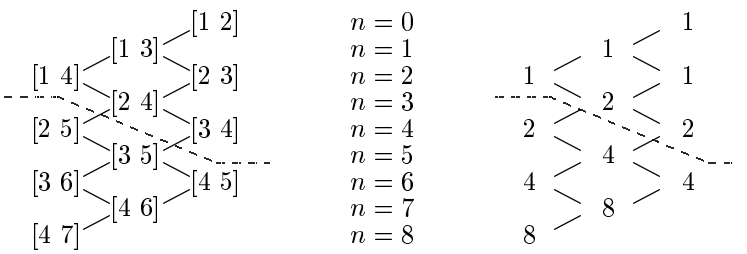


FIG. 5.4. *Combinatorial root count for $p = 2 = m$, $q = 1$. The brackets at the left are the bottom pivots. The top pivots remain* [1 2]. *The trivial localization pattern $X_0(s,t)$ corresponds to* [4 7]. *The dashed line marks the transition between $q = 0$ and $q = 1$. The root count is at the right.*

set of maps which fit this localization pattern has dimension $mp + q(m + p)$ and will account for all solutions of the problem of finding degree q-maps meeting $mp+q(m+p)$ generic m-planes at $mp + q(m + p)$ specified s values. In analogy to the approach used in section 3, we will proceed by successively specializing the input planes to yield problems over smaller localization patterns involving fewer conditions.

When specializing the input planes we also must specialize the values for s by fixing the interpolating points. For bottom pivots we focus on the highest-degree coefficients and move s to ∞ to select those coefficients. For top pivots the lowest-degree coefficients are specialized and we move s to 0. To deal with both situations we add an additional variable t to homogenize the polynomials. So we consider localization patterns of homogeneous polynomials and denote those by $X(s,t)$. The special m-planes of the hypersurface Pieri case are recycled when we evaluate $X(s,t)$ at $(s,t) = (1,0)$ for bottom pivots, or at $(s,t) = (0,1)$ for top pivots. The solutions to the problem in special position can be found in the child patterns that differ from $X(s,t)$ in exactly one position. The analogue to Definition 3.3 is immediate. The main root-counting theorem follows from the correctness of the Pieri homotopies. See Figure 5.4 for an example.

THEOREM 5.3. *Consider a localization pattern $X(s,t)$ with $p + n$ stars. For $n = 0$, $X(s,t)$ counts for one solution. For $n > 0$, the number of maps fitting $X(s,t)$ and meeting n general m-planes at n values for (s,t) equals the sum of the number of solution maps fitting the children of $X(s,t)$ and meeting $n-1$ general m-planes at $n-1$ values for (s,t).*

Note that in the definition of the Pieri homotopies we abuse notation in the sense that t denotes both the continuation parameter and the variable added to homogenize

the maps. This abuse is justified for bottom pivots because we move t from 0 to 1 to move the interpolation point from its specific value $(s,t) = (1,0)$ at ∞ to a general value $(s,t) = (s_n, 1)$. For top pivots we swap the roles of s and t.

DEFINITION 5.4. *Let $X(s,t)$ be a localization pattern and S_X be the special m-plane for bottom pivots at $(s,t) = (1,0)$. Suppose that the children $Y(s,t)$ of $X(s,t)$ meet $n-1$ general m-planes L_i at (s_i,t_i), i.e., $\det(Y(s_i,t_i)|L_i) = 0$ for $i = 1,2,\ldots,n-1$. To satisfy the nth intersection condition with L_n at (s_n,t_n), the* Pieri homotopy *is defined as*

$$H(X(s,t),s,t) = \begin{cases} \det(X(s,t)|(1-t)S_X + tL_n) = 0, \\ (s-1)(1-t) + (s-s_n)t = 0, \\ \det(X(s_i,t_i)|L_i) = 0, \\ i = 1,2,\ldots,n-1, \end{cases} \qquad t \in [0,1]. \tag{5.11}$$

For top pivots s and t swap roles and we consider s_n^{-1} as the target value for s. When we remove entries from both top and bottom, two independent s-variables are used. Theorem 5.3 follows from the correctness of the Pieri homotopy, proven in Lemma 5.5.

LEMMA 5.5. *Consider a localization pattern $X(s,t)$ with $p+n$ stars and n complex m-planes L_i in general position at the values (s_i,t_i). Suppose we are given all maps that meet L_i at (s_i,t_i), $i = 1,2,\ldots,n-1$, and fit one of the children of $X(s,t)$. Then the Pieri homotopy defines regular solution paths as t varies from 0 to 1, that start at the given maps and end at those maps that meet all n general m-planes L_i at (s_i,t_i) while fitting $X(s,t)$.*

Proof. Observe that in the Pieri homotopy (5.11) at $(s,t) = (1,0)$ we select in the first equation with the special m-plane S_X in the expansion of $\det(X(1,0)|S_X) = 0$ those stars that are indexed by the bottom pivots. So the first equation in (5.11) looks like

$$H_1(X(s,t),s,t) = p(X(1,0)) + t(f(X(s,t),s,t) + O(t)), \tag{5.12}$$

where $p(X(1,0))$ denotes the product of all pivot stars and where $f(X(s,t),s,t)$ is a polynomial not divisible by t. Ordering the variables as (s,t,x_{ij}), the matrix of all partial derivatives of the homotopy has the following structure:

$$\begin{bmatrix} \frac{\partial H_1}{\partial s} & \frac{\partial H_1}{\partial t} & A \\ 1 & 1-s_n & B \\ 0 & 0 & C \end{bmatrix}. \tag{5.13}$$

Since the $n-1$ planes L_i are in general position the rank of C is $n-1$. Thus the Jacobian matrix (5.13) is of full rank if and only if $\frac{\partial H_1}{\partial s}\frac{\partial H_1}{\partial t} \neq 0$ at $(s,t) = (1,0)$. We can show that $f(X(1,0),1,0) \neq 0$, which implies $\frac{\partial H_1}{\partial t} \neq 0$. If $f(X(1,0),1,0)$ were equal to 0, then $f(X(s,t),s,t)$ would be identically zero and H_1 could not contain any term that is linear in t. This could happen only for very special choices of L_n. Hence at $(s,t) = (1,0)$, all solutions are regular.

To compactify our space we choose again a multihomogenization according to the columns of the matrices. The regularity of the homotopy is maintained as t moves to 1 since L_n is in general position. Suppose we would have more solutions at $(s,t) = (s_n,1)$ than the number of paths we have started with. Going backward, when in the Pieri homotopy we let t go from 1 to 0, we would arrive at solutions at

$(s,t) = (1,0)$ that do not fit any of the child patterns of $X(s,t)$. But any other child pattern that is not provided in the root count does not allow enough freedom to meet the general intersection conditions L_i at (s_i,t_i) for $i = 1,2,\ldots,n-1$. □

We can reformulate the Pieri rule for the general case to deal with general intersection conditions for any q when the codimensions conditions k_i, $i = 1,2,\ldots,n$, satisfy $k_1 + k_2 + \cdots + k_n = mp + q(m+p)$. In the case where $q > 0$ we work with brackets modulo $(m+p)$ to represent the top and bottom pivots.

Consider a chain C ending at the pattern $X(s,t)$. Let α denote the bracket with the indices of the basis vectors in the intersection of the special m-planes for the patterns in the chain C. Let β denote the pivots modulo $(m+p)$ for a child $Y(s,t)$ of $X(s,t)$. Define γ as the bracket obtained from concatenating α and β, sorted in increasing order. Note that γ may contain repeated entries. We say that $Y(s,t)$ is *a valid child of the chain* C if there exists an index i such that

$$|\gamma_i - \gamma_1| < i - 1 \quad \text{for bottom pivots } \beta \tag{5.14}$$

or

$$|\gamma_l - \gamma_i| < l - i \quad \text{for top pivots } \beta, \tag{5.15}$$

where l is the index of the last element in γ. This rule is consistent with the general Pieri rule for $q = 0$ and allows us to compute intersection numbers for any q and for general codimension conditions.

6. Computational experiments and experiences. The Pieri homotopy algorithm is part of the module of PHCpack [44] that also provides the SAGBI homotopies. We compare with those homotopies and give comments on the organization of the code, numerical aspects, and applications. All timings reported concern a 166 MHz Pentium II processor with 64 Mb internal memory running Linux, unless indicated otherwise.

6.1. Implementation issues. In setting up the Pieri homotopy algorithm, we restricted to the case where the input planes are all in general position. In this case, all the deformations are free from multiplicities. To solve a nongeneric, specific instance of the problem instance we first solve a generic case and then apply a cheater's homotopy to obtain solutions to the original problem. The paradigm of cheater's homotopy [20], [21] or coefficient-parameter polynomial continuation [25], [26] makes it straightforward to set up this homotopy. This paradigm ensures that singularities can occur only at the end of the solution paths and only when the original problem has singular solutions.

The general intersection conditions are represented by overdetermined polynomial systems. A direct method to solve an overconstrained polynomial system of N equations in n unknowns is to multiply it by an $n \times N$-matrix of random complex numbers to construct a square system, as proposed in [36]. This reduction to a square system destroys the geometric structure and creates many excess solution paths to follow. In our situation, the overdetermined polynomial systems we have to deal with are the Pieri homotopies for general intersection conditions, which are overdetermined homotopies. From the geometry we know which solution paths we have to follow so that we maintain the optimal number of solution paths; see [16, section 4]. To control the correctness and to detect numerical failures, we evaluate the original overconstrained system at the computed end points of the paths. The complexity of Newton's method and path-following algorithms for overdetermined systems is analyzed in [9].

The geometric intersection conditions treated in section 4 are expressed by determinantal equations, which are a special kind of polynomial system. Just as sparse polynomial systems allow fewer perturbations on the coefficient structure than general polynomial systems (whence they have a better numerical condition [8] and can be more easily solved), we expect that the polynomial systems arising from these geometric intersection conditions are numerically easier to solve. We have done some tests on working directly with the matrix structures, but we abandoned this approach because the general linear algebra operations turn out to be slower than the nested Horner schemes that are used to evaluate the expanded determinantal forms. Nevertheless, we expect computational progress from Newton's method adapted to this specific geometric situation [11], in combination with secant methods [49].

6.2. Computational and numerical experiments. The description in [16] leads to path following in a chain-by-chain fashion as defined by Pieri trees. Contemplating the posets in Figures 3.2 and 4.1 one sees that one can avoid the tracing of many solution paths when the continuation matches the poset structure. To solve a generic complex instance encoded by $(m, p) = (4, 3)$, the chain-by-chain Pieri algorithm takes 4 h 15 m 58 s 220 ms and has to trace 2310 solution paths. The new poset-oriented Pieri homotopies take only 1 h 7 m 20 s 160 ms (see Table 6.1) to trace 1663 paths. This saving is only partially explained by the reduction in the number of paths. The poset structure allows us to organize the symbolic homotopy constructors in a much more economical way.

TABLE 6.1

Number of paths and timings for $(m, p) = (4, 3)$ at every even level n, up to the total resolution time for a generic complex intersection problem. The first total is added to the time needed for solving a real target instance and results in the last line.

n	#paths	User cpu time
2	18	800 ms
4	72	6 s 918 ms
6	187	1 m 8 s 40 ms
8	462	8 m 27 s 460 ms
10	462	18 m 46 s 390 ms
12	462	38 m 42 s 360 ms
Total	1663	1 h 7 m 20 s 160 ms
Target	462	5 h 4 m 46 s 610 ms
Total	2125	6 h 13 m 15 s 840 ms

In Table 6.1 we summarize the computational results with pivoting done in a mixed fashion. Here we are removing pivots from both the top and the bottom, as in the original chain-by-chain original description in [16]. In this case, the Pieri homotopy has two moving equations:

$$H(X,t) = \begin{cases} \det(X|(1-t)S_X^t + tL_n) = 0, \\ \det(X|(1-t)S_X^b + tL_{n-1}) = 0, \\ \det(X|L_i) = 0, \\ i = 1, 2, \ldots, n-2, \end{cases} \qquad t \in [0,1], \tag{6.1}$$

where S_X^t and S_X^b are the special m-planes, respectively, for top and bottom pivots of X. The input 4-planes for the real target problem are osculating a rational normal curve, as prescribed in one of the Shapiro–Shapiro conjectures [32], [39], [45]. For these inputs, all solution 3-planes turn out to be real on the examples we ran, in

accordance with the Shapiro–Shapiro conjectures. Dividing the total time on the last line in Table 6.1 by 462, we arrive at 48 s 476 ms cpu time as the cost for one solution, which is the cost for one feedback law.

Note that one can change basis to bring two input 4-planes into special position and thus save the time spent at stage $n = 12$ in Table 6.1. Another optimization involves using linear algebra instead of continuation to satisfy the intersection conditions at the top level of the poset; see [16, Proposition 3.1]. But this optimization cuts off only milliseconds.

The input planes are represented by matrices of generators which are equivalent up to a transformation of basis. We have noticed that an orthonormal representation of the input planes—filtered through the QR-algorithm [14]—is responsible for considerable computational savings and an increased numerical accuracy of the solution planes. The difference between random and orthonormal inputs becomes especially significant for higher-dimensional problems. For instance, the case $(m, p) = (5, 3)$ leads to a 15-dimensional polynomial system of cubics with 6006 solutions. A 450 MHz Intel Pentium machine needs 16 h 17 m 16 s cpu time on random inputs, whereas it takes only 10 h 12 m 38 s on orthonormal input planes.

6.3. Comparison with SAGBI homotopies. The SAGBI homotopies were proposed in [16] to compute all p-planes that meet mp given m-planes in general position. The second author implemented those homotopies to verify large instances of the Shapiro–Shapiro conjectures. See [45] for a description and [32] and [39], [41] for other tests and related work on these conjectures.

To solve a generic complex instance, encoded by $(m, p) = (4, 3)$, the Pieri homotopy algorithm takes 1 h 7 m 20 s 160 ms (see Table 6.1). As reported in [45], SAGBI homotopies need 2 h 23 m 27 s 630 ms for the same problem using the localization patterns that lead to polynomials of lowest degree, and 5 h 23 m 36 s 840 ms with localization patterns used by Pieri homotopies. The common optimization of bringing two input planes into special position to diminish the dimension with two forces those localization patterns that lead to higher-degree polynomials, increasing the complexity of SAGBI homotopies.

SAGBI homotopies suffer from two drawbacks, not shared by the Pieri homotopies. First, they cannot solve the more general intersection problem treated in section 4. Second, a polyhedral continuation is needed to start up the SAGBI homotopies. These continuations happen in the top dimension, which makes them computationally intensive. In addition, these homotopies are both highly nonlinear in the continuation parameter, which leads to numerical instabilities as the problems get larger. Although the techniques in [12] help, they always involve a lower bound that may be too high for the machine precision.

To conclude the comparison, we contrast the analytic geometry as used in the SAGBI homotopies with the intrinsic geometry of the Pieri homotopies. With the SAGBI homotopies we lose all geometric information after the transformation of the problem into a polynomial system. The Pieri homotopies use the polynomial equations merely as a means to compute and are better able to exploit the evolving geometry.

6.4. Reality issues. The Pieri homotopies arose from the question [37], [38] whether a problem of enumerative geometry can have all its solutions be real. The developed software may be useful for verifying instances of conjectures such as the Shapiro–Shapiro conjectures; see [32], [39], [41], and [45]. Note that for the quantum Schubert calculus a counterexample to one of the analogues of the Shapiro–Shapiro conjectures is given in [40] for the case $q = 1$, $m = 2 = p$. For engineering applications,

where the machines are specified with real matrices, only real feedback laws are of interest. Several explicit root counting formulas are given in [29] which allow one to quickly determine the parity of the intersection numbers.

We end this paper describing some computational experiences with random real inputs. Table 6.2 gives the number of paths and timings reported for solving the case of $m = 3$, $p = 2$, and $q = 1$. For this random real instance, 25 real and 30 complex maps were found. The timings are for a 350 Mhz Intel II Pentium processor, with 64 Mb internal memory, running SunOs 5.7.

TABLE 6.2

Number of paths and timings for $(m, p) = (3, 2)$ for $q = 1$, at every odd level n. The first total concerns the time needed for the quantum Pieri homotopy algorithm on a random complex instance. Note that this is less than the time needed to solve a random real target system.

n	#paths	User cpu time
1	3	$<$ 1ms
3	8	240ms
5	21	1s 690ms
7	55	18s 390ms
9	55	35s 560ms
11	55	1m 16s 560ms
Total	197	2m 12s 880ms
Target	55	2m 25s 280ms
Total	252	4m 39s 760ms

TABLE 6.3

Number of paths and timings for $(m, p) = (4, 3)$, $q = 0$, and $k_i = 3$, for $i = 1, 2, 3, 4$, listed at every even level. The timings for the general Pieri homotopy algorithm, on a random complex instance are added up into the first total. The last total includes the resolution of a random real target system.

n	#paths	User cpu time
2	1	20ms
4	1	190ms
6	1	570ms
8	1	6s 800ms
10	1	4s 910ms
12	1	22s 780ms
Total	6	38s 190ms
Target	1	19s 500ms
Total	7	1m 25s 340ms

In the last example we indicate what could be done when the number of the solution is even, following the spirit of Sottile's original ideas [37]. In the hypersurface case for $(m, p) = (4, 3)$, there are 462 solution 3-planes. Since 462 is even, there might be a chance that all of them are complex. But suppose you specialize the input somewhat, so that we have 4 sets of three 4-planes that each meet at 1 line (= 2-plane) common to their set. Then we may replace the original hypersurface intersection conditions by the requirement that the solution 3-planes meet those 4 common 2-planes. Each of these 4 2-planes gives a codimension 3 condition and as $12 = 3+3+3+3$ we can apply the general Pieri algorithm which gives exactly 1 solution, which is real for real inputs. Note that this involves the resolution of a polynomial system of 84 cubic equations in 12 variables. Table 6.3 summarizes the computations.

Acknowledgments. We are indebted to Frank Sottile for the many explanations that helped us through the first implementation and for the hint that Pieri homotopies exist for the quantum Schubert calculus. The explanations of Joachim Rosenthal were most helpful. We thank MSRI for the opportunity of sharing an office for one full year. Frank Sottile and Mengnien Wu gave valuable comments on the preprint version of this paper. We are grateful for the remarks of the anonymous referees.

REFERENCES

[1] E. L. Allgower and K. Georg, *Numerical Continuation Methods, an Introduction*, Springer Ser. Comput. Math. 13, Springer-Verlag, Berlin, Heidelberg, New York, 1990.

[2] E. L. Allgower and K. Georg, *Numerical path following*, in Techniques of Scientific Computing Part 2, Handb. Numer. Anal. 5, P. G. Ciarlet and J.-L. Lions, eds., North-Holland, Amsterdam, 1997, pp. 3–207.

[3] S. Ariki, *Generic pole assignment via dynamic feedback*, SIAM J. Control Optim., 36 (1998), pp. 379–390.

[4] A. Bertram, *Quantum Schubert calculus*, Adv. Math., 128 (1997), pp. 289–305.

[5] R. W. Brockett and C. I. Byrnes, *Multivariate Nyquist criteria, root loci, and pole placement: A geometric viewpoint*, IEEE Trans. Automat. Control, 26 (1981), pp. 271–284.

[6] C. I. Byrnes, *Pole assignment by output feedback*, in Three Decades of Mathematical Systems Theory, Lecture Notes in Control and Inform. Sci. 135, H. Nijmacher and J. M. Schumacher, eds., Springer-Verlag, Berlin, 1989, pp. 13–78.

[7] D. A. Cox and S. Katz, *Mirror Symmetry and Algebraic Geometry*, Math. Surveys Monogr. 68, AMS, Providence, RI, 1999.

[8] J.-P. Dedieu, *Condition number analysis for sparse polynomial systems*, in Foundations of Computational Mathematics, F. Cucker and M. Shub, eds., Springer-Verlag, Berlin, 1997, pp. 75–101.

[9] J. P. Dedieu and M. Shub, *Newton's method for overdetermined systems of equations*, Math. Comp., to appear.

[10] D. F. Delchamps, *State Space and Input-Output Linear Systems*, Springer-Verlag, Berlin, 1988.

[11] A. Edelman, T. Arias, and S. T. Smith, *The geometry of algorithms with orthogonality constraints*, SIAM J. Matrix Anal. Appl., 20 (1998), pp. 303–353.

[12] T. Gao, T. Y. Li, J. Verschelde, and M. Wu, *Balancing the lifting values to improve the numerical stability of polyhedral homotopy continuation methods*, Appl. Math. Comput., to appear.

[13] I. Gohberg, M. A. Kaashoek, and F. van Schagen, *Partially Specified Matrices and Operators*: *Classification, Completion, Applications*, Oper. Theory Adv. Appl. 79, Birkhäuser, Boston, 1995.

[14] G. Golub and C. Van Loan, *Matrix Computations*, 3rd ed., The Johns Hopkins University Press, Baltimore, London, 1996.

[15] W. V. D. Hodge and D. Pedoe, *Methods of Algebraic Geometry*, Vol. 2, Cambridge University Press, Cambridge, UK, 1994. Paperback reprint of the 1952 edition.

[16] B. Huber, F. Sottile, and B. Sturmfels, *Numerical Schubert calculus*, J. Symbolic Comput., 26 (1998), pp. 767–788.

[17] T. Kailath, *Linear Systems*, Prentice-Hall, Englewood Cliffs, NJ, 1980.

[18] S. Kleiman and D. Laksov, *Schubert calculus*, Amer. Math. Monthly, 79 (1972), pp. 1061–1082.

[19] T. Y. Li, *Numerical solution of multivariate polynomial systems by homotopy continuation methods*, Acta Numer., 6 (1997), pp. 399–436.

[20] T. Y. Li, T. Sauer, and J. A. Yorke, *The cheater's homotopy: An efficient procedure for solving systems of polynomial equations*, SIAM J. Numer. Anal., 26 (1989), pp. 1241–1251.

[21] T. Y. Li and X. Wang, *Nonlinear homotopies for solving deficient polynomial systems with parameters*, SIAM J. Numer. Anal., 29 (1992), pp. 1104–1118.

[22] A. Morgan, *Solving Polynomial Systems Using Continuation for Engineering and Scientific Problems*, Prentice-Hall, Englewood Cliffs, NJ, 1987.

[23] A. Morgan and A. Sommese, *Computing all solutions to polynomial systems using homotopy continuation*, Appl. Math. Comput., 24 (1987), pp. 115–138.

[24] A. Morgan and A. Sommese, *A homotopy for solving general polynomial systems that respects m-homogeneous structures*, Appl. Math. Comput., 24 (1987), pp. 101–113.

[25] A. Morgan and A. Sommese, *Coefficient-parameter polynomial continuation*, Appl. Math.

Comput., 29 (1989), pp. 123–160. Errata: Appl. Math. Comput., 51 (1992), p. 207.
[26] A. Morgan and A. Sommese, *Generically nonsingular polynomial continuation*, in Computational Solution of Nonlinear Systems of Equations, E. L. Allgower and K. Georg, eds., AMS, Providence, RI, 1990, pp. 467–493.
[27] M. Pieri, *Formule di coincidenza per le serie algebriche ∞^n di coppie di punti dello spazio a n dimensioni*, Rend. Circ. Mat. Palermo, 5 (1891), pp. 252–268.
[28] M. S. Ravi and J. Rosenthal, *A smooth compactification of the space of transfer functions with fixed McMillan degree*, Acta Appl. Math., 34 (1994), pp. 329–352.
[29] M. S. Ravi, J. Rosenthal, and X. Wang, *Dynamic pole placement assignment and Schubert calculus*, SIAM J. Control Optim., 34 (1996), pp. 813–832.
[30] M. S. Ravi, J. Rosenthal, and X. Wang, *Degree of the generalized Plücker embedding of a Quot scheme and quatum cohomology*, Math. Ann., 311 (1998), pp. 11–26.
[31] J. Rosenthal, *On dynamic feedback compensation and compactifications of systems*, SIAM J. Control Optim., 32 (1994), pp. 279–296.
[32] J. Rosenthal and F. Sottile, *Some remarks on real and complex output feedback*, Systems Control Lett., 33 (1998), pp. 73–80. See http://www.nd.edu/~rosen/pole for a description of computational aspects of the paper.
[33] J. Rosenthal and J. C. Willems, *Open problems in the area of pole placement*, in Open Problems in Mathematical Systems and Control Theory, Comm. Control Engrg. Ser., V. D. Blondel, E. D. Sontag, M. Vidyasagar, and J. C. Willems, eds., Springer-Verlag, Berlin, 1998, pp. 181–191.
[34] H. Schubert, *Kalkül der abzählenden Geometrie*, Teubner, Leipzig, 1879. Reprint, Introduction by S. L. Kleiman, Springer-Verlag, Berlin, 1979.
[35] H. Schubert, *Beziehungen zwischen den linearen Räumen auferlegbaren charakteristischen Bedingungen*, Math. Ann., 38 (1891), pp. 588–602.
[36] A. J. Sommese and C. W. Wampler, *Numerical algebraic geometry*, in The Mathematics of Numerical Analysis, Lectures in Appl. Math. 32, J. Renegar, M. Shub, and S. Smale, eds., 1996, pp. 749–763.
[37] F. Sottile, *Enumerative geometry for real varieties*, in Algebraic Geometry—Santa Cruz 1995, University of California, Part I, Proc. Sympos. Pure Math., J. Kollár, R. Lazarsfeld, and D. R. Morrison, eds., AMS, Providence, RI, 1997, pp. 435–447.
[38] F. Sottile, *Pieri's formula via explicit rational equivalence*, Can. J. Math., 49 (1997), pp. 1281–1298.
[39] F. Sottile, *Real Schubert calculus: Polynomial systems and a conjecture of Shapiro and Shapiro*, Experiment. Math., to appear.
[40] F. Sottile, *Real rational curves in Grassmannians*, J. Amer. Math. Soc., 13 (2000), pp. 333–341.
[41] F. Sottile, *The special Schubert calculus is real*, Electronic Research Announcements of the AMS, 5 (1999), pp. 35–39. Also available online from http://www.ams.org/era.
[42] F. Sottile and B. Sturmfels, *A SAGBI basis for the quantum Grassmannian*, J. Pure Appl. Algebra, to appear.
[43] R. P. Stanley, *Some combinatorial aspects of the Schubert calculus*, in Combinatoire et Représentation du groupe symétrique, Lecture Notes in Math. 579, D. Foata, ed., Springer-Verlag, Berlin, 1977, pp. 217–251.
[44] J. Verschelde, *Algorithm* 795: *PHCpack: A general-purpose solver for polynomial systems by homotopy continuation*, ACM Trans. Math. Software, 25 (1999), pp. 251–276.
[45] J. Verschelde, *Numerical evidence for a conjecture in real algebraic geometry*, Experiment. Math., to appear. Paper and software available at the author's web page: http://www.mth.msu.edu/~jan.
[46] X. Wang, M. S. Ravi, and J. Rosenthal, *Algebraic and combinatorial aspects of the dynamic pole assignment problem*, in Systems and Networks: Mathematical Theory and Applications, U. Helmke, R. Mennicken, and J. Saurer, eds., Math. Res. 79, Akademie-Verlag, Berlin, 1994, pp. 547–550.
[47] X. Wang and J. Rosenthal, *A cell structure for the set of autoregressive systems*, Linear Algebra Appl., 1994, pp. 205–206, pp. 1203–1225.
[48] J. C. Willems, *Paradigms and puzzles in the theory of dynamical systems*, IEEE Trans. Automat. Control, 36 (1991), pp. 259–294.
[49] J. C. Yakoubsohn, *Finding zeros of analytic functions: α-theory for secant type methods*, J. Complexity, 15 (1999), pp. 239–281.
[50] H. G. Zeuthen, *Lehrbuch der abzählenden Methoden der Geometrie*, Lehrbücher math. Wissenschaften 39, Teubner, Leipzig, 1914.

SIAM J. CONTROL OPTIM.
Vol. 38, No. 4, pp. 1288–1308

HOMOGENEOUS STATE FEEDBACK STABILIZATION OF HOMOGENOUS SYSTEMS*

LARS GRÜNE†

Abstract. We show that for any asymptotically controllable homogeneous system in euclidean space (not necessarily Lipschitz at the origin) there exists a homogeneous control Lyapunov function and a homogeneous, possibly discontinuous state feedback law stabilizing the corresponding sampled closed loop system. If the system satisfies the usual local Lipschitz condition on the whole space we obtain semiglobal stability of the sampled closed loop system for each sufficiently small fixed sampling rate. If the system satisfies a global Lipschitz condition we obtain global exponential stability for each sufficiently small fixed sampling rate. The control Lyapunov function and the feedback are based on the Lyapunov exponents of a suitable auxiliary system and admit a numerical approximation.

Key words. homogeneous system, state feedback stabilization, control Lyapunov functions, Lyapunov exponents

AMS subject classifications. 93D15, 93D22, 93D30, 93D20

PII. S0363012998349303

1. Introduction.

In this paper we consider the problem of state feedback stabilization of homogeneous control systems in $\mathbb{R}^n$. This problem has been considered by a number of authors during the last few years; see, e.g., [16, 17, 18, 19, 22, 23, 24, 27] to mention just a few examples. Stability in this context will always mean asymptotic stability.

Homogeneous systems appear naturally as local approximations to nonlinear systems; cf., e.g., [15]. In order to make use of this approximation property in the design of locally stabilizing feedbacks for nonlinear systems the main idea lies in the construction of homogeneous feedbacks, i.e., feedback laws that preserve homogeneity for the resulting closed loop system. Utilizing a corresponding homogeneous Lyapunov function, those laws can then be shown to be locally stabilizing also for the approximated nonlinear system; cf. [15, 19, 21]. Regarding the existence of homogeneous stabilizing feedback laws, it was shown in [16] that if the system admits a continuous, but not necessarily homogeneous, stabilizing state feedback law, then there exists a homogeneous dynamic feedback stabilizing the system. Unfortunately, if we are looking for static state feedback laws, it is generally not true that any continuously stabilizable homogeneous system is stabilizable by a continuous *and* homogeneous state feedback law, as the examples in [24] show. Even worse, there exist homogeneous systems, e.g., Brockett's classical example [2], which—although asymptotically controllable—do not admit a stabilizing continuous state feedback law at all.

Brockett's results especially inspired the search for alternative feedback concepts. In the present paper we are going to use discontinuous state feedback laws for which the corresponding closed loop systems are defined as sampled systems. Although this is not a new concept (see, e.g., [13, 14, 25]), it has recently received new attention;

*Received by the editors December 17, 1998; accepted for publication (in revised form) August 27, 1999; published electronically May 2, 2000.
http://www.siam.org/journals/sicon/38-4/34930.html
†Fachbereich Mathematik, J.W. Goethe-Universität, Postfach 111932, D–60054 Frankfurt a.M., Germany (gruene@math.uni-frankfurt.de). This paper was written while the author was visiting the Dipartimento di Matematica, Università di Roma "La Sapienza," Italy, and was supported by DFG-grant GR1569/2-1.

see, e.g., the survey in [26]. In particular, it was shown in [5] that (global) asymptotic controllability is equivalent to the existence of a (globally) stabilizing discontinuous state feedback law for the sampled closed loop system. Stability in this context means asymptotic stability for the sampled trajectories (i.e., the feedback is evaluated only at discrete *sampling times* with the values being used until the next sampling time) where—in general—the intervals between two sampling times have to tend to zero close to the equilibrium and far away from it. A related but slightly different concept of a discontinuous feedback is the notion of discrete feedback introduced in [8]; here sampled trajectories are also considered, but with fixed intersampling times. With this approach it was possible to show in [11] that for semilinear systems asymptotic controllability is equivalent to (exponential) discrete feedback stabilizability.

The goal of the present paper is to provide a link between these two concepts in the framework of homogeneous systems. As in [11] we use a spectral characterization of asymptotic controllability by means of Lyapunov exponents, and obtain stability results for fixed sampling rates; as in [5] we construct the feedback based on a suitable (and here also homogeneous) control Lyapunov function and obtain stability not only for fixed intersampling times but for all sufficiently small ones. Furthermore, and this is a key feature of our construction, the resulting stabilizing state feedback law is homogeneous, thus rendering the corresponding closed loop system homogeneous. All this will be done under just the assumption that the corresponding homogeneous system is asymptotically controllable.

The organization of this paper is as follows. In section 2 we define homogeneous systems and introduce a class of auxiliary systems we call homogeneous-in-the-state. In some sense these systems have a built in homogeneity for each control value and can be simplified by suitable coordinate and time transformations. Section 3 provides the concepts of asymptotic controllability and stabilization by means of sampled feedback laws. After stating our main theorem at the end of this section, in section 4 we turn to the proof of this result for systems homogeneous-in-the-state by characterizing asymptotic controllability by means of Lyapunov exponents and constructing a suitable control Lyapunov function and a stabilizing feedback. After giving some hints about a numerical approximation of this feedback law in section 5, we will return to the homogeneous systems in section 6 and prove the stabilization result by showing that these systems can easily be transformed into systems homogeneous-in-the-state without losing the asymptotic controllability property. Finally, in section 7 we discuss two examples.

2. Homogeneous systems. We consider a class of systems

$$\dot{x}(t) = g(x(t), w(t)) \tag{2.1}$$

on $\mathbb{R}^n$, where $w(\cdot) \in \mathcal{W}$ and $\mathcal{W}$ denotes the space of measurable and locally essentially bounded functions from $\mathbb{R}$ to $W \subset \mathbb{R}^m$. We assume that the vector field g is continuous, $g(\cdot, w)$ is locally Lipschitz on $\mathbb{R}^n \setminus \{0\}$ for each $w \in W$, and g satisfies the following property.

DEFINITION 2.1. *We call* g homogeneous *if there exist* $r_i > 0$, $i = 1, \dots, n$, $s_j > 0$, $j = 1, \dots, m$, *and* $\tau \in (-\min_i r_i, \infty)$ *such that*

$$g(\Lambda_\alpha x, \Delta_\alpha w) = \alpha^\tau \Lambda_\alpha g(x, w) \text{ for all } w \in W, \alpha \geq 0, \tag{2.2}$$

where

$$\Lambda_\alpha = \begin{pmatrix} \alpha^{r_1} & 0 & \cdots & 0 \\ 0 & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \dots & 0 & \alpha^{r_n} \end{pmatrix} \quad \textit{and} \quad \Delta_\alpha = \begin{pmatrix} \alpha^{s_1} & 0 & \cdots & 0 \\ 0 & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \dots & 0 & \alpha^{s_m} \end{pmatrix}$$

are called dilation matrices. *With* $k = \min_i r_i$ *we denote the* minimal power *(of the state dilation) and the value* $\tau \in (-k, \infty)$ *is called the* degree *of the system.*

This definition generalizes the one given in [24] to the case of a multidimensional control input; see, e.g., [15] for an alternative definition (equivalent on $\mathbb{R}^n$) for vector fields on arbitrary manifolds. The use of dilation matrices instead of the usual dilation functions allows a more compact notation in what follows. Observe that if g is Lipschitz in the origin, then $\tau \geq 0$, and if g is globally Lipschitz, then $\tau = 0$. Furthermore, the definition implies $g(0,0) = 0$.

Corresponding to the dilation matrix Λ_α we define a function $N : \mathbb{R}^n \to [0,\infty)$ which can be interpreted as a "dilated norm" with respect to Λ_α. Denoting $d = 2\prod_{i=1}^n r_i$ we define $N(x)$ by

$$N(x) := \left(\sum_{i=1}^n x_i^{\frac{d}{r_i}} \right)^{\frac{1}{d}} \tag{2.3}$$

implying $N(0) = 0$, $N(x) > 0$ if $x \neq 0$, and $N(\Lambda_\alpha x) = \alpha N(x)$.

Note that the trajectories of (2.1) may tend to infinity in finite time if $\tau > 0$ and that uniqueness of the trajectory may not hold if $\tau < 0$; however, it holds away from the origin. As long as uniqueness holds (i.e., if $\tau \geq 0$ or the trajectory does not cross the origin) we denote the (open loop) trajectories of (2.1) by $x(t, x_0, w(\cdot))$ for each $x_0 \in \mathbb{R}^n$ and each $w(\cdot) \in \mathcal{W}$, where $x(0, x_0, w(\cdot)) = x_0$. Then from Definition 2.1 we obtain

$$x(t, \Lambda_\alpha x_0, \Delta_\alpha w(\alpha^\tau \cdot)) = \Lambda_\alpha x(\alpha^\tau t, x_0, w(\cdot)) \tag{2.4}$$

for $x_0 \in \mathbb{R}^n$. If uniqueness fails to hold, $x(\cdot, x_0, w(\cdot))$ shall denote one possible trajectory; in this case we implicitly assume the definitions of section 3 below to be valid for *all* possible trajectories.

In the remainder of this section we introduce and discuss a class of auxiliary systems which are homogeneous-in-the-state and will turn out to be useful for our analysis: Consider the class of systems

$$\dot{x}(t) = f(x(t), u(t)) \tag{2.5}$$

on $\mathbb{R}^n$, where $u(\cdot) \in \mathcal{U}$ and $\mathcal{U}$ denotes the space of measurable functions from $\mathbb{R}$ to some compact set $U \subset \mathbb{R}^m$. We assume that the vector field f is continuous, $f(\cdot, u)$ is locally Lipschitz on $\mathbb{R}^n \setminus \{0\}$ for each $u \in U$, and f satisfies the following property.

Definition 2.2. *We call* f homogeneous-in-the-state *if there exist* $r_i > 0$, $i = 1, \dots, n$ *and* $\tau \in (-\min_i r_i, \infty)$ *such that*

$$f(\Lambda_\alpha x, u) = \alpha^\tau \Lambda_\alpha f(x, u) \text{ for all } u \in U, \tag{2.6}$$

where Λ_α *is the* dilation matrix *as in Definition* 2.1, $k = \min_i r_i$ *is called the* minimal power, *and the value* $\tau \in (-k, \infty)$ *is called the* degree *of the system.*

Note that this definition implies $f(0,u) = 0$ for all $u \in U$. We denote the trajectories of (2.5) with initial value x_0 at the time $t = 0$ and control function $u(\cdot) \in \mathcal{U}$ again by $x(t, x_0, u(\cdot))$. Observe also that the trajectories of (2.5) may escape in finite time if $\tau > 0$ and that uniqueness of the trajectory may not hold in the origin if $\tau < 0$ (here again we use the convention as for the trajectories of (2.1)). As long as the trajectories exist and uniqueness holds we obtain from Definition 2.2 that

$$x(t, \Lambda_\alpha x_0, u(\alpha^\tau \cdot)) = \Lambda_\alpha x(\alpha^\tau t, x_0, u(\cdot)) \tag{2.7}$$

for all $x_0 \in \mathbb{R}^n$.

Besides being useful auxiliary systems for our stabilization problem for homogeneous systems, homogeneous-in-the-state systems themselves form an interesting class of systems. They generalize homogeneous bilinear and semilinear systems (see, e.g., [6, 7, 8, 11]). One interpretation of this structure is that the control affects parameters in the system rather than representing some force acting on the system; cf. the examples in [9, 10]. For this class of systems there also exist examples which are stabilizable but not with a continuous feedback law; see [26, example after Theorem A]. Note that this class can be generalized analogously to the generalization of semilinear systems made in [11]; all results in this paper can easily be adapted to that case.

The connection between homogeneous and homogeneous-in-the-state systems is easily seen: Given some homogeneous system (2.1) satisfying

$$g(\Lambda_\alpha x, \Delta_\alpha u) = \alpha^\tau \Lambda_\alpha (x, u)$$

we define

$$f(x,u) := g(x, \Delta_{N(x)} u). \tag{2.8}$$

Then it is immediate from the property $N(\Lambda_\alpha x) = \alpha N(x)$ of the dilated norm N that

$$\begin{aligned} f(\Lambda_\alpha x, u) &= g(\Lambda_\alpha x, \Delta_{N(\Lambda_\alpha x)} u) = g(\Lambda_\alpha x, \Delta_{\alpha N(x)} u) \\ &= g(\Lambda_\alpha x, \Delta_\alpha \Delta_{N(x)} u) = \alpha^\tau \Lambda_\alpha g(x, \Delta_{N(x)} u) \\ &= \alpha^\tau \Lambda_\alpha f(x,u); \end{aligned}$$

i.e., f is homogeneous-in-the-state.

Homogeneous and homogeneous-in-the-state systems can be considerably simplified applying suitable coordinate and time transformations. We will make use of this procedure for homogeneous-in-the-state systems: Using the dilated norm N from (2.3) the function

$$P(x) := \Lambda^{-1}_{N(x)} x$$

defines a projection from $\mathbb{R}^n \backslash \{0\}$ onto $N^{-1}(1)$ satisfying $P(\Lambda_\alpha x) = P(x)$ for all $\alpha > 0$. We denote the $n-1$ dimensional embedded unit sphere $\{x \in \mathbb{R}^n \mid \|x\| = 1\}$ by $\mathbb{S}^{n-1}$. Then, since $N(tx)$ is strictly increasing in $t \geq 0$ the function $S : N^{-1}(1) \to \mathbb{S}^{n-1}$, $S(x) = x/\|x\|$ is a diffeomorphism between these two manifolds. Thus, we can define a coordinate transformation $y = \Psi(x)$ by

$$\Psi(x) = N(x)^k S(P(x)), \quad \Psi^{-1}(y) = \Lambda_{\sqrt[k]{\|y\|}} S^{-1}\left(\frac{y}{\|y\|}\right),$$

and $\Psi(0) = 0$, $\Psi^{-1}(0) = 0$, which is continuous on $\mathbb{R}^n$ and C^1 on $\mathbb{R}^n \setminus \{0\}$. This definition implies

$$\Psi(\Lambda_\alpha x) = \alpha^k \Psi(x), \quad \Psi^{-1}(\alpha^k y) = \Lambda_\alpha \Psi^{-1}(y)$$

and by differentiation of $\Psi(\Lambda_\alpha x)$ and $\alpha^k \Psi(x)$ one sees

$$D\Psi(\Lambda_\alpha x) = \alpha^k \Lambda_\alpha^{-1} D\Psi(x).$$

Thus defining

$$\tilde{f}(y,u) = D\Psi(\Psi^{-1}(y)) f(\Psi^{-1}(y), u)$$

we obtain (with $x = \Psi^{-1}(y)$)

$$\tilde{f}(\alpha^k y, u) = D\Psi(\Lambda_\alpha x) f(\Lambda_\alpha x, u) = \alpha^k \Lambda_\alpha^{-1} D\Psi(x) \alpha^\tau \Lambda_\alpha f(x,u) = \alpha^\tau \alpha^k \tilde{f}(y,u)$$

implying

$$\tilde{f}(\alpha y, u) = \alpha^{\gamma+1} \tilde{f}(y,u),$$

with $\gamma = \tau/k$; i.e., $\tilde{f}$ is homogeneous-in-the-state with respect to the standard dilation $\Lambda_\alpha = \alpha \mathrm{Id}$, with mimimal power $k = 1$, and with degree $\tau = \gamma$.

Furthermore setting $\bar{f}(y,u) = \tilde{f}(y,u)\|y\|^{-\gamma}$ (which defines a time transformation for $\tilde{f}$) we obtain a system with degree $\tau = 0$.

In section 4 we will therefore first consider systems satisfying

$$f(\alpha x, u) = \alpha f(x,u) \text{ for all } x \in \mathbb{R}^n,\ \alpha \geq 0 \tag{2.9}$$

and will then retranslate the results to the general case. Observe that the transformed f is now globally Lipschitz with a uniform constant which we will denote by L.

3. Asymptotic controllability and feedback stabilization. In this section we give the precise definitions of asymptotic controllability and feedback stabilization. For this purpose we briefly describe the idea of sampling and the concept of control Lyapunov functions. We formulate the concepts for system (2.1) with the obvious modifications; however, all definitions also apply to system (2.5).

DEFINITION 3.1. *We call system* (2.1) asymptotically controllable (to the origin) *if for each* $x_0 \in \mathbb{R}^n$ *there exists* $w_{x_0}(\cdot) \in \mathcal{W}$ *such that* $\|x(t, x_0, w_{x_0}(\cdot))\| \to 0$ *as* $t \to \infty$.

We now discuss the concept of homogeneous state feedbacks. A *state feedback law* is a map $F : \mathbb{R}^n \to W$. A *homogeneous* state feedback law satisfies $F(\Lambda_\alpha x) = \Delta_\alpha F(x)$ for all $x \in \mathbb{R}^n$ and all $\alpha \geq 0$, thus implying $g(\Lambda_\alpha x, F(\Lambda_\alpha x)) = \alpha^\tau \Lambda_\alpha g(x, F(x))$; i.e., the *closed loop system* using F becomes homogeneous. Observe that W needs to satisfy some structural condition in order to allow nontrivial homogeneous feedbacks. In what follows we will assume

$$\Delta_\alpha W \subseteq W \text{ for all } \alpha \geq 0, \quad \text{where } \Delta_\alpha W := \{\Delta_\alpha w \,|\, w \in W\}$$

which gives a necessary and sufficient condition for the fact that given some $c > 0$ any homogeneous map $F : \mathbb{R}^n \to \mathbb{R}^m$ satisfying $F(x) \in W$ on $\{x \in \mathbb{R}^d \,|\, N(x) = c\}$ satisfies $F(x) \in W$ for all $x \in \mathbb{R}^n$.

Note that we do not require any continuity property of F. This is due to the fact that in many examples stabilizing continuous feedbacks cannot exist; cf., e.g.,

[26, section 2.2], where Brockett's classical example [2] is discussed also which—in suitable coordinates—is in fact a homogeneous system. Furthermore, even if stabilizing continuous feedback laws exist, it is possible that no such law is homogeneous, as the examples in [24] show (Brockett's example and the first example from [24] will be discussed in section 7). However, using discontinuous feedbacks for the solutions of the classical closed loop system $\dot{x} = g(x, F(x))$ the usual existence and uniqueness results might not hold. In order to obtain a meaningful solution for the closed loop system we use the following concept of a sampled closed loop system.

DEFINITION 3.2 (sampled closed loop system). *Consider a feedback law $F : \mathbb{R}^n \to W$. An infinite sequence $\pi = (t_i)_{i\in\mathbb{N}_0}$ of times satisfying*

$$0 = t_0 < t_1 < t_2 < \cdots \quad \text{and} \quad t_i \to \infty \text{ as } i \to \infty$$

is called a sampling schedule. *The values*

$$\Delta t_i := t_{i+1} - t_i \quad \text{and} \quad d(\pi) := \sup_{i\in\mathbb{N}_0} \Delta t_i$$

are called the intersampling times *and the* sampling rate, *respectively. For any sampling schedule π the corresponding* sampled *or* π-trajectory $x_\pi(t, x_0, F)$ *with initial value $x_0 \in \mathbb{R}^n$ at initial time $t_0 = 0$ is defined inductively by*

$$x_\pi(t, x_0, F) = x(t - t_i, x_i, F(x_i)) \text{ for all } t \in [t_i, t_{i+1}], i \in \mathbb{N}_0,$$

where $x_i = x_\pi(t_i, x_0, F)$ and $x(t, x_i, F(x_i))$ denotes the (open loop) trajectory of (2.1) *with constant control value $F(x_i)$ and initial value x_i.*

Observe that this definition guarantees the existence and uniqueness of trajectories in positive time on their maximal intervals of existence (except possibly at the origin if $\tau < 0$, in which case we use the same convention as for open loop trajectories). Moreover, the sampled π-trajectories have a meaningful physical interpretation, as they correspond to an implementation of the feedback law F using a digital controller.

The next definition introduces control Lyapunov functions which will be vital for the construction of the feedback. Here we make use of the lower directional derivatives; see, e.g., [4] for an equivalent definition.

DEFINITION 3.3. *A continuous function $V : \mathbb{R}^n \to [0, \infty)$ is called a* control Lyapunov function (CLF) *if it is positive definite (i.e., $V(x) = 0$ if and only if $x = 0$), proper (i.e., $V(x) \to \infty$ as $\|x\| \to \infty$), and there exists a continuous and positive definite function $P : \mathbb{R}^n \to [0, \infty)$ such that for each bounded subset $G \subset \mathbb{R}^n$ there exists a compact subset $W_G \subset W$ with*

$$\min_{v\in \operatorname{cog}(x,W_G)} DV(x; v) \le -P(x) \text{ for all } x \in G.$$

Here $DV(x; v)$ denotes the lower directional derivative

$$DV(x; v) := \liminf_{t\searrow 0, v'\to v} \frac{1}{t} \left(V(x + tv') - V(x)\right),$$

$g(x, W_G) := \{g(x, w) \,|\, w \in W_G\}$, and $\operatorname{cog}(x, W_G)$ denotes the convex hull of $g(x, W_G)$.

The following definition now describes the stability concepts we will use in this paper. For this definition recall that a function $\gamma : [0, \infty) \to [0, \infty)$ is of class $\mathcal{K}$ if it satisfies $\gamma(0) = 0$ and is continuous and strictly increasing, and a function $\beta :$

$[0,\infty)^2 \to [0,\infty)$ is of class $\mathcal{KL}$ if it is decreasing to zero in the second and of class $\mathcal{K}$ in the first argument.

DEFINITION 3.4. *We call the sampled closed loop system from Definition* 3.2

(i) semiglobally practically stable with fixed sampling rate *if there exists a class* $\mathcal{KL}$ *function* β *such that for each open set* $B \subset \mathbb{R}^n$ *and each compact set* $K \subset \mathbb{R}^n$ *satisfying* $0 \in B \subset K$ *there exists* $h > 0$ *such that*

$$x_\pi(t, x_0, F) \notin B \;\;\Rightarrow\;\; \|x_\pi(t, x_0, F)\| \le \beta(\|x_0\|, t)$$

for all $x_0 \in K$ *and all* π *with* $d(\pi) \le h$,

(ii) semiglobally stable with fixed sampling rate *if* (i) *holds and the sampling rate* $h > 0$ *can be chosen independently of* B,

(iii) globally practically stable with fixed sampling rate *if* (i) *holds and the sampling rate* $h > 0$ *can be chosen independently of* K,

(iv) globally stable with fixed sampling rate *if* (i) *holds and the sampling rate* $h > 0$ *can be chosen independently of* K *and* B.

We call the stability in (i)–(iv) exponential *if* β *can be chosen such that the inequality* $\beta(\|x_0\|, t) \le Ce^{-\sigma t}\|x_0\|$ *holds for constants* $C, \sigma > 0$ *which may depend on* K, *and* uniformly exponential *if* $C, \sigma > 0$ *can be chosen independently of* K.

Note that each of the concepts (ii)–(iv) implies (i) which is equivalent to the s-stability property as defined in [5]; also cf. [26, sections 3.1 and 5.1]. Hence any of these concepts imply *global stability* for the (possibly nonunique) limiting trajectories as $h \to 0$. The difference lies "only" in the performance with a *fixed* sampling rate. From the applications point of view, however, this is an important issue, since, e.g., for an implementation of a feedback using some digital controller, arbitrary small sampling rates in general will not be realizable. Furthermore, if the sampling rate tends to zero the resulting stability may be sensitive to measurement errors if the feedback is based on a nonsmooth CLF; see [20, 26]. In contrast to this it is quite straightforward to see that for a fixed sampling rate the stability is in fact robust to small errors in the state measurement (small, of course, relative to the norm of the current state of the system) if the corresponding CLF is Lipschitz; cf. [26, Theorem E].

The main result we will prove in this paper is the following theorem on the existence of a homogeneous CLF V and a homogeneous stabilizing feedback F.

THEOREM 3.5. (a) *Consider system* (2.1) *satisfying Definition* 2.1 *with dilation matrices* Λ_α *and* Δ_α, *minimal power* $k > 0$, *and degree* $\tau \in (-k, \infty)$, *and assume asymptotic controllability. Then there exist* $\mu > 0$ *and a CLF* V *being Lipschitz on* $\mathbb{R}^n \setminus \{0\}$, *satisfying*

$$V(\Lambda_\alpha(x)) = \alpha^{2k} V(x)$$

and

$$\min_{v \in \operatorname{cog}(x, W_x)} DV(x; v) \le -2\mu N^\tau(x) V(x)$$

for the function N *from* (2.3) *and* $W_x = \Delta_{N(x)} U$ *for some suitable compact subset* $U \subset W$.

Furthermore there exists a feedback law $F : \mathbb{R}^n \to W$ *satisfying* $F(x) \in W_x$ *and* $F(\Lambda_\alpha x) = \Delta_\alpha F(x)$ *for all* $x \in \mathbb{R}^n$ *and all* $\alpha \ge 0$ *such that the corresponding sampled closed loop system is either*

(i) *semiglobally stable (if $\tau > 0$), or*
(ii) *globally uniformly exponentially stable (if $\tau = 0$), or*
(iii) *globally practically exponentially stable (if $\tau < 0$) with fixed sampling rate.*

(b) *The analogous result holds for system* (2.5) *satisfying Definition* 2.2. *Here we obtain*

$$\min_{v \in \operatorname{co} f(x,U)} DV(x;v) \le -2\mu N^\tau(x) V(x),$$

$F(x) \in U$, and $F(\Lambda_\alpha x) = F(x)$ for all $x \in \mathbb{R}^n$ and all $\alpha > 0$.

We will prove part (b) in section 4 and then use this result in order to prove part (a) in section 6.

4. Stabilization of systems homogeneous-in-the-state. In this section we will construct a Lyapunov function and a stabilizing feedback for system (2.9). Afterwards we retranslate this stabilization result to general systems homogeneous-in-the-state from Definition 2.2.

We start by characterizing asymptotic controllability of (2.9). For this purpose we introduce the finite time exponential growth rate (cf. [11, 12])

$$\lambda^t(x_0, u(\cdot)) = \frac{1}{t} \ln \frac{\|x(t, x_0, u(\cdot))\|}{\|x_0\|}.$$

It follows immediately from (2.9) that $x(t, \alpha x_0, u(\cdot)) = \alpha x(t, x_0, u(\cdot))$ and thus the growth rates satisfy $\lambda^t(x_0, u(\cdot)) = \lambda^t(\alpha x_0, u(\cdot))$ for all $x_0 \in \mathbb{R}^d \setminus \{0\}$ and all $\alpha > 0$. The meaning of λ^t is described by the following proposition.

PROPOSITION 4.1. *System* (2.9) *is asymptotically controllable if and only if there exist a time $T > 0$ and some $\rho > 0$ such that for each $x \in \mathbb{R}^n \setminus \{0\}$ there exists $u_x(\cdot) \in \mathcal{U}$ with*

$$\lambda^t(x, u_x(\cdot)) < -\rho \text{ for all } t \ge T. \tag{4.1}$$

Proof. Obviously (4.1) implies exponential controllability and thus, in particular, asymptotic controllability.

For the converse implication, since $\lambda^t(x, u(\cdot)) = \lambda^t(\alpha x, u(\cdot))$ it is sufficient to show (4.1) for $\|x\| = 1$, i.e. $x \in \mathbb{S}^{n-1}$. Asymptotic controllability implies that for each $x \in \mathbb{S}^{n-1}$ there exist $\tilde{u}_x(\cdot) \in \mathcal{U}$, $\tilde{t}_x > 0$, and $C_x > 0$ such that $\|\varphi(\tilde{t}_x, x, \tilde{u}_x(\cdot))\| < 1/2$, and $\|\varphi(t, x, \tilde{u}_x(\cdot))\| < C_x$ for all $t \in [0, \tilde{t}_x]$. By compactness of $\mathbb{S}^{n-1}$ and continuous dependence on the initial value we can choose the controls such that $T_1 = \sup_{x \in \mathbb{S}^{n-1}} \tilde{t}_x$ and $C = \sup_{x \in \mathbb{S}^{n-1}} C_x$ are finite. Now for each $x \in \mathbb{S}^{n-1}$ we define $u_x(\cdot)$ and a sequence t_i inductively by $t_0 = 0$ and

$$t_{i+1} = t_i + \tilde{t}_{x_i}, \quad u_x(t) = \tilde{u}_{x_i}(t - t_i),\ t \in [t_i, t_{i+1}],$$

where $x_i = \varphi(t_i, x, u_x(\cdot))/\|\varphi(t_i, x, u_x(\cdot))\|$. Now consider an arbitrary $t > 0$. Choosing t_i maximal with $t_i \le t$ (i.e., $t - t_i < T_1$ and $t_i > t - T_1$) this implies

$$\lambda^t(x, u_x(\cdot)) = \frac{t_i}{t} \lambda^{t_i}(x, u_x(\cdot)) + \frac{t - t_i}{t} \lambda^{t - t_i}(x_i, u_x(t_i + \cdot)) \le \frac{t - T_1}{t} \ln \frac{1}{2} + \frac{T_1}{t} \ln C,$$

where the last expression is independent of x and negative for all $t \ge T$ for $T > 0$ sufficiently large, which yields the assertion. ∎

In fact, we can show something more than just the negativity of the finite time exponential growth rates. We define the *Lyapunov exponent* of each trajectory by

$$\lambda(x, u(\cdot)) := \limsup_{t\to\infty} \lambda^t(x, u(\cdot))$$

and the supremum with respect to the state and infimum with respect to the control over these exponents by

$$\sigma := \sup_{x\in\mathbb{R}^n\setminus\{0\}} \inf_{u(\cdot)\in\mathcal{U}} \lambda(x, u(\cdot)).$$

Lyapunov exponents for control systems have been utilized in the analysis of bilinear systems (see, e.g., [6] for some basic concepts and [7] for a detailed exposition) and for the global stabilization of semilinear and the local stabilization of differentiable nonlinear systems at singular points [11]. In the homogeneous setup we obtain the following characterization.

PROPOSITION 4.2. *Consider the system* (2.9) *and its* sup-inf *Lyapunov exponent* σ. *Then for each* $\rho \in (0, |\sigma|)$ *there exists* $T > 0$ *such that for each* $x \in \mathbb{R}^n \setminus \{0\}$ *there exists* $u_x(\cdot) \in \mathcal{U}$ *with*

$$\lambda^t(x, u_x(\cdot)) < -\rho \text{ for all } t \geq T.$$

Proof. The proof is exactly as [11, Proof of Proposition 3.4]. □

Since by Proposition 4.1 for our class (2.9) of homogeneous systems asymptotic controllability immediately implies $\sigma < 0$, we obtain from Proposition 4.2 that (2.9) is asymptotically controllable if and only if $\sigma < 0$; i.e., we obtain a characterization of asymptotic controllability by means of Lyapunov exponents.

We will now use Proposition 4.2 for the construction of a homogeneous Lyapunov function for system (2.9). First observe that the projection

$$s(t, s_0, u(\cdot)) := \frac{x(t, x_0, u(\cdot))}{\|x(t, x_0, u(\cdot))\|}, \quad s_0 = \frac{x_0}{\|x_0\|}$$

of (2.9) onto $\mathbb{S}^{n-1}$ is well defined due to the homogeneity of the system. A simple application of the chain rule shows that s is the solution of

$$\dot{s}(t) = f_{\mathbb{S}}(s(t), u(t)), \quad f_{\mathbb{S}}(s, u) = f(s, u) - \langle s, f(s, u)\rangle s$$

and that for $s_0 = x_0/\|x_0\|$ the exponential growth rate λ^t satisfies

$$\lambda^t(x_0, u(\cdot)) = \lambda^t(s_0, u(\cdot)) = \frac{1}{t}\int_0^t q(s(\tau, s_0, u(\cdot)), u(\tau))d\tau$$

with $q(s, u) = \langle s, f(s, u)\rangle$. Unfortunately, this averaged integral does not allow a direct construction of a suitable Lyapunov function. Therefore we make use of an approximation by discounted integrals: Defining

$$J_\delta(s_0, u(\cdot)) := \int_0^\infty e^{-\delta\tau} q(s(\tau, s_0, u(\cdot)), u(\tau))d\tau$$

and the corresponding optimal value function

$$v_\delta(s_0) := \inf_{u(\cdot)\in\mathcal{U}} J_\delta(s_0, u(\cdot))$$

from Propositions 4.1 and 4.2 and [11, Lemma 3.5(i)] we obtain that if system (2.9) is asymptotically controllable then for each $\rho \in (0, |\sigma|)$ there exists $\delta_\rho > 0$ such that for all $\delta \in (0, \delta_\rho]$ and all $s_0 \in \mathbb{S}^{n-1}$ the inequality

$$\delta v_\delta(s_0) < -\rho$$

holds.

Note that v_δ is Hölder continuous and bounded for each $\delta > 0$; cf., e.g., [1]. We now fix some $\rho \in (0, \sigma)$ and some $\delta \in (0, \delta_\rho]$ and define

$$V_0(x) := e^{2v_\delta(x/\|x\|)}\|x\|^2.$$

LEMMA 4.3. *The function V_0 is a CLF which is homogeneous with degree* 1 *(with respect to the standard dilation) and satisfies*

$$\min_{v \in \mathrm{co} f(x,U)} DV_0(x; v) \le -2\rho V_0(x).$$

Proof. Homogeneity, positive definiteness, and properness follow immediately from the definition. Now for each $t > 0$ the function v_δ satisfies the dynamic programming principle

$$v_\delta(s_0) = \inf_{u(\cdot)\in\mathcal{U}} \left\{ \int_0^t e^{-\delta\tau} q(s(\tau, s_0, u(\cdot)), u(\tau))d\tau + e^{-\delta t} v_\delta(s(t, s_0, u(\cdot))) \right\};$$

see, e.g., [1]. Abbreviating $q(t, s_0, u(\cdot)) = q(s(t, s_0, u(\cdot)), u(t))$ and using $e^{-\delta t} - 1 \ge -\delta t$ we obtain for the integral part of this equality

$$\int_0^t e^{-\delta\tau} q(\tau, s_0, u(\cdot))d\tau \ge \int_0^t q(\tau, s_0, u(\cdot)) + (e^{-\delta t} - 1)M_q d\tau \ge t\lambda^t(s_0, u(\cdot)) - M_q\delta\frac{t^2}{2},$$

where M_q denotes a bound of $|q|$. Thus with $s_0 = x_0/\|x_0\|$ we obtain

$$\begin{aligned}
V_0(x_0) &\ge \inf_{u(\cdot)\in\mathcal{U}} \exp[2t\lambda^t(x, u(\cdot)) - M_q\delta t^2 + 2e^{-\delta t} v_\delta(s(t, s_0, u(\cdot)))]\|x_0\|^2 \\
&= \inf_{u(\cdot)\in\mathcal{U}} e^{2t\lambda^t(x,u(\cdot))} e^{-M_q\delta t^2} e^{2(e^{-\delta t}-1)v_\delta(s(t,s_0,u(\cdot)))} e^{2v_\delta(s(t,s_0,u(\cdot)))}\|x_0\|^2 \\
&= \inf_{u(\cdot)\in\mathcal{U}} \frac{\|x(t, x_0, u(\cdot))\|^2}{\|x_0\|^2} e^{-M_q\delta t^2} e^{2(e^{-\delta t}-1)v_\delta(s(t,s_0,u(\cdot)))} e^{2v_\delta(s(t,s_0,u(\cdot)))}\|x_0\|^2 \\
&= \inf_{u(\cdot)\in\mathcal{U}} e^{-M_q\delta t^2 + 2(e^{-\delta t}-1)v_\delta(s(t,s_0,u(\cdot)))} V_0(x(t, x_0, u(\cdot))) \\
&\ge \inf_{u(\cdot)\in\mathcal{U}} e^{-M_q\delta t^2 + 2(1-e^{-\delta t})\rho/\delta} V_0(x(t, x_0, u(\cdot))).
\end{aligned}$$

Now for each $t > 0$ we choose $u_t(\cdot) \in \mathcal{U}$ such that the infimum of the last expression is attained up to t^2. Using $b - b^2 \le 1 - e^{-b} \le b$ for $b > 0$ we can conclude

$$\begin{aligned}
V_0(x(t, x_0, u_t(\cdot))) - V_0(x_0) &\le (1 - e^{-M_q\delta t^2 + 2(1-e^{-\delta t})\rho/\delta}) V_0(x(t, x_0, u_t(\cdot))) + t^2 \\
&\le (1 - e^{-M_q\delta t^2 + 2t\rho - 2\delta t^2\rho}) V_0(x(t, x_0, u_t(\cdot))) + t^2 \\
&\le (-2t\rho + (M_q + 2\rho)\delta t^2) V_0(x(t, x_0, u(\cdot))) + t^2
\end{aligned}$$

for all $t > 0$ sufficiently small. Denoting $v_t = (x(t, x_0, u_t(\cdot)) - x_0)/t$ we obtain

$$\frac{1}{t}\left(V_0(x_0 + tv_t) - V_0(x_0)\right) \leq -2\rho V_0(x(t, x_0, u(\cdot))) + (M_q + 2\rho)\delta t V_0(x(t, x_0, u(\cdot))) + t$$

and since by compactness of U there exists a $v \in \mathrm{co} f(x_0, U)$ and a sequence $t_i \to 0$ such that $v_{t_i} \to v$ as $i \to \infty$ the assertion follows by the definition of DV_0. □

Based on V_0 and using the techniques from [5] (but with a more careful evaluation of the constants involved) we can now construct the stabilizing feedback law for system (2.9). To this end, for $\beta > 0$ we consider the approximation of V_0 via the inf-convolution

$$V_\beta(x) = \inf_{y \in \mathbb{R}^n} \left\{ V_0(y) + \frac{\|x - y\|^2}{2\beta^2} \right\}. \tag{4.2}$$

Observe that V_β is locally Lipschitz and $V_\beta \to V_0$ as $\beta \to 0$.

PROPOSITION 4.4. *For each $\mu \in (0, \rho)$ there exists $\beta > 0$ such that the function V_β is a Lipschitz continuous CLF which is homogeneous with degree 1 (with respect to the standard dilation) and satisfies*

$$\min_{v \in \mathrm{co} f(x,U)} DV_\beta(x; v) \leq -2\mu V_\beta(x).$$

Furthermore there exists a feedback law $F : \mathbb{R}^n \to U$ satisfying $F(\alpha x) = F(x)$ for all $x \in \mathbb{R}^n$, $\alpha > 0$ and constants $h > 0$ and $C > 0$ such that any π-trajectory corresponding to some partition π with $d(\pi) \leq h$ satisfies

$$\|x_\pi(t, x_0, F)\| \leq Ce^{-\mu t}\|x_0\|. \tag{4.3}$$

Proof. By its definition V_β is obviously positive definite. Now for each $x \in \mathbb{R}^n$ we denote by $y_\beta(x)$ a point realizing the minimum on the right-hand side of (4.2). Since V_0 is homogeneous with degree 1 we have that

$$\left\{ V_0(\alpha y) + \frac{\|\alpha x - \alpha y\|^2}{2\beta^2} \right\} = \alpha^2 \left\{ V_0(y) + \frac{\|x - y\|^2}{2\beta^2} \right\}$$

and thus in particular V_β is also homogeneous with degree 1 and hence proper, and we can choose $y_\beta(x)$ in such a way that $y_\beta(\alpha x) = \alpha y_\beta(x)$. Since V_0 is strictly increasing along the rays αx in $\alpha > 0$ it follows that $\|y_\beta(x)\| \leq \|x\|$.

Now we define

$$\zeta_\beta(x) := \frac{x - y_\beta(x)}{2\beta^2}$$

which implies $\zeta_\beta(\alpha x) = \alpha \zeta_\beta(x)$.

By [5, Lemma III.1 and III.2] (or by straightforward calculations) for this vector we can deduce the inequalities

$$V_\beta(x + \tau v) \leq V_\beta(x) + \tau\langle \zeta_\beta(x), v\rangle + \frac{\tau^2\|v\|^2}{2\beta^2} \tag{4.4}$$

and

$$V_0(y_\beta(x) + \tau v) \geq V_0(y_\beta(x)) + \tau\langle \zeta_\beta(x), v\rangle - \frac{\tau^2\|v\|^2}{2\beta^2}; \tag{4.5}$$

i.e., $\zeta_\beta(x)$ is a proximal supergradient of V_β in x and a proximal subgradient of V_0 in $y_\beta(x)$ (see, e.g., [3] for an exposition of these concepts). We choose the feedback $F(x)$ in such a way that

$$\langle \zeta_\beta(x), f(x, F(x))\rangle = \inf_{u\in U}\langle \zeta_\beta(x), f(x,u)\rangle$$

and $F(\alpha x) = F(x)$ for all $x \in \mathbb{R}^n \setminus \{0\}$ and all $\alpha > 0$. The value $F(0)$ can be chosen arbitrarily.

Now consider points $x \in \mathbb{R}^n$ with $\|x\| = 1$, i.e. $x \in \mathbb{S}^{n-1}$. For these points the Hölder continuity of V_0 (which is inherited from the Hölder continuity of v_δ) and the definitions of V_β and ζ_β imply

$$\frac{1}{2\beta^2}\|y_\beta(x) - x\|^2 \le V_0(x) - V_0(y_\beta(x)) \le H\|y_\beta(x) - x\|^\nu$$

and thus

$$\|\zeta_\beta(x)\|\|y_\beta(x) - x\| \le \bar{H}\beta^{\frac{2\nu}{2-\nu}} \quad \text{with} \quad \bar{H} = 2^{\frac{2\nu-2}{2-\nu}} H^{\frac{2}{2-\nu}}, \tag{4.6}$$

where $H > 0$ and $\nu \in (0, 1]$ denote the Hölder constant and exponent of V_0 on $\{x \in \mathbb{R}^n \mid \|x\| \le 1\}$. From (4.6) and the definition of V_β we immediately obtain

$$|V_0(y_\beta(x)) - V_\beta(x)| \le \bar{H}\beta^{\frac{2\nu}{2-\nu}}. \tag{4.7}$$

Now the Lipschitz continuity of f implies that

$$\langle \zeta_\beta(x), f(x, F(x))\rangle \le \min_{u\in U}\langle \zeta_\beta(x), f(y_\beta(x), u)\rangle + L\|\zeta_\beta(x)\|\|y_\beta(x) - x\|$$

and by (4.5) and the definition of DV_0 it follows that $\langle \zeta_\beta(x), v\rangle \le DV_0(y_\beta(x), v)$ for all $v \in \mathbb{R}^n$. Thus by the linearity of the scalar product and Lemma 4.3 we can conclude

$$\min_{u\in U}\langle \zeta_\beta(x), f(y_\beta(x), u)\rangle = \min_{v\in \operatorname{co} f(y_\beta(x),U)}\langle \zeta_\beta(x), v\rangle \le -2\rho V_0(y_\beta(x)).$$

Combining these inequalities with (4.6) and (4.7) yields

$$\langle \zeta_\beta(x), f(x, F(x))\rangle \le -2\rho V_\beta(x) + 2\rho\bar{H}\beta^{\frac{2\nu}{2-\nu}} + L\bar{H}\beta^{\frac{2\nu}{2-\nu}}. \tag{4.8}$$

Defining

$$f_x^\tau := \frac{1}{\tau}\int_0^\tau f(x(t, x, F(x)), F(x))dt$$

and using $M := \sup_{\|x\|\le 2, u\in U} f(x, u)$ and the Lipschitz continuity of f for $\tau > 0$ sufficiently small we obtain

$$\|f_x^\tau - f(x, F(x))\| \le ML\tau, \quad \|f_x^\tau\| \le M.$$

Thus by (4.4), (4.8), and the fact that $\|\zeta_\beta(x)\| \le C_\beta$ for all $x \in \mathbb{S}^{n-1}$ and some suitable $C_\beta > 0$ we can conclude

$$\begin{aligned} V_\beta(x(\tau, x, F(x))) - V_\beta(x) &= V_\beta(x + \tau f_x^\tau) - V_\beta(x) \\ &\le \tau\langle \zeta_\beta(x), f_x^\tau\rangle + \frac{\tau^2\|f_x^\tau\|^2}{2\beta^2} \\ &\le \tau\langle \zeta_\beta(x), f(x, F(x))\rangle + ML\tau^2\|\zeta_\beta(x)\| + \frac{\tau^2 M^2}{2\beta^2} \\ &\le \tau(-2\rho V_\beta(x) + (2\rho + L)\bar{H}\beta^{\frac{2\nu}{2-\nu}}) + \tau^2\left(MLC_\beta + \frac{M^2}{2\beta^2}\right). \end{aligned}$$

Denoting

$$\gamma_\beta := \sup_{x\in\mathbb{S}^{n-1}} \frac{(2\rho+L)\bar{H}\beta^{\frac{2\nu}{2-\nu}}}{V_\beta(x)}, \quad \tilde{C}_\beta := \sup_{x\in\mathbb{S}^{n-1}} \frac{MLC_\beta}{V_\beta(x)} + \frac{M^2}{2\beta^2 V_\beta(x)}$$

and exploiting homogeneity of $x(\cdot,x,F(x))$ and V_β, we obtain for arbitrary $x\neq 0$

$$V_\beta(x(\tau,x,F(x))) - V_\beta(x) \le \tau(-2\rho+\gamma_\beta)V_\beta(x) + \tau^2\tilde{C}_\beta V_\beta(x).$$

Now let $\varepsilon = \rho-\mu$, choose $\beta>0$ such that $\gamma_\beta\le\varepsilon$, and $\Delta t>0$ such that $\Delta t\tilde{C}_\beta\le\varepsilon$. Then we obtain

$$V_\beta(x(\tau,x,F(x))) - V_\beta(x) \le -2\tau\mu V_\beta(x)$$

for all $\tau\in(0,\Delta t]$, which implies the first assertion, and by

$$V_\beta(x(\tau,x,F(x))) \le V_\beta(x) - 2\tau\mu V_\beta(x) \le e^{-2\tau\mu}V_\beta(x)$$

we obtain the second assertion by the homogeneity of degree 1 of V_β. □

This proposition shows the stabilization for systems of type (2.9). In order to prove Theorem 3.5(b) it remains to retranslate this result to general homogeneous-in-the-state systems.

Proof of Theorem 3.5(b). Obviously, if the system defined by f is asymptotically controllable, then the transformed system defined by $\bar{f}$ is asymptotically controllable. Thus from Proposition 4.4 we obtain $\bar{V}=V_\beta$ and $\bar{F}=F$ satisfying the assertion for $\bar{f}$ which is homogeneous-in-the-state with $\Lambda_\alpha=\alpha\mathrm{Id}$, $k=1$, and $\tau=0$.

We start by showing the result for the system defined by $\tilde{f}(x,u)=\bar{f}(x,u)\|x\|^\gamma$ being homogeneous-in-the-state with $\Lambda_\alpha=\alpha\mathrm{Id}$, $k=1$, and $\tau=\gamma$. Let $\tilde{V}(x)=\bar{V}(x)$. Then we immediately obtain

$$\min_{v\in\mathrm{co}\tilde{f}(x,U)} D\tilde{V}(x;v) = \|x\|^\gamma \min_{v\in\mathrm{co}\bar{f}(x,U)} D\tilde{V}(x;v) \le -\|x\|^\gamma 2\mu\tilde{V}(x).$$

Now observe that for each control function $u(\cdot)\in\mathcal{U}$ the trajectories $\tilde{x}$ and $\bar{x}$ of these systems satisfy

$$\tilde{x}(t,x_0,u(\cdot)) = \bar{x}(\bar{t}(t),x_0,u(\tilde{t}(\cdot))), \tag{4.9}$$

where $\tilde{t}(t)$ denotes the inverse of $\bar{t}(t)$ which is defined by

$$\bar{t}(t) = \int_0^t \|\tilde{x}(\tau,x_0,u(\cdot))\|^\gamma d\tau$$

and thus is well defined as long as the solution $\tilde{x}(t,x_0,u(\cdot))$ exists. If both $\tilde{x}$ and $\bar{x}$ uniquely exist for all $t\ge 0$ it is immediate that $\bar{t}(t)\to\infty$ as $t\to\infty$.

Setting $\tilde{F}(x)=\bar{F}(x)$ a $\tilde{\pi}$-trajectory $\tilde{x}_{\tilde{\pi}}(t,x_0,\tilde{F})$ of

$$\dot{\tilde{x}} = \tilde{f}(\tilde{x},\tilde{F}(\tilde{x})) \tag{4.10}$$

on some interval $[0,T]$ on which $\tilde{x}_\pi$ exists becomes a $\bar{\pi}$-trajectory $\bar{x}_{\bar{\pi}}(\bar{t}(t),x_0,\bar{F})$ of

$$\dot{\bar{x}} = \bar{f}(\bar{x},\bar{F}(\bar{x})), \tag{4.11}$$

where $\bar\pi = (\bar t_i)_{i\in\mathbb{N}_0}$ is given by $\bar t_i = \bar t(\tilde t_i)$ with $\tilde\pi = (\tilde t_i)_{i\in\mathbb{N}_0}$. Now we distinguish the three cases.

(i) $\gamma > 0$: By the choice of $\bar F$ there exist $C, \mu, h > 0$ such that inequality (4.3) holds for each $\bar\pi$-trajectory $\bar x_{\bar\pi}$ of (4.11) with $d(\bar\pi) \le h$ and each $x \in \mathbb{R}^n$. Now consider a compact set $K \subset \mathbb{R}^n$ with $0 \in \operatorname{int} K$. Let $C_K := \sup_{x\in K} \|x\|$, consider a $\tilde\pi$-trajectory $\tilde x_{\tilde\pi}(t, x_0, \tilde F)$ of (4.10) with $d(\tilde\pi) \le h(CC_K)^{-\gamma}$ and $x \in K$, and assume that there exists a (minimal) time $t^* > 0$ such that $\|\tilde x_{\tilde\pi}(t^*, x_0, \tilde F)\| = C\|x\|$. Without loss of generality we may assume $t^* = \tilde t_l \in \tilde\pi$ for some $l > 0$, otherwise we may reduce the sampling interval containing t^*. Then since $\|\tilde x_{\tilde\pi}(t, x_0, \tilde F)\| \le CC_K$ for all $t \in [0, \tilde t_l]$ the rescaled $\bar\pi$ satisfies $\bar t_i - \bar t_{i-1} \le h$ for all $i = 1, \dots, l$; thus we obtain

$$\|\tilde x_{\tilde\pi}(\tilde t_l, x_0, \tilde F)\| = \|\bar x_{\bar\pi}(\bar t_l, x_0, \bar F)\| \le Ce^{-\mu\bar t_l}\|x_0\| < C\|x_0\|$$

contradicting the choice of $t^* = \tilde t_l$. Thus $\|\tilde x_{\tilde\pi}(t, x_0, \tilde F)\| \le C\|x\|$ holds for all $t \ge 0$, and hence $d(\bar\pi) \le h$, implying

$$\|\tilde x_{\tilde\pi}(t, x_0, \tilde F)\| \le Ce^{-\mu\bar t(t)}\|x_0\|$$

which implies the desired stability estimate with $\beta(\|x\|, t) = Ce^{-\mu\bar t(t)}\|x\|$ which is of class $\mathcal{KL}$ because the corresponding trajectories stay inside some compact set, thus exist for all $t \ge 0$, and are unique since $\gamma > 0$, hence $\bar t(t) \to \infty$ as $t \to \infty$.

(ii) $\gamma = 0$: In this case the assumption follows immediately from Proposition 4.4.

(iii) $\gamma < 0$: As in case (i) there exist $C, \mu, h > 0$ such that inequality (4.3) holds for each $\bar\pi$-trajectory $\bar x_{\bar\pi}$ of (4.11) with $d(\bar\pi) \le h$ and each $x \in \mathbb{R}^n$. Consider a compact set $K \subset \mathbb{R}^n$ and an open set $B \subset \mathbb{R}^n$ with $0 \in B \subset K$. Let $C_K = \sup_{x\in K} \|x\|$, $C_B = \inf_{x\notin B} \|x\|/2 > 0$. By continuous dependence on the initial value and compactness we can choose $s > 0$ such that $\|\tilde x(t, x_0, u)\| \le 2C^{-1}C_B$ for all $\|x_0\| = C^{-1}C_B$, all $u \in U$, and all $t \in [0, s]$. Then (2.7) implies (recall $\tau = \gamma < 0$ and $\Lambda_\alpha = \alpha\mathrm{Id}$) the inequality $\|\tilde x(t, x_0, u)\| \le 2\|x_0\|$ for all $\|x_0\| \ge C^{-1}C_B$, all $u \in U$, and all $t \in [0, s]$.

Now pick an arbitrary $\tilde\pi$-trajectory $\tilde x_{\tilde\pi}(t, x_0, \tilde F)$ of (4.10) with $\tilde\pi$ satisfying $d(\tilde\pi) \le \min\{h(C^{-1}C_B)^{-\gamma}, s\}$ and $x_0 \in K$, $\|x_0\| \ge C^{-1}C_B$, and consider an interval $[t_*, t^*]$ such that $\|\tilde x_{\tilde\pi}(t, x, \tilde F)\| \ge C^{-1}C_B$ for all $t \in [t_*, t^*]$. Then we either have $t_* = 0 = \tilde t_0$ or there exist sampling times such that $t_* \in [\tilde t_{i-1}, \tilde t_i]$. In this case by $d(\tilde\pi) \le s$ and the choice of s we obtain $\|\tilde x_{\tilde\pi}(\tilde t_i, x_0, \tilde F)\| \le 2\|\tilde x_{\tilde\pi}(t_*, x_0, \tilde F)\|$.

Analogously to the case $\gamma > 0$ the choice of $d(\tilde\pi)$ now implies $d(\bar\pi) \le h$ and thus

$$\|\tilde x_{\tilde\pi}(t, x_0, \tilde F)\| \le Ce^{-\mu(\bar t(t) - \bar t(\tilde t_i))}\|\tilde x_{\tilde\pi}(\tilde t_i, x_0, \tilde F)\| \tag{4.12}$$

for all $t \in [\tilde t_i, t^*]$. This estimate implies that for each trajectory there exists a (minimal) time $T \ge 0$ such that the trajectory hits the set $\{x \in \mathbb{R}^n \mid \|x\| \le C^{-1}C_B\}$, and up to that time (4.12) implies the desired estimate with $\beta(\|x_0\|, t) = Ce^{-\mu\bar t(t)}\|x\| \le Ce^{-\mu(CC_K)^\gamma t}\|x_0\|$. After that time T, whenever the trajectory leaves this set at some time $t_* \ge T$ inequality (4.12) implies that it will enter again at some time $t^* > t_*$ and satisfies $\|\tilde x_{\tilde\pi}(t, x_0, \tilde F)\| \le 2CC^{-1}C_B = 2C_B$ for all $t \in [t_*, t^*]$. Hence $\|\tilde x_{\tilde\pi}(t, x_0, \tilde F)\| \le 2C_B$ for all $t \ge T$, and since $\|x\| \le 2C_B$ implies $x \in B$ we obtain the practical stability property.

So far we have shown the existence of $\tilde V$ and $\tilde F$ satisfying the assumptions of the theorem for $\tilde f$; hence it remains to translate the results to f. To this end we define $V(x) = \tilde V(\Psi(x))$ and $F(x) = \tilde F(\Psi(x))$. This implies

$$DV(x; f(x, u)) = D\tilde V(\Psi(x))\tilde f(\Psi(x), u)$$

and

$$x_\pi(t,x,F) = \tilde{x}_\pi(t,\Psi(x),\tilde{F})$$

and thus immediately the assertion since $\|\Psi(x)\| = N^k(x)$. □

5. Numerical approximation of V and F. In this section we briefly explain how a numerical approximation to the CLF V and the feedback law F can be computed.

Unfortunately, up to now no numerical method for the approximation of v_δ, V_0, or V_β is known, which also gives an approximation of the super- or subgradients and thus allows the approximation of F. However, if we slightly change our feedback concept (or, more precisely, the notion of a closed loop system) an approximation is possible. For this purpose we introduce the following definition.

DEFINITION 5.1. *Let $h : \mathbb{R}^n \to \mathbb{R}^+$ be an arbitrary map. A feedback law $F^h : \mathbb{R}^n \to U$ is called a* discrete feedback *if we apply it as a sampled feedback according to Definition* 3.2 *with fixed state dependent intersampling times $\Delta t_i = h(x_\pi(t_i, x_0, F^h))$.*

This definition generalizes the one given in [8] in the sense that the time step h may now depend on x. The name "discrete feedback" is motivated by the fact that the resulting system can be written as a discrete time system $x_{i+1} = x(h(x_i), x_i, F^h(x_i))$ for which F^h is a feedback law in the classical sense.

Returning to our simplified system (2.9), and again fixing some $\rho > 0$ and $\delta \in (0,\rho)$, we can apply the results from [8, 9], observing that the structural assumptions on the system in these references (i.e., bi- or semilinearity, accessibility, convexity of U) are needed only in order to show $v_\delta(x) < 0$. In particular, all the numerical approximation results remain valid; thus we can proceed as in [8, 9] and (i) approximate $\mathcal{U}$ by piecewise constant control functions, (ii) approximate the trajectories and the integral by numerical schemes, and (iii) compute an approximation of v_δ on a grid discretizing the state space $\mathbb{S}^{n-1}$. Proceeding this way for any given $\varepsilon > 0$ we find $h > 0$ and numerically computable functions $v_\delta^h : \mathbb{S}^{n-1} \to \mathbb{R}$ and $F_{\mathbb{S}}^h : \mathbb{S}^{n-1} \to U$ such that

$$v_\delta^h(s) \le -\rho + \varepsilon \tag{5.1}$$

and

$$v_\delta^h(s) + \varepsilon \ge \int_0^h q(s(t,s,F_{\mathbb{S}}^h(s)), F_{\mathbb{S}}^h(s))dt + e^{-\delta h} v_\delta^h(s(h,s,F_{\mathbb{S}}^h(s))) \tag{5.2}$$

hold for each $s \in \mathbb{S}^{n-1}$. This function v_δ^h is the function $\tilde{v}_\delta^a$ from [8]; inequality (5.1) follows from [8, Theorems 3.3, 5.3, and inequality (5.4)], inequality (5.2) is easily extracted from the proofs of [8, Lemma 5.1, Lemma 5.2, and Proposition 5.4] using again [8, inequality (5.4)]. The feedback $F_{\mathbb{S}}^h$ is defined by choosing a control value minimizing the right-hand side of (5.2) using the numerical approximations of the trajectory and the integral. Observe that the state space $\mathbb{S}^{n-1}$ to be discretized here is somewhat more difficult to handle than the projective space $\mathbb{P}^{n-1}$ appearing in [8, 9], since for $n \ge 3$ a single map cannot be sufficient for the parametrization of $\mathbb{S}^{n-1}$ without introducing singularities. Hence, numerically, one either needs to work directly on $\mathbb{S}^{n-1}$, or one has to compute the solution using two parametrizations (e.g., the stereographic projection from the north and south pole) and consequently two grids for the representation of v_δ^h in local coordinates. This method was used for the second example in section 7.

Defining $F^h(x) = F^h_{\mathbb{S}}(x/\|x\|)$ analogous to Proposition 4.3 we can conclude that the function

$$V^h(x) := e^{2v^h_\delta(x/\|x\|)}\|x\|$$

is homogeneous, proper, and positive definite and satisfies

$$V^h(x(h,x,F^h(x))) \le (1 - 2h\rho + M(\varepsilon + \varepsilon h + h^2))V^h(x)$$

for some suitable constant $M > 0$ independent of h and ε; i.e., for any $\rho' \in (0,\rho)$ there exist sufficiently small $h > 0$ and $\varepsilon > 0$ such that

$$V^h(x(h,x,F^h(x))) \le (1 - 2h\rho')V^h(x). \tag{5.3}$$

Thus the function V^h is a (discrete time) Lyapunov function for the system controlled by the discrete feedback F^h according to Definition 5.1 with $h(x) \equiv h$, which immediately implies (exponential) stability.

As in the proof of Theorem 3.5(b) we can retranslate this result to arbitrary homogeneous-in-the-state systems. Analogous to the proof of this theorem denote the functions obtained for (2.9) by $\bar F^h$, $\bar V^h$, and $\bar h$. For the retranslation from $\bar f$ to $\tilde f$ we can use $\tilde F^h = \bar F^h$ and $\tilde F^h = \bar V^h$; however, following (4.9) we now have to use $\tilde h(x) = \tilde t(\bar h, x)$ as intersampling times, where

$$\tilde t(\bar h, x) = \int_0^{\bar h} \|\bar x(\tau, x, F^h(x))\|^{-\gamma} d\tau.$$

Passing from $\tilde f$ to f we define—again analogously to the proof of Theorem 3.5(b)—the feedback $F^h(x) = \tilde F^h(\Psi(x))$, the Lyapunov function $V^h(x) = \tilde V^h(\Psi(x))$, and the intersampling time $h(x) = \tilde h(\Psi(x))$. This way it is straightforward to see that (5.3)—now for the x-dependent h—remains valid and thus stability follows.

Observe that the time steps $h(x)$ are bounded from below by some positive constant on each compact set if and only if $\tau \ge 0$, and they are bounded on each open set not containing the origin if and only if $\tau \le 0$. In this way they behave just like the sampling rate for the theoretical feedback law from Theorem 3.5(b); however, here the stability is only guaranteed for these fixed intersampling times $h(x)$ and not for smaller ones as allowed in Definition 3.4.

6. Stabilization of homogeneous systems. We now return to the homogeneous system from Definition 2.1. The idea of the proof of Theorem 3.5(a) lies in the fact that for any asymptotically controllable homogeneous system we can find an asymptotically controllable homogeneous-in-the-state system. For this we find a CLF and a stabilizing feedback law by Theorem 3.5(b), which—retranslated to the homogeneous system—have the properties as stated in Theorem 3.5(a).

For this purpose recall that for each homogeneous system (2.1) we find an associated homogeneous-in-the-state system by (2.8). The following proposition shows that this system inherits the asymptotic controllability property.

PROPOSITION 6.1. *Consider a system* (2.1) *satisfying Definition* 2.1. *Assume that the system is asymptotically controllable. Then there exists a compact set of control values $U \subset W$ such that the homogeneous-in-the-state system* (2.8) *is asymptotically controllable using control functions with values in U.*

Proof. First observe that due to (2.7) it is sufficient to show that there exists a compact $U \subset W$ and a time $T > 0$ such that any initial value x_0 with $N(x_0) = 1$ can

be steered to some point x_1 with $N(x_1) \le 1/2$ in some time $t < T$ using a measurable control $u(\cdot)$ with $u(t) \in U$ for almost all $t \ge 0$. With this property asymptotic controllability easily follows by induction from (2.7).

In order to show the existence of this U first observe that, denoting the trajectories of f and g by x_f and x_g, respectively, the equality

$$x_f(t, x_0, u(\cdot)) = x_g(t, x_0, w(\cdot)) \quad \text{with } u(t) = \Delta^{-1}_{N(x_g(t,x_0,w(\cdot)))} w(t) \tag{6.1}$$

holds. Now consider the initial values $x_0 \in N^{-1}(1)$. For each of these points there exists a control $w_{x_0}(\cdot) \in \mathcal{W}$ such that $N(x_g(t_{x_0}, x_0, w_{x_0}(\cdot))) = 1/3$. Now by continuous dependence of the solution on the initial value we obtain that for each x_0 there exists an open neighborhood $B_{x_0} \ni x_0$ such that

$$N(x_g(t_{x_0}, x, w_{x_0}(\cdot))) \le 1/2$$

for all $x \in B_{x_0}$. Since $N^{-1}(1)$ is compact and is covered by the B_{x_0} we find a finite number $M \in \mathbb{N}$ of points x_0^i, $i = 1, \dots, M$, such that the sets $U_{x_0^i}$, $i = 1, \dots, M$, cover $N^{-1}(1)$. Thus setting $t_i = t_{x_0^i}$ and $w_i(\cdot) = w_{x_0^i}(\cdot)$ for each $x_0 \in N^{-1}(1)$ there exists a number $i \in \{1, \dots, M\}$ such that

$$N(x_g(t_i, x_0, w_i(\cdot))) \le 1/2.$$

Now we choose $u_i(t) = \Delta^{-1}_{N(x_g(t,x_0,w_i(\cdot)))} w_i(t)$ for all $t \in [0, t_i]$, $u_i(t) \in W$ arbitrary for $t > t_i$. Then by (6.1) we immediately obtain

$$N(x_f(t_i, x, u_i(\cdot))) \le 1/2$$

for each $x_0 \in N^{-1}(1)$ and some suitable $i \in \{1, \dots, M\}$. Since the functions $w_i(\cdot)$ are locally essentially bounded, i.e., essentially bounded on $[0, t_i]$, we can conclude that the functions $u_i(\cdot)$ are essentially bounded. Thus $\|u_i(\cdot)\|_\infty$ is finite for each $i = 1, \dots, M$ and also $\sup_{i=1,\dots,M} \|u_i(\cdot)\|_\infty$ is finite; hence there exists a compact $U \subset W$ such that $u_i(t) \in U$ for almost all $t > 0$ and all $i = 1, \dots, M$. ∎

Now we can complete the proof of Theorem 3.5.

Proof of Theorem 3.5(a). Consider the system homogeneous-in-the-state as defined by (2.8) with $U \subset W$ from Proposition 6.1. For this system from part (b) of this theorem we obtain a CLF V_f and a feedback F_f. Setting $V = V_f$ and $F(x) = \Delta_{N(x)} F_f(x)$ we immediately obtain the assertion. ∎

7. Examples. Finally, let us illustrate our results by two examples. The first example

$$g(x, w) = \begin{pmatrix} x_1 + w \\ 3x_2 + x_1 w^2 \end{pmatrix} \tag{7.1}$$

for $x = (x_1, x_2)^T \in \mathbb{R}^2$, $w \in W = \mathbb{R}$, is taken from [24], where it has been shown that a stabilizing continuous and homogeneous feedback law cannot exist for this system.

The vector field g is homogeneous with $\Lambda_\alpha = \operatorname{diag}(\alpha, \alpha^3)$ and $\Delta_\alpha = \alpha$. Thus we obtain $N(x) = (x_1^6 + x_2^2)^{1/6}$. For system (7.1) a stabilizing discrete feedback F^h has been computed numerically using the techniques of section 5. Analyzing the switching curves of the numerical feedback in this case, it was easy to derive the feedback

$$F(x) = \begin{cases} N(x), & x_1 \le -x_2^3, \\ -N(x), & x_1 > -x_2^3, \end{cases}$$

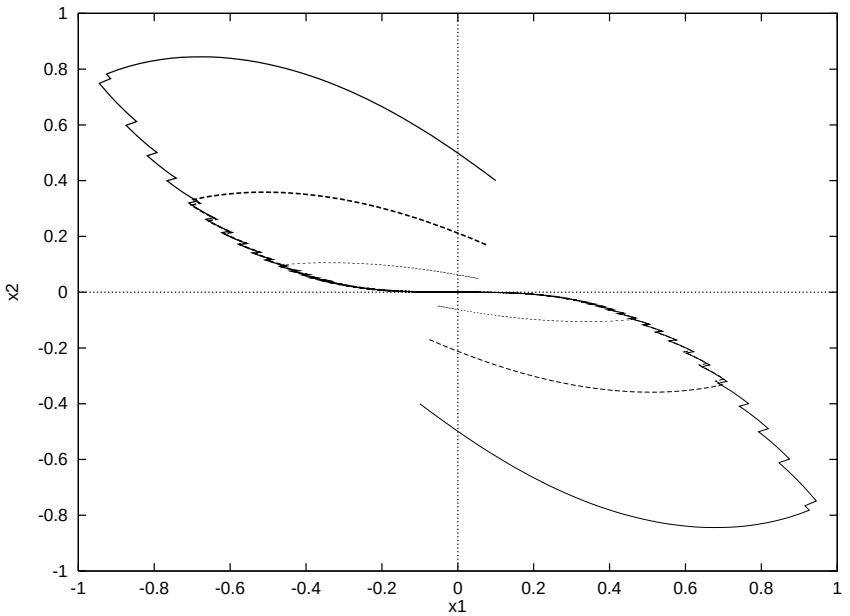


FIG. 7.1. *Trajectories for stabilized system* (7.1).

stabilizing the sampled system for all sufficiently small sampling rates. Figure 7.1 shows the corresponding (numerically simulated) sampled trajectories for some initial values; here the intersampling times have been chosen as $\Delta t_i = 0.01$ for all $i \in \mathbb{N}_0$.

The second example is the nonholonomic integrator given by Brockett [2] as an example for a system being asymptotically null controllable but not stabilizable by a continuous feedback law.

In suitable coordinates (cf. [26], where the physical meaning is also discussed) it reads

$$g(x,w) = \begin{pmatrix} w_1 \\ w_2 \\ x_1 w_2 \end{pmatrix} \tag{7.2}$$

for $x = (x_1, x_2, x_3)^T \in \mathbb{R}^3$, $w = (w_1, w_2)^T \in W = \mathbb{R}^2$. For this g we obtain homogeneity with $\Lambda_\alpha = \operatorname{diag}(\alpha, \alpha, \alpha^2)$ and $\Delta_\alpha = \operatorname{diag}(\alpha, \alpha)$, hence $N(x) = (x_1^4 + x_2^4 + x_3^2)^{1/4}$. Again a stabilizing discrete feedback law F_h has been computed numerically following section 5.

In this example it is also possible, in principle, to derive an explicit formula from the numerical results; it is, however, considerably more complicated, since a number of switching surfaces have to be identified. Hence we directly used the numerical approximation F^h of F for the simulation shown in Figures 7.2–7.4 in different projections; the time step is $h = 0.01$, and the control values were chosen as $U = \{-1, 1\}$.

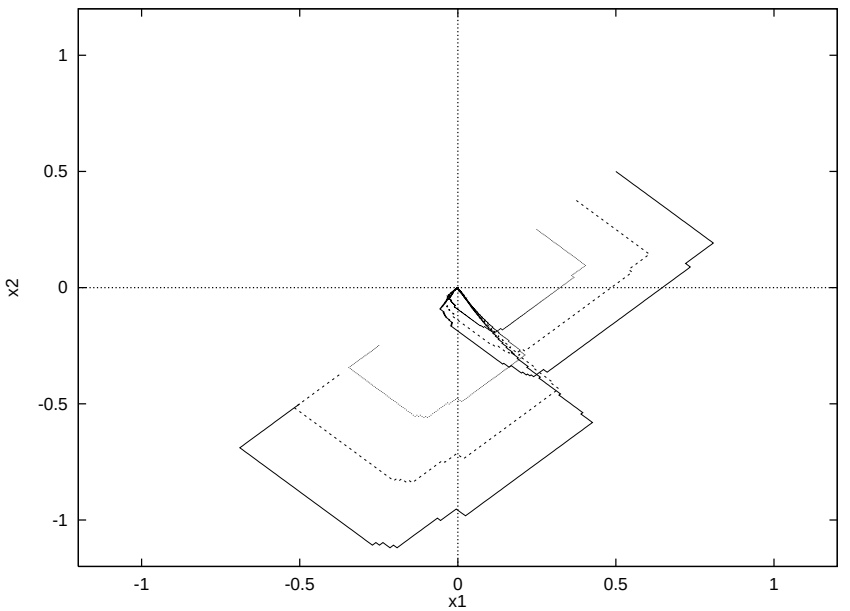


FIG. 7.2. *Trajectories for stabilized system* (7.2)*, projected to the* (x_1, x_2) *plane.*

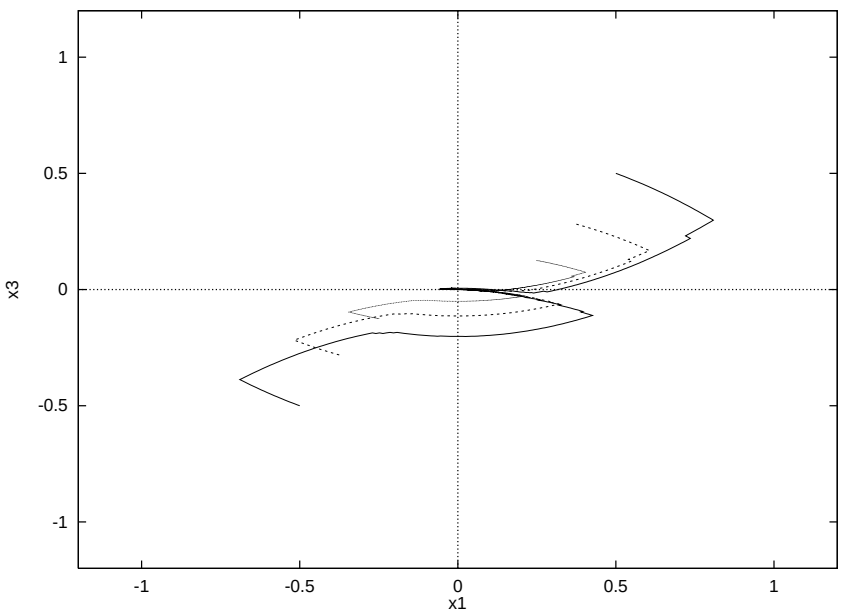


FIG. 7.3. *Trajectories for stabilized system* (7.2)*, projected to the* (x_1, x_3) *plane.*

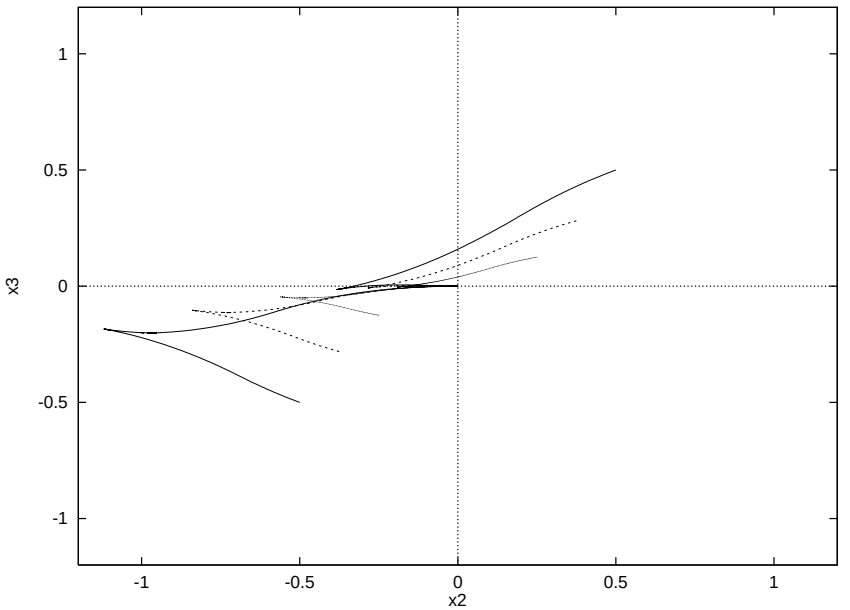


FIG. 7.4. *Trajectories for stabilized system* (7.2)*, projected to the* (x_2, x_3) *plane.*

REFERENCES

[1] M. BARDI AND I. CAPUZZO DOLCETTA, *Optimal Control and Viscosity Solutions of Hamilton-Jacobi-Bellman Equations*, Birkhäuser, Boston, 1997.

[2] R. BROCKETT, *Asymptotic stability and feedback stabilization*, in Differential Geometric Control Theory, R. Brockett, R. Millman, and H. Sussmann, eds., Birkhäuser, Boston, 1983, pp. 181–191.

[3] F. CLARKE, *Methods of Dynamic and Nonsmooth Optimization*, CBMS-NSF Regional Conf. Ser. Appl. Math. 57, SIAM, Philadelphia, 1989.

[4] F. CLARKE, Y. LEDYAEV, L. RIFFORD, AND R. STERN, *Feedback stabilization and Lyapunov functions*, SIAM J. Control Optim., to appear.

[5] F. CLARKE, Y. LEDYAEV, E. SONTAG, AND A. SUBBOTIN, *Asymptotic controllability implies feedback stabilization*, IEEE Trans. Automat. Control, 42 (1997), pp. 1394–1407.

[6] F. COLONIUS AND W. KLIEMANN, *Maximal and minimal Lyapunov exponents of bilinear control systems*, J. Differential Equations, 101 (1993), pp. 232–275.

[7] F. COLONIUS AND W. KLIEMANN, *The Dynamics of Control*, Birkhäuser, Boston, 1999.

[8] L. GRÜNE, *Discrete feedback stabilization of semilinear control systems*, ESAIM Control Optim. Calc. Var., 1 (1996), pp. 207–224.

[9] L. GRÜNE, *Numerical stabilization of bilinear control systems*, SIAM J. Control Optim., 34 (1996), pp. 2024–2050.

[10] L. GRÜNE, *Discrete feedback stabilization of nonlinear control systems at a singular point*, in Proceedings of the 4th European Control Conference, Brussels, 1997.

[11] L. GRÜNE, *Asymptotic controllability and exponential stabilization of nonlinear control systems at singular points*, SIAM J. Control Optim., 36 (1998), pp. 1485–1503.

[12] L. GRÜNE, *A uniform exponential spectrum for linear flows on vector bundles*, J. Dynam. Differential Equations, to appear.

[13] H. HERMES, *On stabilizing feedback attitude control*, J. Optim. Theory Appl., 31 (1980), pp. 373–384.

[14] H. HERMES, *On the synthesis of a stabilizing feedback control via Lie algebraic methods*, SIAM J. Control Optim., 18 (1980), pp. 352–361.

[15] H. HERMES, *Nilpotent and high-order approximations of vector field systems*, SIAM Rev., 33 (1991), pp. 238–264.

[16] H. Hermes, *Homogeneous feedback control for homogeneous systems*, Systems Control Lett., 24 (1995), pp. 7–11.
[17] H. Hermes, *Smooth homogeneous asymptotically stabilizing feedback controls*, ESAIM Control Optim. Calc. Var., 2 (1997), pp. 13–32.
[18] A. Iggidr and J.-C. Vivalda, *Global stabilization of homogeneous polynomial systems*, Nonlinear Anal., 18 (1992), pp. 1181–1186.
[19] M. Kawski, *Homogeneous feedback stabilization*, in New Trends in Systems Theory (Genova, 1990), Progr. Systems Control Theory 7, Birkhäuser, Boston, 1991, pp. 464–471.
[20] Y. Ledyaev and E. Sontag, *A Lyapunov characterization of robust stabilization*, Nonlinear Anal., 37 (1999), pp. 813–840.
[21] L. Rosier, *Homogeneous Lyapunov function for continuous vector fields*, Systems Control Lett., 19 (1992), pp. 467–473.
[22] E. Ryan, *Universal stabilization of a class of nonlinear systems with homogeneous vector field*, Systems Control Lett., 26 (1995), pp. 177–184.
[23] R. Sepulchre and D. Aeyels, *Homogeneous Lyapunov functions and necessary conditions for stabilization*, Math. Control Signals Systems, 9 (1996), pp. 34–58.
[24] R. Sepulchre and D. Aeyels, *Stabilizability does not imply homogeneous stabilizability for controllable homogeneous systems*, SIAM J. Control Optim., 34 (1996), pp. 1798–1813.
[25] E. Sontag, *Nonlinear regulation: The piecewise linear approach*, IEEE Trans. Automat. Control, 26 (1981), pp. 346–358.
[26] E. Sontag, *Stability and stabilization: Discontinuities and the effect of disturbances*, in Nonlinear Analysis, Differential Equations, and Control, NATO Sci. Ser. C Math. Phys. Sci. 528, Kluwer, Dordrecht, The Netherlands, 1999, pp. 551–598.
[27] J. Tsinias, *Remarks on feedback stabilizability of homogeneous systems*, Control Theory Adv. Tech., 6 (1990), pp. 533–542.

SIAM J. CONTROL OPTIM.
Vol. 38, No. 4, pp. 1309–1321

ON THE EFFECT OF NEGLECTING SENSOR DYNAMICS IN PARAMETER IDENTIFICATION PROBLEMS*

ASEN L. DONTCHEV†, MICHAEL P. POLIS‡, AND VLADIMIR M. VELIOV§

Abstract. In this paper, we consider a deterministic parameter identification problem for a nonlinear control system subject to disturbances and with output measured by a sensor with fast dynamics. We study the effect of neglecting sensor dynamics on the set of parameter values that are consistent with the measurements. We show that when neglecting dynamics we must significantly modify the corresponding static model of the sensor in order to obtain a set of parameter values that contains the true value and is consistent with the measurements. We describe a correct model of the static sensor which is in a sense minimal.

Key words. deterministic parameter identification, sensor dynamics, nonlinear control, singular perturbations, invariance

AMS subject classifications. 93B30, 34E15, 49K40, 93C10

PII. S0363012998348085

1. Introduction.

In deterministic parameter identification the goal is to find a set of values of an unknown parameter vector which are consistent with the measurements; for a collection of recent papers in this field see [12]. In virtually all work in this field the dynamics of the sensor are ignored by assuming instantaneous response. This assumption considerably simplifies the model and improves the numerical tractability of the problem. In this paper, we show that ignoring the fast dynamics may lead to unsatisfactory effects. More specifically, we consider a nonlinear control system subject to disturbances whose output is measured by a sensor with fast dynamics described by a singularly perturbed ordinary differential equation. We show that when neglecting fast dynamics we must significantly modify the model of the sensor in order to obtain a set of parameter values that contains the true value. We describe a model of the static sensor which is in a sense minimal to obtain correct parameter identification results.

The effects of neglecting sensor/actuator dynamics have been studied by Leitmann, Ryan, and Steinberg [10], Corless, Leitmann, and Ryan [5], and Kenneth and Taylor [8], in the context of stabilization and observation problems. In the present paper we explore this effect from the point of view of parameter identification, a problem which, to the authors' knowledge, has not been considered in the literature. In our analysis, we use tools from the recently developed proximal approach in invariance (viability) theory [3, 4, 13].

We consider a nonlinear control system where the state equation is of the form

$$\dot{x}(t) = f(p, x(t), \mathcal{F}(y(\cdot))(t), v(t)), \quad x(0) = x_0. \tag{1.1}$$

*Received by the editors November 20, 1998; accepted for publication (in revised form) October 29, 1999; published electronically May 2, 2000.
http://www.siam.org/journals/sicon/38-4/34808.html

†Mathematical Reviews, American Mathematical Society, 416 Fourth Street, Ann Arbor, MI 48107 (ald@ams.org).

‡School of Engineering and Computer Science, Oakland University, Rochester, MI 48309-4478 (polis@oakland.edu).

§Institute of Mathematics and Informatics, Bulgarian Academy of Sciences, 1113 Sofia, Bulgaria and Vienna University of Technology, Wiedner Hauptstr. 8–10/115, A-1040 Vienna, Austria (veliov @uranus.tuwien.ac.at).

Here $t \in [0,1]$ is the independent variable representing time, $x(t) \in \mathbb{R}^n$ is the state vector, $y(t) \in \mathbb{R}^m$ represents the output signal fed back into the system, $\mathcal{F}$ describes a feedback law, $v(t) \in \mathbb{R}^r$ is a disturbance, and $p \in \mathbb{R}^q$ is a parameter whose value is to be estimated. We assume that the initial state x_0 is known and the only information about the disturbance $v(\cdot)$ is that it is a Lebesgue measurable function on $[0,1]$ with values in V, a given closed subset of $\mathbb{R}^r$. The set of disturbance functions is denoted by $\mathcal{V}$. The state trajectory $x(\cdot)$ is regarded as an absolutely continuous function satisfying (1.1) almost everywhere (a.e.) in $[0,1]$. The output $y(t)$ is obtained through a sensor with dynamics described by the output equation

$$\varepsilon \dot{y}(t) = g(x(t), y(t), v(t)), \quad y(0) \in Y_0, \tag{1.2}$$

where ε is a "small" positive parameter describing the "fast" reaction of the sensor to changes of the state and the disturbance, and the set $Y_0 \subset \mathbb{R}^m$. The sensor with "neglected" dynamics is modeled by (1.2) with $\varepsilon = 0$, that is,

$$0 = g(x(t), y(t), v(t)). \tag{1.3}$$

We consider the following parameter identification problem. Let the a priori information for the value of the parameter p be represented by a set $P \subset \mathbb{R}^q$; that is, we know that $p \in P$. The goal is to find a subset of P, ideally, a single element—the true value of p, by measuring the output y. A "true" state trajectory $\hat{x}$ of the system associated with a function $\hat{v} \in \mathcal{V}$ representing the uncertainty (both unknown) results in a sensor "trajectory" $\hat{y}$ according to (1.2). We assume that a subset Δ of the interval $[0,1]$ is known such that the values of $\hat{y}(\cdot)$ are available at each moment of time $t \in \Delta$. In other words, the restriction $\hat{y}_{|\Delta}$ is the *exactly measured output* and, correspondingly, we suppose that the feedback law $\mathcal{F}(y(\cdot))$ uses only the values of y for $t \in \Delta$. Normally, $\mathcal{F}$ is a causal operator: its values $\mathcal{F}(y(\cdot))(t)$ depend only on the restriction of $y(\cdot)$ on $[0,t] \cap \Delta$, but the causality will not be essential for the considerations below. The data $\hat{y}(t)$, $t \in \Delta$ are used to reduce the uncertainty in the parameter p replacing the a priori estimation $p \in P$ by the a posteriori deterministic estimation $p \in P^{\varepsilon}_{\hat{y}(\cdot)}$, where

$$\begin{aligned} P^{\varepsilon}_{\hat{y}(\cdot)} \stackrel{def}{=} \{p \in P \mid &\text{there exist } v(\cdot) \in \mathcal{V} \text{ and a solution } (x(\cdot), y(\cdot)) \\ &\text{of (1.1)–(1.2) such that } y(t) = \hat{y}(t) \text{ for all } t \in \Delta\}. \end{aligned} \tag{1.4}$$

Observe that this estimation depends on the speed of the response represented by ε. Smaller values of ε correspond to faster responses. The sensor with an instantaneous response is modeled by (1.3), and applying this model to the parameter identification problem leads to the estimation $p \in P^0_{\hat{y}(\cdot)}$, where $P^0_{\hat{y}(\cdot)}$ is defined in the same way as $P^{\varepsilon}_{\hat{y}(\cdot)}$ but with (1.2) replaced by (1.3). Note that, due to the assumptions on the disturbance $v(\cdot)$, the output of the static sensor $y(\cdot)$ may be not absolutely continuous in time t.

If the set Δ consists of finitely many points (or more generally, has Lebesgue measure zero), then the values $v(t)$ in (1.3) can be chosen completely independently of the function $v(\cdot)$ in (1.1). In such a case, the static model can be written as

$$0 \in g(x(t), y(t), V) \tag{1.5}$$

or, equivalently, as

$$y \in K_0(x), \tag{1.6}$$

where

$$K_0(x) \stackrel{def}{=} \{y|\ 0 \in g(x,y,V)\}. \tag{1.7}$$

Taking into account that $\mathcal{F}(y(\cdot))(t)$ depends only on the values $y(s),\ s \in \Delta$, we can then describe the set $P^0_{\hat{y}(\cdot)}$ in the following way:

$$P^0_{\hat{y}(\cdot)} \stackrel{def}{=} \{p \in P|\ \text{there exist } v(\cdot) \in \mathcal{V} \text{ and an absolutely continuous solution } x(\cdot) \text{ of (1.1) corresponding to } \hat{y}(\cdot), v(\cdot) \text{ such that } \hat{y}(t) \in K_0(x(t)) \text{ for } t \in \Delta\}. \tag{1.8}$$

The estimation (1.8), however, can be totally wrong in certain situations. As shown later in this paper, the set $P^0_{\hat{y}(\cdot)}$ may be very different from the true estimation $P^{\varepsilon}_{\hat{y}(\cdot)}$ or even empty for arbitrarily small ε. The reason for such an effect is the fact that the values of the measurement $\hat{y}(t)$ may happen to be away from the (equilibrium) set $K_0(x(t))$ for some solutions $(x(\cdot), \hat{y}(\cdot))$ of (1.1)–(1.2) corresponding to some $v(\cdot) \in \mathcal{V}$. This is a consequence of possible "singular" behavior of the singularly perturbed equation (1.2) which occurs when the disturbance function $v(\cdot)$ changes rapidly in time, at least as fast as the dynamics of the sensor (1.2). Such a singular behavior is related to the discontinuity of the reachable set of singularly perturbed control systems first observed in [7]. In the "regular" case, as opposed to the singular one, the relation $\hat{y}(t) \in K_0(x(t))$ is satisfied approximately, with error proportional to ε and then the sensor model (1.6) needs to be ε-modified only.

The following is a simple example illustrating singular behavior. Consider two identical systems with states x_1 and x_2 in $\mathbb{R}$, described by

$$\dot{x}_1(t) = px_1(t), \quad x_1(0) = 1,$$
$$\dot{x}_2(t) = px_2(t), \quad x_2(0) = 1.$$

Suppose that each of the states is measured by a sensor, and that the second sensor is faster than the first, namely,

$$\varepsilon\dot{y}_1(t) = -y_1(t) + x_1(t) + v(t), \quad y_1(0) = 0, \tag{1.9}$$
$$\frac{\varepsilon}{2}\dot{y}_2(t) = -y_2(t) + x_2(t) + v(t), \quad y_2(0) = 0. \tag{1.10}$$

We assume that the disturbance vector v affecting the two sensors has bounded magnitude, i.e., $|v(t)| \le 1$. Let the set Δ consist of a finite number of points $t_i = ih$, $i = 1, \dots, k$, $h = 1/k$. Suppose that the true value of the parameter is $p = 0$ (the systems are static) and that the realization of the disturbance $v(\cdot)$ has the form $v(t) = v_0(t + \alpha\varepsilon)$, where (for convenience) $\alpha = \ln 2$ and

$$v_0(t) = \begin{cases} -1 & \text{for } t \in [t_i, t_i + \frac{h}{2}], \\ 1 & \text{for } t \in (t_i + \frac{h}{2}, t_{i+1}]. \end{cases}$$

The solutions of (1.9) and (1.10) associated with $x_1(t) = x_2(t) = 0$ and the disturbance $v(\cdot)$ just defined satisfy

$$\hat{y}_1(t_i) = 1 + O(\varepsilon), \quad \hat{y}_2(t_i) = \frac{3}{2} + O(\varepsilon).$$

Neglecting the dynamics of the sensors results in two identical static sensor models,

$$y_j(t_i) = x_j(t_i) + v_i, \quad |v_i| \le 1, \ \ i = 1, \dots, k, \quad j = 1, 2. \tag{1.11}$$

Obviously, the set of parameters p that are consistent with the measurements $\hat{y}_j(t_i)$ according to the sensor model (1.11) is empty. A possible way to handle this problem is to "regularize" the static model by artificially introducing new disturbances w_1 and w_2, each bounded by a constant d. Then we obtain the static sensor model

$$y_j(t_i) = x_j(t_i) + v_i + w_i^j, \quad |v_i| \le 1, \ \ |w_i^j| \le d, \ \ i = 1, \dots, k, \ j = 1, 2. \tag{1.12}$$

Since $|\hat{y}_1(t_i) - \hat{y}_2(t_i)| = 1/2 + O(\varepsilon)$, it is clear that to ensure $p = 0 \in P^0_{\hat{y}}$ one has to take $d \ge 1/4 - O(\varepsilon)$. That is, even for arbitrarily small ε, the static model (1.11) must be significantly modified to be usable for parameter identification.

In the following section we obtain a general sufficient condition for a static sensor model of the form $y \in K(x)$ to provide correct results when used instead of the "true" model (1.2) in the parameter identification problem. In section 3 we obtain conditions for regular behavior, where a slight modification, proportional to ε, of the instantaneous response model (1.6) suffices to obtain correct results. In section 4 we give a complete answer to the question of how to define the mapping $K(\cdot)$ of the static sensor, in the case of singular behavior.

2. From the dynamic to a static sensor model. We use the following notation. The space $\mathbb{R}^m$ is endowed with any Hilbert norm $|\cdot|$ with the scalar product denoted by $\langle \cdot, \cdot \rangle$. The corresponding unit ball is denoted by $\mathcal{B}$. For a closed set $Z \subset \mathbb{R}^m$ and a point $y \in \mathbb{R}^n$ we denote $\operatorname{dist}(y, Z) \stackrel{def}{=} \min\{|y - z| \mid z \in Z\}$. $\mathcal{P}_Z(y) \stackrel{def}{=} \{z \in Z \mid |z - y| = \operatorname{dist}(y, Z)\}$ is the set of projections of y on the closed set Z. The proximal normal cone to the closed set $Y \subset \mathbb{R}^m$ at the point $y \in Y$ is defined as

$$N_Y^{\perp}(y) \stackrel{def}{=} \{l \in \mathbb{R}^m \mid y \in \mathcal{P}_Y(y + \alpha l) \text{ for some } \alpha > 0\}.$$

In this section we study the following abstract model of the static sensor:

$$y(t) \in K(x(t)) + c\varepsilon\mathcal{B}, \tag{2.1}$$

where $K : \mathbb{R}^n \Rightarrow \mathbb{R}^m$ is a set-valued map and c is a positive constant. Given the measurement $\hat{y}(\cdot)$ of the system (1.1)–(1.2) for $t \in \Delta$, the deterministic estimation set of p based on the model (2.1) is described by

$$\begin{aligned} P_{\hat{y}(\cdot)} \stackrel{def}{=} \{p \in P \mid & \text{ there exist } v(\cdot) \in \mathcal{V} \text{ and a solution } x(\cdot) \text{ of (1.1) corresponding} \\ & \text{to } \hat{y}(\cdot), v(\cdot) \text{ such that (2.1) holds for } y(t) = \hat{y}(t), \ \ t \in \Delta\}. \end{aligned} \tag{2.2}$$

We use the following assumptions.

(A) There is a compact set $S \subset \mathbb{R}^n$ such that for each $p \in P$ and every $v(\cdot) \in \mathcal{V}$, every solution of the differential equation

$$\dot{x}(t) = f(p, x(t), \mathcal{F}(\hat{y}(\cdot))(t), v(t)), \ \ x(0) = x_0,$$

has values in the interior of S for all $t \in [0, 1]$. The functions $f : P \times \mathbb{R}^n \times \mathbb{R}^m \times V \to \mathbb{R}^n$ and $g : \mathbb{R}^n \times \mathbb{R}^m \times V \to \mathbb{R}^m$ are locally Lipschitz continuous. There exists a constant M such that $|f(p, x, \mathcal{F}(\hat{y}(\cdot))(t), v)| \le M$ for all $p \in P$, $x \in S$, $t \in \Delta$, and $v \in V$.

(B) For each $x \in S$ the set $K(x)$ is nonempty and closed. There exists a constant L such that $K(x'') \subset K(x') + L|x'' - x'|\mathcal{B}$ for every $x', x'' \in S$. In addition, we assume $Y_0 \subset K(x_0)$.

(C) There exists $\lambda > 0$ such that for every $x \in S$, $y \in K(x)$, $l \in N^{\perp}_{K(x)}(y)$, and $v \in V$

$$\langle g(x, y + l, v), l\rangle \le -\lambda |l|^2.$$

We will see further that assumption C plays a crucial role for passing from the dynamic sensor model to a static one. This assumption is inspired by the recently developed proximal approach to stability and invariance; see [3, 4, 13]. Here, in the general case, its form is fairly abstract; in the following sections we show that in specific cases it reduces to a computationally tractable condition.

The following theorem is the central result of the paper.

THEOREM 2.1. *Let assumptions* A, B, *and* C *hold, let the constant* c *in* (2.1) satisfy $c \ge LM/\lambda$, and let $\varepsilon > 0$. Then

$$P^{\varepsilon}_{\hat{y}(\cdot)} \subset P_{\hat{y}(\cdot)}.$$

The meaning of this result is clear: if we apply the static sensor model (2.1) under the above assumptions, then the estimation set $P_{\hat{y}(\cdot)}$ obtained on the basis of the measured output must contain the true value of the parameter p.

In the proof of the theorem we use the notion of contingent derivative (see, e.g., Aubin and Frankowska [2]) and the following lemma. We recall that $\xi \in \mathbb{R}^m$ is an element of the contingent derivative $DK(x, z; \eta)$ of the set-valued mapping $K : \mathbb{R}^n \Rightarrow \mathbb{R}^m$ at the point (x, z) on the graph of $K(\cdot)$ in the direction η if

$$\liminf_{h\to 0+} \frac{1}{h} \operatorname{dist}(z + h\xi, K(x + h\eta)) = 0.$$

LEMMA 2.2. *Let the functions* $x : [0, 1] \to \mathbb{R}^n$ *and* $y : [0, 1] \to \mathbb{R}^m$ *be absolutely continuous and let the set-valued mapping* $K : \mathbb{R}^n \Rightarrow \mathbb{R}^m$ *satisfy assumption* B. *Then the function* $\rho(t) \stackrel{def}{=} \operatorname{dist}(y(t), K(x(t)))$ *is absolutely continuous and*

$$\dot\rho(t) \le \min_{z\in \mathcal{P}_{K(x(t))}(y(t))} \min_{\xi \in DK(x(t),z;\dot x(t))} \frac{\langle \dot y(t) - \xi, y(t) - z\rangle}{\rho(t)}$$

for a.e. t *for which* $\rho(t) > 0$.

Proof. The absolute continuity of $\rho(\cdot)$ is obvious. Let t be a point where $\rho(t) > 0$ and each of the functions $x(\cdot)$, $y(\cdot)$, and $\rho(\cdot)$ is differentiable. Let $z \in \mathcal{P}_{K(x(t))}(y(t))$ and let $\xi \in DK(x(t), z; \dot x(t))$. Then there exist sequences $h_k \to 0+$ and $\omega_k \to 0$ such that

$$z + h_k\xi + h_k\omega_k \in K(x(t) + h_k\dot x(t)) \subset K(x(t + h_k)) + o(h_k)\mathcal{B},$$

where $o(\cdot)$ is any function with $o(h)/h \to 0$ when $h \to 0$. Then $z + h_k\xi + h_k\omega'_k \in K(x(t + h_k))$ for some $\omega'_k \to 0$. Hence,

$$\rho(t + h_k) \le |y(t + h_k) - z - h_k\xi| + o(h_k) \le |y(t) - z + h_k(\dot y(t) - \xi)| + o(h_k).$$

Since $|y(t) - z| = \rho(t) > 0$, taking the directional derivative of the norm in the right-hand side we obtain

$$\rho(t + h_k) \le \rho(t) + h_k \frac{\langle \dot y(t) - \xi, y(t) - z\rangle}{|y(t) - z|} + o(h_k).$$

Rearranging the last inequality and passing to the limit with k we obtain the desired inequality. □

Proof of Theorem 2.1. Let $p \in P^{\varepsilon}_{\hat{y}(\cdot)}$ and let $v^p_\varepsilon(\cdot), x^p_\varepsilon(\cdot)$ correspond to p according to (1.4). This means that there exists a solution $y(\cdot)$ of (1.2) with $x = x^p_\varepsilon$ and $v = v^p_\varepsilon$ such that $y(t) = \hat{y}(t)$ for $t \in \Delta$. Since $P^{\varepsilon}_{\hat{y}(\cdot)}$ and $P_{\hat{y}(\cdot)}$ depend only on the values of $\hat{y}(\cdot)$, these two sets will not change if we extend the domain of $\hat{y}$ to $[0,1]$ by taking $\hat{y}(t) = y(t)$ for all $t \in [0,1]$. In this way we ensure that $\hat{y}(\cdot)$ satisfies (1.2) with $x = x^p_\varepsilon$ and $v = v^p_\varepsilon$. Then, to prove the theorem, it is enough to establish the inclusion $\hat{y}(t) \in K(x^p_\varepsilon(t)) + c\varepsilon\mathcal{B}$ with $c = ML/\lambda$ or, equivalently, the inequality $\rho(t) \stackrel{def}{=} \operatorname{dist}(\hat{y}(t), K(x^p_\varepsilon(t))) \le c\varepsilon$.

According to the last part of assumption B, $\rho(0) = 0$. Lemma 2.2 implies that $\rho(\cdot)$ is absolutely continuous and for almost every t for which $\rho(t) > 0$ and at which $\dot{x}^p_\varepsilon(t)$ and $\dot{\hat{y}}(t)$ exist and (1.1)–(1.2) are satisfied, we have

$$\dot{\rho}(t) \le \min_{z \in \mathcal{P}_{K(x^p_\varepsilon(t))}(\hat{y}(t))} \min_{\xi \in DK(x^p_\varepsilon(t), z; \dot{x}^p_\varepsilon(t))} \frac{\langle \dot{\hat{y}}(t) - \xi, \hat{y}(t) - z\rangle}{\rho(t)}. \tag{2.3}$$

The term $\langle \dot{\hat{y}}(t), \hat{y}(t) - z\rangle$ can be estimated by

$$\langle \varepsilon^{-1} g(x^p_\varepsilon(t), \hat{y}(t), v^p_\varepsilon(t)), \hat{y}(t) - z\rangle \le -\frac{\lambda}{\varepsilon}(\rho(t))^2,$$

where in the last estimation we use assumption C and the relations

$$\hat{y}(t) - z \in N^{\perp}_{K(x^p_\varepsilon(t))}(z), \quad |\hat{y}(t) - z| = \rho(t).$$

Hence,

$$\begin{aligned} \dot{\rho}(t) &\le -\frac{\lambda}{\varepsilon}(\rho(t))^2 + \min_{z \in \mathcal{P}_{K(x^p_\varepsilon(t))}(\hat{y}(t))} \min_{\xi \in DK(x^p_\varepsilon(t), z; \dot{x}^p_\varepsilon(t))} \frac{|\langle \xi, \hat{y}(t) - z\rangle|}{\rho(t)} \\ &\le -\frac{\lambda}{\varepsilon}(\rho(t))^2 + \min_{z \in \mathcal{P}_{K(x^p_\varepsilon(t))}(\hat{y}(t))} \min_{\xi \in DK(x^p_\varepsilon(t), z; \dot{x}^p_\varepsilon(t))} |\xi|. \end{aligned} \tag{2.4}$$

In order to estimate the last term we prove that assumption B implies

$$\min_{\xi \in DK(x,z;l)} |\xi| \le L|l| \tag{2.5}$$

for every $x \in S$, $z \in K(x)$, and $l \in \mathbb{R}^n$. Indeed, for every $h > 0$ there exists $z_h \in K(x+hl)$ such that $|z - z_h| \le Lh$. Thus the sequence $\xi_h = (z_h - z)/h$ has a subsequence ξ_{h_k} converging to a point ξ, and $|\xi| \le L$. Since $z + h_k\xi_{h_k} = z_{h_k} \in K(x + h_k l)$, we obtain that $\xi \in DK(x, z; l)$; hence (2.5) holds.

Taking into account assumption A and (2.5), from (2.4) we obtain

$$\dot{\rho}(t) \le -\frac{\lambda}{\varepsilon}\rho(t) + LM.$$

The above inequality need not be fulfilled for those t for which $\rho(t) = 0$, but it is sufficient (together with $\rho(0) = 0$) to verify that the desired inequality $\rho(t) \le c\varepsilon$ holds. The proof is complete. □

The two coupled equations (1.1)–(1.2) represent a singular perturbation model. In the realm of singular perturbations, one usually assumes a certain type of stability

for the fast dynamics. For our model such an assumption means that the function g describing the dynamic sensor must ensure a certain tracking property for the solution of (1.2). Here we use the following assumption.

(D) There exists a positive constant λ such that for every $x \in S$, $y_1, y_2 \in \mathbb{R}^m$, and $v \in V$,

$$\langle g(x, y_2, v) - g(x, y_1, v), y_2 - y_1 \rangle \leq -\lambda |y_1 - y_2|^2.$$

Assumption D is related to the so-called one-sided Lipschitz condition which is used in the analysis of stiff differential equations. For various versions of this condition, see [6, 14]. In the case of a function g which is invertible and differentiable in y, assumption D is implied by the (uniform) negative definiteness of the matrix $\partial g/\partial y$ with respect to the scalar product $\langle \cdot, \cdot \rangle$. Note that this scalar product was arbitrarily chosen, and therefore the familiar, in the linear case, requirement that the real parts of the eigenvalues be negative is covered.

Under assumption D, assumption C has the following equivalent form which will be used in the next section.

PROPOSITION 2.3. *If assumptions* A, B, *and* D *are satisfied, then assumption* C *is equivalent to the following:*

$$\langle g(x, y, v), l \rangle \leq 0 \quad \textit{for all } x \in S,\ y \in K(x),\ v \in V,\ l \in N^{\perp}_{K(x)}(y). \tag{2.6}$$

Proof. Suppose that assumption C is satisfied and assume that, on the contrary,

$$\langle g(x, y, v), l \rangle = \delta > 0$$

for some $x \in S$, $y \in K(x)$, $v \in V$, and $l \in N^{\perp}_{K(x)}(y)$, say, with $|l| = 1$. Then

$$\langle g(x, y + \alpha l, v), \alpha l \rangle \geq \langle g(x, y, v), \alpha l \rangle - L' \alpha^2 = \delta \alpha - L' \alpha^2 > 0$$

for all sufficiently small positive α. Here L' is the Lipschitz constant of g in a neighborhood of the point (x, y, v). This contradicts assumption C, since $\alpha l \in N^{\perp}_{K(x)}(y)$.

Now, let (2.6) be fulfilled. Then for every x, y, v, l chosen as in assumption C, we have $\langle g(x, y, v), l \rangle \leq 0$. Applying assumption D for $y_1 = y \in K(x)$ and $y_2 = y_1 + l$ we obtain

$$\langle g(x, y + l, v), l \rangle \leq \langle g(x, y + l, v), l \rangle - \langle g(x, y, v), l \rangle$$

$$= \langle g(x, y_2, v) - g(x, y_1, v), y_2 - y_1 \rangle \leq -\lambda |l|^2.$$

This completes the proof. ∎

3. The regular case. In this section we obtain conditions for regularity of the sensor model. By regularity we mean the following: there exists a small, that is, proportional to ε, modification of the static sensor model (1.5), obtained by neglecting the sensor dynamics, which provides correct results when used for deterministic parameter identification. Specifically, we answer the question under what circumstances assumptions B and C are satisfied by the equilibrium mapping $K_0(\cdot)$ defined in (1.7).

First, note that $K_0(x)$ is a closed set, provided that V is compact. Further, the Lipschitz-type assumption in B for K_0 is standard in the set-valued framework, for sufficient conditions; see [2]. From [14] it follows that if assumption D is fulfilled, then

$K_0(x)$ is nonempty for every $x \in S$. The remaining condition for the initial value of the observation is a natural requirement, meaning that at the beginning of the process the sensors are in equilibrium ($\dot{y} = 0$). An alternative condition is to suppose that some interval $[0, \delta], \delta > 0$, does not contain points from the set Δ.

Thus, the critical requirement for regularity is represented by assumption C.

In the following we suppose that assumption D holds. Then (2.6) is equivalent to C, according to Proposition 2.3. We focus on (2.6) in the special case where the function g has the form

$$g(x, y, v) = g_0(x, y) + C(x)v.$$

Here $g_0 : S \times \mathbb{R}^m \to \mathbb{R}^m$ and $C : S \to \mathbb{R}^{n\times r}$ are Lipschitz continuous functions with respect to x, uniformly in y. We suppose that for all $x \in S$ the function $g_0(x, \cdot)$ is invertible, differentiable, and with locally Lipschitz derivative, and the derivative $\partial g_0/\partial y(x, y)$ is invertible for all y. We also suppose that the set V is convex and compact.

PROPOSITION 3.1. *Under the assumptions listed in the above paragraph, the map $K_0(\cdot)$ is nonempty and compact-valued and Lipschitz continuous. Moreover, assumption* C *is equivalent to the following:*

$$-\frac{\partial g_0}{\partial y}(x, \varphi(x, v))C(x)T_V(v) \subset C(x)T_V(v) \quad \textit{for all } x \in S, \textit{ for all } v \in V, \tag{3.1}$$

where $\varphi(x, v)$ is the unique solution of the equation $0 = g_0(x, y) + C(x)v$ and $T_V(v)$ is the tangent cone to V at v.

Proof. We already know that the set $K_0(x)$ is closed and nonempty. Since $K_0(x) = g_0^{-1}(x, -C(x)V)$, where the inversion is with respect to the second argument, the first claim follows from the classical inverse function theorem.

By Proposition 2.3, assumption C is equivalent to (2.6). According to the proximal characterization of the strong invariance property with respect to a differential inclusion obtained in [9] (see also [3]) and the tangential characterization of the same property [1], the relation (2.6) (for the set $K_0(x)$) is equivalent to

$$g(x, \bar{y}, V) \subset T_{K_0(x)}(\bar{y}) \quad \text{for all } x \in S,\ \bar{y} \in K_0(x). \tag{3.2}$$

Since x is fixed, we skip the dependence of x in the rest of the proof.

Let $\bar{v} \in V$ and $\bar{y}$ be arbitrary vectors satisfying the relation $0 = g(\bar{y}, \bar{v})$, that is, $\bar{y} = \varphi(\bar{v})$. Then (3.2) can be rewritten as $C(V - \bar{v}) \subset T_{K_0(x)}(\bar{y})$ or, equivalently, as

$$CT_V(\bar{v}) \subset T_{K_0(x)}(\bar{y}) \quad \text{for all } \bar{v} \in V.$$

It remains to prove that

$$T_{K_0}(\bar{y}) = -\left[\frac{\partial g_0}{\partial y}(\bar{y})\right]^{-1} CT_V(\bar{v}). \tag{3.3}$$

The inclusion of $T_{K_0}(\bar{y})$ in the right-hand side can be verified by applying the definition of the contingent cone. To prove the inverse inclusion we first define the function

$$\omega_h(\eta) \overset{def}{=} \frac{g_0(\bar{y} + h\eta) - g_0(\bar{y})}{h} - \frac{\partial g_0}{\partial y}(\bar{y})\eta,$$

where h is a positive parameter and $\eta \in \mathbb{R}^m$. A standard estimation implies that $\omega_h(\cdot)$ is Lipschitz continuous with Lipschitz constant $L'h/2$, where L' is the Lipschitz constant of $\frac{\partial g_0}{\partial y}(\cdot)$ at $\bar{y}$. Then for every fixed $\xi \in \mathbb{R}^m$ with $|\xi| \leq 1$ and for all sufficiently small h, the operator

$$\eta \mapsto \xi - \left[\frac{\partial g_0}{\partial y}(\bar{y})\right]^{-1} \omega_h(\eta)$$

maps $2\mathcal{B}$ into $2\mathcal{B}$ and is contractive; it therefore has a fixed point $\hat{\eta}_h(\xi)$ and obviously

$$|\hat{\eta}_h(\xi) - \xi| \leq \frac{L'}{2} \left| \left[\frac{\partial g_0}{\partial y}(\bar{y})\right]^{-1} \right| h. \tag{3.4}$$

Now take an arbitrary element ξ, $|\xi| \leq 0.5$ from the right-hand side of (3.3). There exist a sequence $h_k \to 0$ and a sequence $v_k \in V$ such that

$$\xi = -\left[\frac{\partial g_0}{\partial y}(\bar{y})\right]^{-1} C \lim_k \frac{v_k - \bar{v}}{h_k}.$$

Then $\xi_k \to \xi$ for $\xi_k \stackrel{def}{=} -[\frac{\partial g_0}{\partial y}(\bar{y})]^{-1} C(v_k - \bar{v})/h_k$. Since $|\xi_k| \leq 1$ for all sufficiently large k we can define $\eta_k = \hat{\eta}_{h_k}(\xi_k)$. We have

$$h_k \frac{\partial g_0}{\partial y}(\bar{y})\xi_k = -C(v_k - \bar{v}) = -g_0(\bar{y}) - Cv_k.$$

From the definition of η_k, we obtain

$$g(\bar{y} + h_k\eta_k, v_k) = g_0(\bar{y}) + h_k \frac{\partial g_0}{\partial y}(\bar{y})\eta_k + h_k\omega_{h_k}(\eta_k) + Cv_k$$

$$= g_0(\bar{y}) + h_k \frac{\partial g_0}{\partial y}(\bar{y})\xi_k + Cv_k = 0.$$

This means that $\xi \in T_{K_0}(\bar{y})$ since $\eta_k \to \xi$ according to (3.4). The proof is now complete. □

To better understand condition (3.1) we denote $A = (\partial g_0/\partial y)(x, y)$, $C = C(x)$ and consider the case where V is a box, $V = [\alpha_1, \beta_1] \times \cdots \times [\alpha_r, \beta_r]$.

PROPOSITION 3.2. Condition (3.1) is equivalent to the following: each column of the matrix C is an eigenvector of the matrix A corresponding to an eigenvalue that is real and nonpositive.

Proof. Let (3.1) be fulfilled. We shall prove the claim for the first column of C. The case $r = 1$ is obvious; therefore we suppose that $r > 1$.

Denote by $e_i \in \mathbb{R}^r$ the vector with the ith component equal to one and all the other components equal to zero. Then $e_1 \in T_V((\alpha_1, v_2, \ldots, v_r))$ for every $v_i \in [\alpha_i, \beta_i]$. Condition (3.1) implies that $l \stackrel{def}{=} -ACe_1 \in CT_V((\alpha_1, v_2, \ldots, v_r))$. Hence, given $\sigma = (\sigma_2, \ldots, \sigma_r)$ with all $\sigma_i \in \{-1, 1\}$, by choosing appropriate $v_2, \ldots, v_r$ we conclude that l has the representation $l = Cz_\sigma$, where $\langle e_1, z_\sigma \rangle \geq 0$ and $\operatorname{sign}\langle e_i, z_\sigma \rangle = \sigma_i$ for $i = 2, \ldots, r$. Taking an arbitrary $\sigma' = (\sigma_3, \ldots, \sigma_r)$ (if $r > 2$) and a convex combination of $z_{(-1,\sigma')}$ and $z_{(1,\sigma')}$, we can represent l as $l = Cz'_{\sigma'}$, where now $\langle e_1, z'_{\sigma'} \rangle \geq 0$, $\langle e_2, z'_{\sigma'} \rangle = 0$, and $\operatorname{sign}\langle e_i, z'_{\sigma'} \rangle = \sigma'_i$ for $i = 3, \ldots, r$ (if $r > 2$). Proceeding by induction we obtain the representation $l = Cz$, where $\langle e_1, z \rangle \geq 0$ and $\langle e_i, z \rangle = 0$ for

$i = 2, \dots, r$. That is, $l = -AC_1 = \gamma C_1$, where C_1 is the first column of C and $\gamma \geq 0$. This completes the proof of the necessity. The proof of the sufficiency is straightforward. □

Let us consider again the example in the introduction. Here $r = 1$ and

$$A = \begin{pmatrix} -1 & 0 \\ 0 & -2 \end{pmatrix}, \quad C = \begin{pmatrix} 1 \\ 2 \end{pmatrix}.$$

The vector C is not an eigenvector; therefore, as was argued in the introduction, the dynamic sensor model (1.9)–(1.10) is not regular. The regularity condition would be fulfilled, however, if the second sensor (1.10) were identical to the first one.

Let us consider the same example, but for a moment forget that the disturbance $v \in [-1, 1]$ is the same for the two sensors. That is, we pass to the case $r = 2$ with

$$C = \begin{pmatrix} 1 & 0 \\ 0 & 2 \end{pmatrix}.$$

Here the two columns of C are obviously eigenvectors of A corresponding to the eigenvalues -1 and -2, respectively; therefore the use of the static sensor model

$$y_1 = x_1 + v_1, \quad y_2 = x_2 + v_2, \quad v_1, v_2 \in [-1 - c\varepsilon, 1 + c\varepsilon]$$

is justified. This model, however, is not minimal; it is "too pessimistic." In the next section we show a way to find a minimal static model providing correct results for the parameter identification problem.

4. The singular case. In the preceding section we showed that the static sensor model (1.6) obtained by simply neglecting the sensor dynamics can be applied for deterministic estimation of the parameter p in very special cases only. Specifically, for a linear time-invariant system, the condition described in Proposition 3.2 is not generic. In this section we show how to construct a static sensor model, represented by a set $K(x)$, larger than $K_0(x)$, which satisfies the requirements of Theorem 2.1.

First we recall some notions from [1]. A set $Y \subset \mathbb{R}^m$ is *strongly invariant* (merely "invariant" in the terminology in [1]) with respect to the differential inclusion $\dot{y} \in G(y)$ if every trajectory $y(\cdot)$ of this differential inclusion starting from a point of Y never leaves Y. For any set Y and for a mapping G that is Lipschitz continuous, there exists a unique minimal closed set—the so-called *invariance envelope* of Y (denoted $\mathrm{Inv}_G(Y)$) that contains Y and is strongly invariant with respect to G.

Going back to the identification problem, let us suppose that the measurement data $\hat{y}(t)$, $t \in \Delta$ are given, that assumptions A and D from section 2 are fulfilled, and that the equilibrium mapping $K_0(\cdot)$ satisfies assumption B. Denote $\bar{K}(x) \stackrel{def}{=} \mathrm{Inv}_{G(x,\cdot)}(K_0(x))$, where $G(x, y) = g(x, y, V)$. We also suppose that the function g has the following Lipschitz property with respect to x: there is a constant L_x such that

$$|g(x', y, v) - g(x'', y, v)| \leq L_x |x' - x''| \quad \text{for all } x', x'' \in S \text{ for all } y \in \bar{K}(x').$$

THEOREM 4.1. *Let the conditions listed in the last paragraph hold. Then $\bar{K}$ is the unique minimal (with respect to pointwise inclusion) mapping from S to the subsets of $\mathbb{R}^m$ satisfying assumptions* B *and* C *such that $K_0(x) \subset \bar{K}(x)$ for all $x \in S$.*

Proof. For $x \in S$ the set $\bar{K}(x)$ is closed by definition and nonempty, since $K_0(x) \subset \bar{K}(x)$ is nonempty according to [14]. Let us prove the Lipschitz property.

Let $x', x'' \in S$, $y'' \in \bar{K}(x'')$. We prove that

$$\operatorname{dist}(y'', \bar{K}(x')) \le L + \frac{L_x}{\lambda}. \tag{4.1}$$

The invariance envelope $\operatorname{Inv}_{G(x,\cdot)}(K_0(x))$ is the closure of the set of all points $y(t)$, $t \ge 0$, where $y(\cdot)$ is a trajectory of $\dot{y} \in G(x, y)$ starting from a point of $K_0(x)$. Then for an arbitrary $\delta > 0$ there exist $y_0'' \in K_0(x'')$, $T > 0$, and a measurable $v(\cdot) : [0, T] \mapsto V$ such that the (unique) solution y_2 of the equation

$$\dot{y}(t) = g(x'', y(t), v(t)), \quad y(0) = y_0'', \tag{4.2}$$

satisfies $|y'' - y_2(T)| \le \delta$. There is $y_0' \in K_0(x')$ such that $|y_0' - y_0''| \le L|x' - x''|$. Denote by $y_1(\cdot)$ the solution of (4.2) with x', y_0' instead of x'', y_0''. Then $y_1(T) \in \bar{K}(x')$, hence $\operatorname{dist}(y'', \bar{K}(x')) \le |y_1(T) - y_2(T)| + \delta$. Denote $d(t) = |y_1(t) - y_2(t)|$. Whenever differentiable and nonzero, $d(\cdot)$ satisfies

$$\dot{d}(t) = \frac{\langle \dot{y}_1(t) - \dot{y}_2(t), y_1(t) - y_2(t) \rangle}{d(t)}$$

$$= \frac{\langle g(x', y_1(t), v(t)) - g(x'', y_2(t), v(t)), y_1(t) - y_2(t) \rangle}{d(t)} \le L_x|x' - x''| - \lambda d(t).$$

Here we used the Lipschitz property of g and assumption D. From here we conclude that $d(t) \le (L + L_x/\lambda)|x' - x''|$ which implies (4.1) since δ was arbitrarily chosen.

The condition $\hat{y}(0) \in \bar{K}(x_0)$ is fulfilled since it is assumed for $K_0(x_0)$. It remains to show that assumption C holds. Under D, assumption C is equivalent to (2.6), according to Proposition 2.3. From [3, 9], the latter is equivalent to the strong invariance of $\bar{K}(x)$. In addition, $\bar{K}(x)$ is the smallest strongly invariant set containing $K_0(x)$. This completes the proof. □

Since the set $K_0(x)$ consists only of equilibrium points of the mapping $G(x, \cdot)$, its invariant envelope can be described as the limit (in the sense of Painlevé–Kuratowski)

$$\bar{K}(x) = \lim_{T \to +\infty} R_x(T),$$

where $R_x(T)$ is the reachable set at T of the system

$$\dot{y}(t) = g(x, y(t), v(t)), \quad y(0) \in K_0(x), \quad v(t) \in V.$$

The reachable set can be analytically found only in exceptional cases (like in the example below), but there are efficient methods for its enclosure and approximation; see, e.g., the survey [11].

Let us compute the set $\bar{K}(x)$ for the example considered in the introduction and in section 3. Easy calculations give us

$$R_x(t) = e^{\begin{pmatrix} -t & 0 \\ 0 & -2t \end{pmatrix}} K_0(x) + \int_0^t e^{\begin{pmatrix} -t+s & 0 \\ 0 & -2t+2s \end{pmatrix}} \left(\begin{pmatrix} x_1 \\ 2x_2 \end{pmatrix} + \begin{pmatrix} 1 \\ 2 \end{pmatrix} [-1, 1] \right) ds.$$

Integrating and taking the limit for $t \to +\infty$ we get

$$\bar{K}(x) = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} + \int_0^{+\infty} \begin{pmatrix} e^{-s} & 0 \\ 0 & e^{-2s} \end{pmatrix} \begin{pmatrix} 1 \\ 2 \end{pmatrix} [-1, 1]\, ds.$$

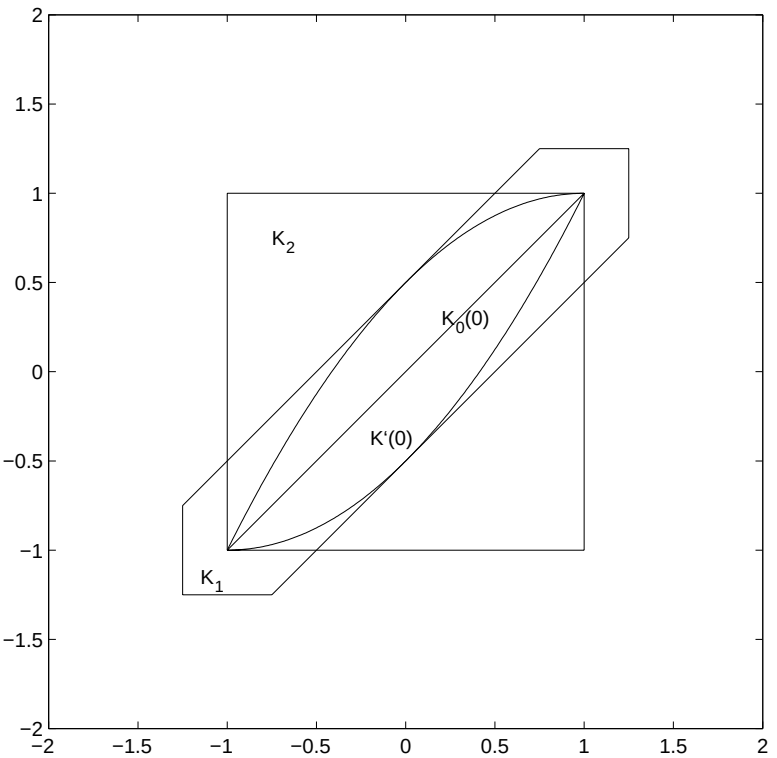


FIG. 1. *Various models of the static sensor.*

The last integral of a set-valued mapping is in the sense of Aumann; see, e.g., [2]. Its calculation gives the set $R = \{(y_1, y_2) \in \mathbb{R}^2 |\ 0.5y_1^2 + y_1 - 0.5 \le y_2 \le -0.5y_1^2 + y_1 + 0.5\}$. Finally we have

$$\bar{K}(x) = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} + R.$$

The set $\bar{K}(x)$ is given on Figure 1 for $x = 0$. Note the difference with the formally obtained static model $K_0(0)$ as well as with the two regularizations, K_1 as discussed in the introduction and K_2 in section 3.

REFERENCES

[1] J.-P. AUBIN, *Viability Theory*, Birkhäuser, Basel, Boston, 1991.
[2] J.-P. AUBIN AND H. FRANKOWSKA, *Set-Valued Analysis*, Birkhäuser, Basel, Boston, 1990.
[3] F. CLARKE, Y. LEDYAEV, R. STERN, AND P. WOLENSKI, *Nonsmooth Analysis and Control Theory*, Springer-Verlag, New York, 1998.
[4] P. CANNARSA AND C. SINESTRARI, *Convexity properties of the minimum time function*, Calc. Var. Partial Diff. Equations, 3 (1995), pp. 273–298.
[5] M. CORLESS, G. LEITMANN, AND E. O. RYAN, *Control of uncertain systems with neglected dynamics*, in Deterministic Control of Uncertain Systems, IEE Control Engrg. Ser. 40, A. S. I. Zinober, ed., Pergamon, Oxford, 1990, pp. 252–268.
[6] A. L. DONTCHEV AND I. SLAVOV, *Singular perturbation in a class of nonlinear differential inclusions*, in System Modelling and Optimization, Lecture Notes in Control and Inform. Sci. 143, H.-J. Sebastian and K. Tammer, eds., Springer-Verlag, New York, 1991, pp. 273–280.

[7] A. L. Dontchev and V. M. Veliov, *Singular perturbations in Mayer's problem for linear systems*, SIAM J. Control Optim., 21 (1983), pp. 566–581.

[8] S. R. Kenneth and D. G. Taylor, *Discrete-time observers for singularly perturbed continuous-time systems,* IEEE Trans. Automat. Control, 40 (1995), pp. 224–235.

[9] M. Krastanov, *Forward invariant sets, homogeneity and small-time controllability*, in Geometry in Nonlinear Control and Differential Inclusions 32, Banach Center Publications, Warsaw, 1995, pp. 287–300.

[10] G. Leitmann, E. P. Ryan, and A. Steinberg, *Feedback control of uncertain systems: Robustness with respect to neglected actuator and sensor dynamics*, Internat. J. Control, 43 (1986), pp. 1243–1256.

[11] F. Lempio and V. M. Veliov, *Discrete approximations to differential inclusions*, Mitteilungen der GAMM, 21 (1998), pp. 103–135.

[12] M. Milanese, J. Norton, H. Piet-Lahanier, and E. Walter, eds., *Bounding Approaches to System Identification*, Plenum Press, New York, 1996.

[13] V. M. Veliov, *Attractiveness and invariance: The case of uncertain measurement*, in Modeling Techniques for Uncertain Systems, Progr. Systems Control Theory 18, A. Kurzhanski and V. M. Veliov, eds., Birkhäuser, Basel, Boston, 1994, pp. 277–288.

[14] V. M. Veliov, *Generalization of the Tikhonov theorem for singularly perturbed differential inclusions*, J. Dynam. Control Systems, 3 (1997), pp. 291–319.

SIAM J. CONTROL OPTIM.
Vol. 38, No. 5, pp. 1323–1333

ADMISSIBILITY OF CONTROL OPERATORS FOR SOLUTION FAMILIES TO VOLTERRA INTEGRAL EQUATIONS*

MICHAEL JUNG†

Abstract. We reduce the admissibility of finite-dimensional control operators for an evolution system (satisfying certain mild regularity conditions) to that for a semigroup control system, for which there are plenty of results available in the literature. One such necessary and sufficient condition is the Carleson measure criterion. We prove this condition explicitly for evolution systems without the regularity assumption mentioned above. We also look at some examples which show that our respective hypotheses are necessary.

Key words. admissibility, evolutionary equations, control theory, Carleson measure, completely positive

AMS subject classifications. 45K05, 45N05, 93B99

PII. S0363012997328191

1. Introduction. This paper deals with control theory for nonautonomous, infinite-dimensional dynamical systems which arise from Volterra integral equations. Classical theory for C_0-semigroups considers the (autonomous) system

$$\begin{aligned} x'(t) &= Ax(t) + Bu(t), \qquad t \geq 0, \\ x(0) &= x_0, \end{aligned} \tag{1}$$

where x is a function with values in the state space X, u is a function with values in the control space, and B is the control operator. In order that the uncontrolled system has a solution, we assume that A is the generator of a C_0-semigroup $T(\cdot)$ in X. Particularly in boundary control problems unbounded control operators arise. A detailed example is given in [6].

We will be interested in the case where the control space is one-dimensional. Any finite-dimensional control operator can be obtained as a sum of one-dimensional operators. Consequently, the solution to (1) will be obtained as a sum if the individual solutions exist. We will write b instead of B, where b is a nonzero element of the range space of B.

This equation was studied in [7] and [10] and later generalized to B of higher rank in [4]. Interesting cases arise when B is unbounded, in other words, when B cannot be identified by an element b of its range space in X, but only in a larger space; see also [2]. This paper also deals with the preliminary question of which operators B are admissible, and not which operators B make the system controllable [3]. Necessary and sufficient criteria were established for admissibility of the control element b in the C_0-semigroup case in the aforementioned papers.

We are interested in a generalized version of this setting, namely, the generalized (Volterra) system:

$$\begin{aligned} x'(t) &= (da * Ax)(t) + bu(t), \\ x(0) &= x_0 \end{aligned}$$

*Received by the editors October 6, 1997; accepted for publication May 5, 1999; published electronically May 11, 2000. This research was supported by a Deutsche Forschungsgemeinschaft Fellowship.

http://www.siam.org/journals/sicon/38-5/32819.html

†Fakultät für Mathematik, Technische Universität Berlin, Straße des 17. Juni 135, 10623 Berlin (mjung@math.tu-berlin.de).

for a a scalar-valued function of bounded variation. Integrating these equations with respect to time, we are thus led to consider the following problem:

$$x(t) = x_0 + (a * Ax)(t) + (1 * bu)(t), \qquad t \geq 0. \tag{2}$$

We wrote the equation in a derivative-free way. This will be of convenience later on and indeed is slightly more general with the appropriate requirement on a than had we chosen to write it in its original setting (1).

It is the purpose of this paper to determine necessary and sufficient conditions under which b will be a suitable control element for (2) in an L^2-setting. In section 3 we will reduce the admissibility for an evolution system to that for a semigroup system under certain regularity conditions. This, in particular, extends the known Carleson measure criterion. In section 2 we will prove this with less regularity needed. These will depend on A and a, of course. The corresponding problem for semigroups was considered in [7] and [10]. We will see that the semigroup conditions remain valid if a is completely positive. Completely positive functions were studied intensively in [9]. Consequently, many control operators known for semigroups can also be used in the Volterra integral equation case.

Let us start by presenting the terminology used in this setting: $\mathbb{R}_+ = (0, \infty)$ and $\mathbb{C}_+ = \{\lambda \in \mathbb{C} : \Re(\lambda) > 0\}$. Assume that X is the (Banach) state space and $u \in L^2(\mathbb{R}_+)$. We have a complete solution theory for the uncontrolled equation in the case that A is a closed operator, a is of subexponential growth, and some compatibility conditions are met (see [9, section 1]). For the time being, we assume these conditions are met. If $S(\cdot)$ is the solution family to the uncontrolled equation (mapping x_0 onto the solution), the (mild) solution to (2) is formally obtained via the variation of constants formula and can be expressed as

$$x(t) = \frac{d}{dt}(S * (x_0 + 1 * bu))(t) = S(t)x_0 + (S * bu)(t), \qquad t \geq 0.$$

We therefore call b admissible if the above equation can be made rigorous. (We understand that to mean that b is an admissible control element.) Let X^{-1} be the completion of X with the norm $\|(\lambda I - A)^{-1} \cdot\|$, where λ is in the resolvent set $\rho(A)$; different values of $\lambda \in \rho(A)$ yield equivalent norms. S extends to a solution family in that space. We assume in the present paper that $b \in X^{-1}$. In the semigroup case it was proven (see [11]) that this has to be the case if b is to be admissible.

DEFINITION 1.1. *If there exist a $t_0 > 0$ and $c > 0$, such that $\|(S*bu)(t_0)\| \leq c\|u\|_2$ for all $u \in L_2([0, t_0])$, then b is called* admissible *(for time $t_0 > 0$).*

This is a direct generalization of the standard definition, e.g., given in [11]. Note that if this definition is true for $t_0 > 0$, then it is also true for all $0 < t < t_0$, while it need not be true for $t > t_0$ (in contrast to the semigroup case). This definition is unaffected by a reflection of u about the ordinate, so we may integrate over the product Sbu where convenient, instead of calculating the convolution. The set X^S denotes the set of admissible vectors that are admissible for all times.

We make some notational conventions. Functions defined on some subset of the real axis are trivially extended without stating this fact each time. For functions $f \in L^p(\mathbb{R})$ and $g \in BV(\mathbb{R})$, we have for their convolution $f * dg \in L^p(\mathbb{R})$. A complex-valued measure $\mu \in \mathcal{M}(\mathbb{R}_+)$ is called *Laplace transformable* if

$$\int_0^\infty e^{-\epsilon t} |d\mu(t)| < \infty$$

for some $\epsilon > 0$. We define the Laplace transform $\mathcal{L}$ for all Laplace transformable measures

$$\mathcal{L}(d\mu)(z) = \int_0^\infty e^{-zt} d\mu(t), \qquad \Re(z) > \epsilon. \tag{3}$$

We extend the transform of a function analytically in the component of $\mathbb{C}$ which contains (ϵ, ∞) and denote this extension with the same symbol. The Laplace transform $\mathcal{L}$ and the Fourier transform $\mathcal{F}$ are related via $\mathcal{L}(d\mu)(iz) = \mathcal{F}(d\mu)(z)$ for appropriate transformable measures $\mu \in \mathcal{M}(\mathbb{R}_+)$ and $z \in \mathbb{R}$. Let

$$e_0(t) = \begin{cases} 1 & \text{for } t > 0, \\ 0 & \text{for } t \leq 0, \end{cases}$$

the Heaviside function.

We recall a definition from [1]. A function $a \in L^1_{\mathrm{loc}}(\mathbb{R}_+)$ is called *completely positive* if there exists a nonnegative, nondecreasing, and concave function $k : \mathbb{R}_+ \to \mathbb{R}$ such that

$$z\mathcal{L}(a)(z)\mathcal{L}(dk)(z) = 1, \qquad z \geq 0. \tag{4}$$

A measure $\mu \in \mathcal{M}(\mathbb{R}_+)$ is called completely positive if it is Laplace transformable and $z\mathcal{L}(d\mu)(z)\mathcal{L}(dk)(z) = 1$, accordingly. The function k can be decomposed as

$$k(t) = k_0 + k_\infty t + \int_0^t k_1(\tau) d\tau, \qquad t > 0, \tag{5}$$

with k_1 nonnegative, nonincreasing, and $k_1(\infty) = 0$. See section 4 in [9] for more details on completely positive functions.

A family S is called *solution family* to (2) if it is strongly continuous on X and on $X_1 := (D(A), \|\cdot\| + \|A\cdot\|)$, $S(0) = I$ and satisfies $S(t)x = x + A(a * S)(t)x$ for $x \in X$; in particular, $(a * S)(t)x \in X_1$.

2. Admissibility for diagonal operators. We restrict ourselves to let a be a completely positive, Laplace transformable function and A be a diagonal operator in this section. More specifically, let $X = l^2(\mathbb{N})$ and $Ae_n = \lambda_n e_n$ for e_n the nth unit vector. We also assume, without loss of generality, that $\sigma := \sup_{n\in\mathbb{N}} \Re(\lambda_n) < 0$. There are no requirements with regard to accumulation points. Many operators can be represented in this way; the most important class is normal generators with a compact resolvent in a Hilbert space. Then (2) has a unique strongly continuous diagonal solution family $S(t)e_n = s_{-\lambda_n}(t)e_n$ (see [9, section 4]). The scalar function s_λ satisfies the equation

$$s_\lambda(t) = 1 + \lambda(a * s_\lambda)(t), \qquad t \geq 0.$$

We now introduce the notion of a Carleson measure (see [8, chapter 8], [7]). Note, that the element b can be interpreted as a vector $(b_n)_n$, although generally not in $l^2(\mathbb{N})$. The vector b has an associated measure μ_b, where $\mu_b(M) = \sum_{-\lambda_n \in M} |b_n|^2$, which depends on $(\lambda_n)_n$; we will also denote that measure with b, in the hope of not having introduced too much ambiguity.

Definition 2.1. *Let $R(h,k) = \{z \in \mathbb{C}_+ : 0 < \Re(z) < h, |\Im(z) - k| < h\}$. Then b is called a* Carleson measure*, if there exists a $M > 0$, such that for all $h > 0$ and $k \in \mathbb{R}$: $\sum_{-\lambda_n \in R(h,k)} |b_n|^2 \leq Mh$.*

We may now formulate and prove the main theorem of this section, giving a necessary condition for admissibility of (2) in the sense of Definition 1.1. A sufficient condition is given in the next section (Theorem 3.2). The condition here is essentially the same—the Carleson measure criterion—as in the semigroup case. It has been proved that the Carleson measure criterion is equivalent to admissibility for diagonal, analytic, and invertible semigroups (see, e.g., [11]). The new results encompass the semigroup case, but the theorem below yields the criterion's necessity without the restrictions given in the next section for general operators (Theorem 3.3), whereas it is sufficient for admissibility, as we shall also see in the next section. Why the restrictions $\kappa > 0$ and $\alpha < \infty$ (see below) are required will be seen from the examples; what proved to be a necessary and sufficient criterion in the semigroup case cannot be applied without precaution in the Volterra integral equation case.

We assume the same properties on A and a as before. We define three characteristic values for a. Let $k : \mathbb{R}_+ \to \mathbb{R}$ be the function related with a as in (4). Then

$$\begin{aligned}\kappa &:= \lim_{t\to 0} k(t) = k_0,\\ \omega &:= \lim_{t\to\infty} k(t)/t = k_\infty,\\ \alpha &:= \lim_{t\to 0} k'(t) = k_1(0+) + k_\infty.\end{aligned}$$

Note that all these values are nonnegative but α may be infinite. The limit (of the increasing sequence) for α need not exist.

We use the subordination principle to represent s_λ: there exists a function $w : \mathbb{R}_+^2 \to \mathbb{R}_+$, such that

$$\begin{aligned}s_\lambda(t) &= -\int_0^\infty e^{\lambda s} d_s w(t,s)\\ &= -\mathcal{L}(d_\cdot w(t,\cdot))(-\lambda), \qquad t \geq 0,\ \Re(\lambda) < 0.\end{aligned}$$

(See [9, Theorem 4.1 and Corollary 4.5].) The function w has certain properties we will now exploit.

We set $w_0(t,s) = w(t,s) - e_0(t - \kappa s)$, $t, s \geq 0$. We state the following lemma without proof. It is not hard to check, given the properties of w listed in [9, section 4].

LEMMA 2.2. *If $\alpha < \infty$, then $\lim_{(t,s)\to 0} w_0(t,s) = w_0(0,0) = 0$ and for each $\epsilon > 0$ there exists $t_0 > 0$, such that for all $0 < t \leq t_0$ holds $\|w_0(t,\cdot)\|_{BV[0,t_0]} < \epsilon$.*

THEOREM 2.3. *Let $\kappa > 0, \alpha < \infty$. If b is admissible for $T > 0$, then b is a Carleson measure.*

Proof. We have already stated that admissibility holds for all $0 < t_0 \leq T$. We shall prove that there exists $t_0 > 0$ and $m > 0$, such that for all $h > 0$ and $k \in \mathbb{R}$ there exists $u \in L^2[0,t_0]$ with $\|u\|_2 = 1$, such that

$$\left|\int_0^{t_0} s_\lambda(t)u(t)dt\right| = \left|\int_0^{t_0}\int_0^\infty e^{\lambda s} d_s w(t,s)u(t)dt\right| \geq m\, h^{-1/2}$$

for $-\lambda \in R(h,k) \cap \sigma(-A)$. This is sufficient to prove the claim, because

$$\sum_{-\lambda_n \in R(h,k)} |b_n|^2 \leq \sum_{-\lambda_n \in R(h,k)} \left|b_n \int_0^{t_0} s_{\lambda_n}(t)u(t)dt\right|^2 / (m^2 h^{-1})$$

$$\begin{aligned}
&\le m^{-2}h\sum_{n\in\mathbb{N}}\left|b_n\int_0^{t_0}s_{\lambda_n}(t)u(t)dt\right|^2\\
&\le m^{-2}h\left\|\int_0^{t_0}S(t)bu(t)dt\right\|^2\\
&\le cm^{-2}\,h.
\end{aligned}$$

Therefore let $\epsilon=\inf_{z\in R(0,1)}|z^{-1}(1-e^{-z})|$ and choose $0<t_0\le T$, such that

$$\sup_{t\in[0,t_0]}\|w_0(t,\cdot)\|_{BV[0,t_0]}<\epsilon/2.$$

Then let $\widetilde{u}(t)=e_0(\kappa/h-t)e^{ikt/\kappa}$ and $u(t)=\widetilde{u}(t)/\|\widetilde{u}\|_2=h^{-1/2}\kappa^{1/2}\widetilde{u}(t)$. We then obtain with $w_0(t,s)=w(t,s)-e_0(t-\kappa s)$, $t,s\ge 0$, from Lemma 2.2

$$\begin{aligned}
\left|\int_0^{t_0}s_\lambda(t)\widetilde{u}(t)dt\right|&=\left|\int_0^{t_0}\int_0^{\infty}e^{\lambda s}d_sw(t,s)\widetilde{u}(t)dt\right|\\
&\ge\left|\int_0^{t_0}e^{\lambda t/\kappa}\widetilde{u}(t)dt\right|-\frac{\epsilon}{2}\int_0^{t_0}|\widetilde{u}(t)|dt\\
&\ge\left|\int_0^{\kappa/h}e^{(\lambda+ik)t/\kappa}dt\right|-\frac{\epsilon}{2}\int_0^{\kappa/h}dt\\
&=\left|\kappa\frac{1-e^{(\lambda+ik)/h}}{\lambda+ik}\right|-\frac{\epsilon\kappa}{2h}\\
&\ge\frac{\epsilon\kappa}{2h}.
\end{aligned}$$

We have silently assumed $\kappa/h\le t_0$. If this does not hold, the corresponding claim holds with t_0 replacing κ/h. Since all λ_n fulfill $\Re(\lambda_n)>-\sigma$, we can safely assume $h\ge\sigma$ and there exists a constant $c>0$, such that $\kappa/h\le ct_0$. We find

$$\left|\int_0^{t_0}s_\lambda(t)u(t)dt\right|\ge\min(c,1)\frac{\epsilon\kappa}{2h\|\widetilde{u}\|_2}\ge\min(c,1)\frac{\epsilon\kappa^{1/2}}{2}h^{-1/2}.$$

This finishes the proof. □

A converse of this theorem is shown in the next section (Theorem 3.2).

We now take a look at the situation in which $\kappa=0$. Then, if $\alpha<\infty$, we find by [9, p. 95, case 2] that $a(t)dt$ has a jump at zero. But this implies that a is not absolutely continuous. Consequently, a cannot be completely positive in that case. We are thus not able to use the subordination principle as we did in the theorems above. A minor extension is possible however. Although the sufficiency theorem, Theorem 2.3, can be proved, if a is just a (completely positive) measure, necessity is not true; the following provides a counterexample if $\kappa=0$.

Example 2.4. Let $Ae_n=(-1+in^{5/3})e_n$ and let $b\in X^{-1}$. Choose $a(t)=\delta_0$; then $a(t)dt$ is completely positive. (The corresponding k of (4) equals $k(t)=t$.) We have $w(t,s)=e^{-s}$ and $s_\lambda(t)=(1-\lambda)^{-1}$. Regarding admissibility, we obtain from (2) $x(t)=Ax(t)+(1*bu)(t)$, which yields $(S*bu)(t)=(I-A)^{-1}b(1*u)(t)$, which satisfies Definition 1.1 obviously (for $u\in L_2([0,T])$). Thus all b are admissible. But let $b_n=n\in\mathbb{N}$, for instance; then $(b_n)_n\in X^{-1}$. We have for $h=n^{5/3}>1$

$$h^{-1}\sum_{-\lambda_n\in R(h,\omega)}|b_n|^2=n^{-5/3}\sum_{j=0}^{n}j=n^{-5/3}(n^2-n)/2,$$

which is obviously not bounded, i.e., b is not a Carleson measure.

The next example shows that $\kappa > 0$ and $\alpha < \infty$ are a necessary requirement in Theorem 2.3.

Example 2.5. Let $Ae_n = -ne_n$ and let $b_n = n^\gamma$ for $n \in \mathbb{N}$ with $0 < \gamma < 1/2$. Choose $a(t) = (\pi t)^{-1/2}$ (the corresponding k of (4) equals $k(t) = 2(t/\pi)^{1/2}$); then we infer from [9, section 4] that

$$s_n(t) = \frac{2}{\sqrt{\pi}} e^{n^2 t} \int_{n\sqrt{t}}^{\infty} e^{-r^2} dr.$$

Let $0 < \beta < 1$. First we estimate the $L^2([0,T])$-norm of s_n:

$$\begin{aligned}
\|s_n\|_2^2 &= \frac{4}{\pi} \int_0^T \left(e^{n^2 t} \int_{n\sqrt{t}}^{\infty} e^{-r^2} dr \right)^2 dt \\
&\le \frac{4}{\pi} \int_0^T \left(e^{n^2 t} \int_{n\sqrt{t}}^{\infty} e^{-rn\sqrt{t}} dr \right)^{2\beta} \left(\int_{n\sqrt{t}}^{\infty} e^{n^2 t - r^2} dr \right)^{2-2\beta} dt \\
&\le \frac{4}{\pi} \int_0^T \left(\frac{1}{n^2 t} \right)^{\beta} \left(\int_0^{\infty} e^{-r^2} dr \right)^{2-2\beta} dt \\
&= C(T, \gamma, \beta) n^{-2\beta}.
\end{aligned}$$

Now we show that b is indeed admissible:

$$\begin{aligned}
\sum_{n=0}^{\infty} |b_n|^2 \left| \int_0^T s_{-n}(t) u(t) dt \right| &\le \sum_{n=0}^{\infty} |b_n|^2 \|s_{-n}\|_2^2 \|u\|_2^2 \\
&\le C(T, \gamma, \beta) \|u\|_2^2 \sum_{n=0}^{\infty} n^{2(\gamma - \beta)},
\end{aligned}$$

which is bounded if $\beta - \gamma > 1/2$. But b is not a Carleson measure

$$h^{-1} \sum_{-\lambda_n \in R(h,\omega)} |b_n|^2 = (1/n) \sum_{k=0}^{n} k^{2\gamma} = O(n^{2\gamma}), \quad n \to \infty,$$

for $h = n\,(n \in \mathbb{N})$. Moreover, by choosing γ, β appropriately, it is easy to see that b may lie outside any fractional power space between X and X^{-1}.

Moreover, we can improve the above example, if we choose $a(t) = t^\delta$ with $\delta < -1/2$. Since $S(\cdot)$ is analytic and $\|AS(\cdot)\| \in L_2(\mathbb{R}_+)$, we have $S * bu = AS * A^{-1}u \in C(\mathbb{R}_+, X)$ (see [9, Example 2.1 and Theorem 2.2]). Again, all $b \in X^{-1}$ are admissible.

We can improve the example in another direction and show that $\alpha < \infty$ is necessary. Choose $k(t) = 1 + 2(t/\pi)^{1/2}$, the corresponding a; then $\kappa > 0$ but $\alpha < \infty$. A similar but more lengthy calculation yields the same result as above. Note that

$$s_n(t) = n\pi^{-1} \int_0^{\infty} e^{-rt} r^{-1/2} (r + (r-n)^2)^{-1} dr$$

in this case.

3. Admissibility for general operators. For nondiagonal operators some significant results are available with respect to admissibility. In fact, admissibility is equivalent to $\|(\lambda - A)^{-1} b\|^2 \le K/|\Re(\lambda)|$ for some $K > 0$ in some right half-plane

in the case of normal or left-invertible semigroups (see [4]). For more results in that line, see [5]. However, a general criterion for admissibility seems not to be known even in the autonomous case. We will therefore try to relate the Volterra integral equation to the underlying Cauchy problem, thereby reducing the question to that problem. We again assume A is the generator of a C_0-semigroup in a Banach space X. Then it is known (for the semigroup case) that any admissible b is a vector in X^{-1} (see [11]). X^{-1} can also be viewed as the "smallest" space Y containing X, such that $A : D(A) \subset X \to Y$ has a continuous extension. Again we assume a is completely positive. Then the solution family to (2) for $b = 0$ can be represented similarly as in the section before as

$$S(t) = \int_0^\infty T(s) d_s w(t,s), \qquad t \geq 0, \tag{6}$$

where T is the semigroup generated by A.

The space of admissible vectors X^T for the semigroup case is continuously embedded in X^{-1}. If we can present $S(t)$ as a perturbation of $T(t)$ with the perturbing part mapping X^{-1} into X, we have equality for the space of admissible vectors, $X^T = X^S$. This idea is inspired by the fact that $S(t)$ is an integral of $T(\cdot)$ for $t \geq 0$ and integrals have smoothing properties. For this to work we need regularity of the involved integral kernels as we will see.

We thus need to decompose w in (6) into regular and singular parts. An informal calculation yields

$$S(t) = \int_0^\infty T(s) w_r(t,s) ds + \sum w_s^j(t) T(t_j),$$

where w_s^j gives the level of the jumps of $w(t,\cdot)$ at t_j. This is made precise in the following lemma.

LEMMA 3.1. *The function w can be decomposed as*

$$w(t,s) = e^{-\omega s} w_0(t - \kappa s, s) + e^{-\alpha s} e_0(t - \kappa s)$$

for $t, s > 0$. (a) *Let k_1 be the nonnegative, nonincreasing function given in* (5). *w_0 is nonnegative, $w_0(\cdot, s)$ is nondecreasing, $w_0(t,\cdot)$ is nonincreasing, and $w_0(0+, s) = 0$. If $k_1 \in W^{1,2}_{\mathrm{loc}}([0,\infty))$, then $w_0 \in W^{1,2}_{\mathrm{loc}}([0,\infty)^2)$.* (b) *If $k_1 \in W^{2,2}_{\mathrm{loc}}([0,\infty))$, then also $w_0 \in W^{2,2}_{\mathrm{loc}}([0,\infty)^2)$.*

Proof. Let $T > 0$; we will calculate all L^2-norms on $[0,T]$. From [9, Proposition 4.10] we infer that $w_0 \in W^{1,1}_{\mathrm{loc}}(\mathbb{R}^2_+)$. Note that we may set $\kappa = \omega = 0$, since this will not change the regularity. Moreover, let $l(t) = -\int_0^t s dk_1(s)$. Then $l'(t) = t k_1{}'(t)$ and

$$\partial_t w_0(t,s) = s t^{-1} \int_0^t l'(t-r) d_r w(r,s) + t^{-1} \lim_{h \to 0} h^{-1} \int_t^{t+h} l(t+h-r) d_r w(r,s).$$

To estimate the L^2-norm of the first term we use a simple convolution estimate and $|r/t| \leq 1$ for $r \in [0,t]$. But for the second term we get, using the fact that k_1 is nonincreasing, the following estimate:

$$\begin{aligned} &\left| \lim_{h \to 0} h^{-1} \int_t^{t+h} l(t+h-r) d_r w(r,s) \right| \\ &\leq \lim_{h \to 0} h^{-1} \int_0^h \int_0^{h-r} u |dk_1(u)| \, |d_r w(t+r,s)| \end{aligned}$$

$$\leq \lim_{h\to 0} \int_0^h |k_1(h-r) - k_1(0)|\, |d_r w(t+r,s)|$$
$$\leq \lim_{h\to 0} \sup_{s\in[0,h]} |k_1(s) - k_1(0)|.\|w(\cdot,s)\|_{BV} \;=\; 0.$$

On the other hand, we have

$$-\partial_s w_0(t,s) = (k_1 * \partial_t w_0)(t,s) + e^{-\alpha s}(k_1(t) - \alpha)$$

for almost all $t, s > 0$. We thus get for calculating the norm of w_0 in $W^{1,2}([0,T]^2)$

$$\|\partial_t w_0(\cdot,s)\|_2 \leq s\|{k_1}'\|_2 \|w(\cdot,s)\|_{BV} \leq s\|{k_1}'\|_2,$$
$$\|\partial_s w_0(\cdot,s)\|_2 \leq T^{1/2}\|k_1\|_2 \|\partial_t w_0(\cdot,s)\|_2 + e^{-\alpha s}\|k_1\|_2 + T^{1/2}\alpha e^{-\alpha s}$$

from which follows the claim.

We now calculate the second derivatives in case (b) for $t > 0$. First we get

$$s^{-1}\partial_t^2 w_0(t,s) = t^{-1}\int_0^t ({k_1}'(r) - rt^{-1}{k_1}'(r)) d_r w(t-r,s) \tag{7}$$
$$+ t^{-1}\int_0^t r{k_1}''(r) d_r w(t-r,s) \tag{8}$$
$$+ \lim_{h\to 0} h^{-1}\int_t^{t+h} l'(t+h-r) d_r w(r,s). \tag{9}$$

Note that from the first part of the proof we may infer with $k_1 \in W^{2,2}([0,T])$ we have

$$\|\partial_t w_0(\cdot,s)\|_\infty \leq s\|{k_1}'\|_\infty \|w(\cdot,s)\|_{BV} \leq s\|{k_1}'\|_\infty < \infty.$$

We estimate the second part of the right side of (7) (the first part is clearly bounded) and obtain

$$\int_0^T \left| \int_0^t rt^{-2}{k_1}'(t-r) d_r w(r,s) \right|^2 dt$$
$$= \int_0^T \left| \int_0^t rt^{-2}{k_1}'(t-r)\partial_r w_0(r,s) dr \right|^2 dt$$
$$\leq \int_0^T t^{-2} \int_0^t |{k_1}'(t-r)|^2 dr \int_0^t |\partial_r w_0(r,s)|^2 dr\, dt$$
$$< C\|k_1\|_{W^{2,2}}$$

(uniformly in $s \in [0,T]$). Now we calculate the desired bound for (8) using a standard estimate for convolutions.

$$\int_0^T \left| \int_0^t rt^{-1}{k_1}''(r) d_r w(t-r,s) \right|^2 dt$$
$$= \int_0^T \int_0^t |{k_1}''(r)|^2 d_r w(t-r,s) dt$$
$$\leq C\|{k_1}''\|_2.$$

We estimate (9) using the fact that $w(\cdot, s)$ is differentiable with bounded derivative for $t > 0$.

$$\lim_{h\to 0}\left|h^{-1}\int_t^{t+h}(t+h-r)k_1{}'(t+h-r)d_r w(r,s)\right| \le \lim_{h\to 0}\int_0^h |k_1{}'(h-r)||d_r w(t+r,s)| \;=\; 0.$$

Thus $\|\partial_t^2 w_0(\cdot,s)\|_2 \le 3Cs\|k_1\|_{W^{2,2}}$, which concludes the proof for $\partial_t^2 w_0$. We have for the remaining derivatives for almost all $t, s > 0$

$$-\partial_t\partial_s w_0(t,s) = (k_1 * \partial_t^2 w_0)(t,s) + e^{-\alpha s}k_1{}'(t) + \partial_t w_0(0,s)k_1(t),$$
$$-\partial_s^2 w_0(t,s) = (k_1 * \partial_s\partial_t w_0)(t,s) - \alpha e^{-\alpha s}(k_1(t) - \alpha).$$

Consequently,

$$\|\partial_t\partial_s w_0(\cdot,s)\|_2 \le \|k_1\|_1\|\partial_t^2 w_0(\cdot,s)\|_2 + e^{-\alpha s}\|k_1{}'\|_2 + \partial_t w_0(0,s)\|k_1\|_2,$$
$$\|\partial_s^2 w_0(\cdot,s)\|_2 \le \|k_1\|_1\|\partial_s\partial_t w_0(\cdot,s)\|_2 + \alpha e^{-\alpha s}(\|k_1\|_1 + \alpha T^{1/2}),$$

which concludes the proof for the remaining derivatives.

Note that we have omitted the proof of the measurability for the derivatives in question. This can be proven by standard mollifier arguments. □

We are now in a position to prove some relationships that hold between the spaces of admissible vectors X^T and X^S.

THEOREM 3.2. *Suppose that $a \in L^1_{\mathrm{loc}}(\mathbb{R}_+)$ is a completely positive function with $\kappa > 0$ and let $k_1 \in W^{1,2}_{\mathrm{loc}}([0,\infty))$. Then b is admissible for the Volterra case (i.e., (2)) if it is admissible for the autonomous case (i.e., (1)).*

Proof. $w(t,\cdot)$ can be decomposed according to Lemma 3.1 into a regular part with density $w_r(t,\cdot) \in L^2([0,t/\kappa])$ and a singular part $w_s(t,\cdot)$ with a jump at t/κ of level $e^{-\alpha t/\kappa}$. We thus may estimate two terms in the admissibility equation which correspond to the regular and singular parts, respectively. For the first we get (using $C_{T/\kappa}$ as the $L_2([0,T/\kappa])$-norm of $T(\cdot)b$)

$$\begin{aligned}
&\left\|\int_0^T\int_0^\infty T(s)bu(t)w_r(t,s)ds\,dt\right\| \\
&= \left\|\int_0^T\int_0^{t/\kappa} T(s)bu(t)w_r(t,s)ds\,dt\right\| \\
&= \left\|\int_0^{T/\kappa} T(s)b\int_{\kappa s}^T u(t)\,w_r(t,s)dt\,ds\right\| \\
&\le C_{T/\kappa}\left\|\int_{\kappa(\cdot)}^T u(t)w_r(t,\cdot)dt\right\|_2 \\
&\le C_{T/\kappa}\left(\int_0^T\left\|\int_{\kappa s}^T u(t)w_r(t,s)dt\right\|^2 ds\right)^{1/2} \\
&\le C_{T/\kappa}\left(\|u\|_2^2\int_0^T\|w_r(\cdot,s)\|_2 ds\right)^{1/2} \\
&\le C_{T/\kappa}\|w_r\|_2\|u\|_2.
\end{aligned}$$

For the singular part we have

$$\begin{aligned}&\left\|\int_0^T\int_0^\infty T(s)bu(t)dw_s(t,s)\,dt\right\|\\&=\left\|\int_0^T e^{-\alpha t/\kappa}T(t/\kappa)bu(t)dt\right\|\\&\le C_{T/\kappa}\kappa\|e^{-\alpha(\cdot)}u(\kappa\cdot)\|_2\le C_{T/\kappa}\|u\|_2.\qquad\square\end{aligned}$$

Note that we have proven a partial converse of Theorem 2.3 in view of the results from [11].

This theorem can be generalized in that one may consider control operators with a range space of higher dimension. Moreover, functions w with more jumps than just at zero can be dealt with on the same basis of proof, but giving a priori conditions on a is more difficult. We thus have $X^T\subset X^S$. Be reminded that Theorem 2.3 does not require $\kappa>0$ for the same result in the diagonal case.

As to the necessity of the condition on k_1, we observe Example 2.5 again. There, we found b that was not admissible for the semigroup generated by A, but admissible for the equation

$$x(t)=x_0+(a*Ax)(t)+(1*bu)(t).$$

In that case $k(t)=2\sqrt{t/\pi}$ and thus $k_1(t)=1/\sqrt{t\pi}$. But $k_1\notin W^{1,2}_{\mathrm{loc}}([0,\infty))$.

Assume that the hypothesis of Theorem 3.2 holds, except that $\kappa=0$. Then the decomposition of w into a regular part with density w_r and a singular part yields a trivial singular part and $w_r(t,\cdot)\in L^2([0,T])$. Now assume $w_r(t,\cdot)\in W^{1,2}([0,T])$. Thus

$$\|(S*bu(T-\cdot))(T)\|=\left\|\int_0^T\int_0^\infty T(s)bu(t)w_r(t,s)ds\,dt\right\|.$$

We remark that $\int_0^t v(s)T(s)x\,ds\in D(A)$ for all $x\in X$, if $v\in W^{1,1}([0,t])$, which is easy to prove. (Use $AT(\cdot)=T'(\cdot)$ and partial integration.) We then need to see that

$$v(\cdot)=\int_0^T u(t)w_r(t,\cdot)dt\in W^{1,1}([0,T]).$$

But this is true, since $w_r(t,\cdot)\in W^{1,2}([0,T])\subset W^{1,1}([0,T])$. Therefore any $b\in X^{-1}$ is admissible in this situation ($\kappa=0$).

THEOREM 3.3. *Suppose that $a\in L^1_{\mathrm{loc}}(\mathbb{R}_+)$ is a completely positive function with $\kappa>0$ and let $k_1\in W^{2,2}_{\mathrm{loc}}([0,\infty))$. Then b is admissible for the Volterra case (i.e., (2)) if and only if it is admissible for the autonomous case (i.e., (1)).*

Proof. By Lemma 3.1, $S(t)$ can be expressed as

$$S(t)=e^{-\alpha t/\kappa}T(t/\kappa)+\int_0^{t/\kappa}T(s)w_r(t,s)ds,\tag{10}$$

where $w_r\in W^{1,1}_{\mathrm{loc}}([0,\infty)^2)$. Now let $\kappa>0$. Since $X^T\subset X^{-1}$, for a given $t>0$,

$$b\mapsto\int_0^\infty w_r(\cdot,s)T(s)b\,ds$$

maps X^T into $L^2([0,t],X)$. Observe that the integral vanishes outside of a compact interval. □

Thus $X^S = X^T$ if we require more regularity on k_1. The representation (10) is however interesting in its own right. The subordinated solution family under this condition is an additive perturbation of the (shifted and rescaled) semigroup with a very "smooth" operator. The perturbation maps not only X^{-1} into X and, similarly, X into X^1 but also is uniformly continuous (cf. [9, section 4]).

In closing we would like to remark that extensions of the presented results to infinite-dimensional control operators are possible. In particular, Theorems 3.2 and 3.3 remain valid even for infinite-dimensional B. An eigenvector expansion in a Hilbert space is also possible for certain constellations. A full-fledged analysis of the infinite-dimensional case is however beyond the scope of this paper; we refer to [12] and [5] for operator generalizations of the Carleson measure criterion in the autonomous case.

Acknowledgment. I would like to thank an unnamed referee for some invaluable suggestions, improving, in particular, the examples.

REFERENCES

[1] Ph. Clément and J. A. Nohel, *Asymptotic behavior of solutions of nonlinear Volterra equations with completely positive kernels*, SIAM J. Math. Anal., 12 (1981), pp. 514–535.
[2] R. F. Curtain and A. J. Pritchard, *Infinite Dimensional Linear Systems Theory*, Springer-Verlag, New York, 1978.
[3] R. F. Curtain and H. J. Zwart, *An Introduction to Infinite-Dimensional Linear Systems Theory*, Springer-Verlag, New York, 1995.
[4] S. Hansen and G. Weiss, *The operator Carleson measure criterion for admissibility of control operators for diagonal semigroups on l^2*, Systems Control Lett., 16 (1991), pp. 219–227.
[5] S. Hansen and G. Weiss, *New results on the operator Carleson measure criterion*, IMA J. Math. Control Inform., 14 (1997), pp. 3–32.
[6] S. Hansen and E. Zuazua, *Exact controllability and stabilization of a vibrating string with an interior point mass*, SIAM J. Control Optim., 33 (1995), pp. 1357–1391.
[7] L. F. Ho and D. L. Russell, *Admissible input elements for systems in Hilbert space and a Carleson measure criterion*, SIAM J. Control Optim., 21 (1983), pp. 614–640.
[8] P. Koosis, *Introduction to H_p Spaces*, Cambridge University Press, Cambridge, UK, 1980.
[9] J. Prüss, *Evolutionary Integral Equations and Applications*, Birkhäuser-Verlag, Basel, 1993.
[10] G. Weiss, *Admissibility of input elements for diagonal semigroups on l^2*, Systems Control Lett., 10 (1988), pp. 79–82.
[11] G. Weiss, *Admissibility of unbounded control operators*, SIAM J. Control Optim., 27 (1989), pp. 527–545.
[12] G. Weiss, *Two conjectures on the admissibility of control operators*, in Estimation and Control of Distributed Parameter Systems, Internat. Ser. Numer. Math. 100, Birkhäuser, Basel, 1993, pp. 367–378.

SIAM J. CONTROL OPTIM.
Vol. 38, No. 5, pp. 1334–1352

ABSTRACT OPTIMAL LINEAR FILTERING*

VLADIMIR N. FOMIN† AND MICHAEL V. RUZHANSKY‡

Abstract. The linear optimal filtering problems in infinite dimensional Hilbert spaces and their extensions are investigated. The quality functional is allowed to be a general quadratic functional defined by a possibly degenerate operator. We describe the solution of the stable and the causal filtering problems. In the case of causal filtering, we establish the relation with a relaxed causal filtering problem in the extended space. We solve the last problem in continuous and discrete cases and give the necessary and sufficient conditions for the solvability of the original causal problem as well as the conditions for the analogue of the Bode–Shannon formula to define an optimal filter.

Key words. linear filtering, stable filtering, causal filtering, optimal filter, Bode–Shannon formula

AMS subject classifications. 93E11, 60G35, 49K45, 60G10, 60G20

PII. S036301299834778X

1. Introduction.

We consider the linear optimal filtering problems in infinite dimensional Hilbert spaces and their extensions. Briefly, the problem is as follows. Let H', H'' be Hilbert spaces and $z = \left[\begin{smallmatrix} x \\ y \end{smallmatrix}\right]$ a random element in $H = H' \times H''$, where x and y are unobservable and observable components of z in H' and H'', respectively. The correlation operator of z is assumed to be bounded in H and we denote by $\mathbb{H}$ a subset of all linear operators $h : H'' \to H'$. The $\mathbb{H}$-optimal linear filtering problem is a problem of the estimations of the unobservable component x based on the realizations of the observable component y in the form

$$\hat{x} = hy \tag{1.1}$$

solving the minimization problem in $\mathbb{H}$

$$J(h) \to \inf_{h \in \mathbb{H}}, \tag{1.2}$$

where the quality functional J is defined by

$$J(h) = E||D(x - \hat{x})||^2 \tag{1.3}$$

with a suitable norm in (1.3) and a linear operator $D : H' \to H'$. The operator D here is an arbitrary operator in general. In that which follows we will sometimes assume that it has an adjoint D^* and that it is continuous if the other operators in the problem are continuous. The choice of operator D allows us to perform the minimization of the quality functional $J(h)$ with respect to some of the variables. In this case D is not bijective and its choice is dictated by the problem at hand. If D is bijective, then, as we shall see, the minimization in (1.2) is performed with respect to *all* variables with certain weights assigned. This leads to the linear transformation of

*Received by the editors December 2, 1998; accepted for publication (in revised form) November 10, 1999; published electronically May 11, 2000.
http://www.siam.org/journals/sicon/38-5/34778.html

†The author is deceased. Former address: Faculty of Mathematics and Mechanics, St. Petersburg University, Bibliotechnaya pl. 2, Peterhof, St. Petersburg 198904, Russia.

‡Department of Mathematics and Statistics, University of Edinburgh, King's Buildings, Mayfield Road, Edinburgh EH9 3JZ, UK (ruzh@maths.ed.ac.uk).

the solution for $D = I$, I being the identity operator. If D is degenerate, the solution of (1.2) is not unique and we will provide formulas for it.

If $\mathbb{H}$ consists of all continuous linear operators h, the problem (1.1), (1.2), (1.3) is called *stable*. If H' and H'' are Hilbert resolution spaces, one has a time structure in H and in its terms defines an "independent of the future" class of the causal continuous operators $\mathbb{H}$. In this case the problem is called *causal*.

If H' and H'' are finite dimensional (the time set $\mathbb{T}$ is discrete and finite), the problem (1.1), (1.2), (1.3) is quite trivial: the causal operators become the lower triangular matrices in a natural basis. In the case of the nondegenerate correlation matrix R_y of the random vector y and the identity matrix D, the problem (1.2) is solvable in the class of the causal weight operators, the solution is unique, and it can be effectively expressed in the terms of the Holetsky factorization [2] of R_y. This result finds various applications [11], [9].

In the case of the discrete time $\mathbb{T} = \mathbb{Z}$ and a stationary partially observable process (time series) z, the problem (1.2) was first treated by Kolmogorov [8]. In the case of the continuous time its solution was first obtained by Wiener [11], who also developed a method for the synthesis of the transfer function of the optimal filter. The Wiener–Kolmogorov optimal filtering theory of the stationary processes was universally accepted partly due to the interpretation of the optimal filter given by Bode and Shannon [1], where the signal y is being "prewhitened" first and the result is optimally processed. The solution representation for optimization problem (1.2), allowing the mentioned interpretation, bears the name of the Bode–Shannon formula.

However, in many applications one estimates only the specific components of x or their combination, which is represented by the degenerate matrix D in the quality functional (1.3) and the solutions of the generalized finite dimensional problems can be found in [10]. In this case the solution need not be unique and there are conditions on the degeneracy of D for which the Bode–Shannon formula still defines an optimal filter.

On the other hand, the infinite dimensional applications required the development of the filtering theory in Hilbert [3] and sometimes Banach spaces [4], [7]. The unobservable component of a process can be formed by the infinite dimensional filter. For example, it can be given by several differential equations with delay or by differential integral equations. In this case it is not possible to use the methods developed for the finite dimensional state space. The method discussed in the paper allows us to formulate such infinite dimensional problems and provides techniques for their solutions. For the applications of this theory to the problem of the linear estimation of the parameters of a signal based on the observations of its realizations see, for example, [4]. In this paper the stable filtering problem will be solved for the general quadratic quality functional (1.3). The solution of the causal filtering problem need not exist in general. We will establish necessary and sufficient conditions of solvability by relaxing the problem, thus allowing a slightly general class of the weight operators in (1.2). The relaxed problem can be solved and the analysis of its solution can be used for the construction of minimizing sequences. The solutions will be given for continuous and discrete resolutions. The conditions for the analogue of the Bode–Shannon formula of [4] to define an optimal filter will also be given.

Some of the literature we are referring to is in Russian, and for the sake of completeness, the results of [4] needed in this paper will be briefly reviewed. We will not give the complete proofs of them in order to avoid technicalities which are unimportant to the nature of the results of this paper. One can consult [5] for some of

the details. A preliminary version of this paper appeared as a preprint [6] while the second author worked at Utrecht University.

In section 2 we fix the notation related to the concept of the extended Hilbert space and random elements in it. In section 3 we formulate and give the solution of the general linear filtering problem in extended Hilbert spaces in Theorem 3.1. The stable linear filtering problem is solved in section 4 (Theorem 4.1). Section 5 is devoted to the causal filtering problem. In subsections 5.1 and 5.2 we discuss Hilbert resolution spaces, their extensions, and linear operators in extended spaces. In subsection 5.3 we formulate the problem. The corresponding relaxed problem is solved in subsection 5.4 (Theorem 5.8). In subsection 5.5 we treat the case of the discrete resolution of the identity (Theorem 5.11) and give necessary and sufficient conditions of the solvability of the original problem (Theorem 5.12). In section 6 the concept of the spectral factorization will be discussed, and the conditions for the Bode–Shannon formula to define an optimal filter will be given in Theorem 6.4.

2. Preliminaries. In this section we will briefly introduce several constructions, which will be used throughout this paper.

2.1. Extended Hilbert spaces and linear operators. Here we will briefly review a concept of the extended Hilbert space (see also [5]). Let H be a complex Hilbert space with an inner product $\langle\cdot,\cdot\rangle$ and let $F \subset H$ be a linear dense subset of H. We will need the notion of F-weak convergence in H.

DEFINITION 2.1. *A sequence $\psi_l \in H$ is called an* F-weakly Cauchy sequence *if*

$$\lim_{l,m\to\infty} \langle \psi_l - \psi_m, \phi\rangle = 0 \text{ for all } \phi \in F.$$

Let Ψ be a set of all F-weakly Cauchy sequences $\psi = \{\psi_l\}$, $\psi_l \in H$, $l \in \mathbb{N}$. Let $\overset{F}{\cong}$ be an equivalence relation in Ψ: $\psi \overset{F}{\cong} \xi$ if $\lim_{m\to\infty}\langle \psi_m - \xi_m, \phi\rangle = 0$ for all $\phi \in F$. Then it is not difficult to check that the quotient space $H_F \equiv \Psi/\overset{F}{\cong}$ of Ψ with respect to the equivalence relation $\overset{F}{\cong}$ is a linear Hausdorff topological space.

Every element $\bar\psi \in H_F$ defines a functional $\bar\psi^* : F \to \mathbb{C}$ by $\bar\psi^*(\phi) = \lim_{l\to\infty}\langle \psi_l, \phi\rangle$, where $\{\psi_l\}$ is a sequence from $\bar\psi$. This duality is an extension of the inner product in H and we will denote this also by $\bar\psi^*(\phi) = \langle \bar\psi, \phi\rangle$. The following relation is obvious.

PROPOSITION 2.1. *H_F is complete in F-weak topology and $F \subset H \subset H_F$.*

The pair (F, H_F) is called an equipment of H, and H with such an equipment is called an equipped Hilbert space. We will also use a construction which gives a space equivalent to H_F.

DEFINITION 2.2. *The space F^* is a space of all the elements f for which there exists a sequence $f_l \in H^*$ such that $f(\phi) = \lim_{l\to\infty} f_l(\phi)$ for all $\phi \in F$.*

Obviously, F^* is complete with respect to the topology of componentwise convergence on F. This implies the completeness of H_F in view of the following.

PROPOSITION 2.2.

(i) *F^* is isomorphic to H_F.*

(ii) *Let $\bar\psi \in H_F$. Define for the corresponding $\bar\psi^* \in F^*$ a "norm" $|\bar\psi^*|_{F^*} = \sup_{\phi\in F} \frac{|\langle\bar\psi,\phi\rangle|}{|\phi|_H}$. Then $\bar\psi \in H$ if and only if $|\bar\psi^*|_{F^*} < \infty$. In this case $|\bar\psi^*|_{F^*} = |\bar\psi|_H$.*

The proof easily follows from the definitions above. Let $A : H \to H$ be a linear operator defined in a dense subspace $D(A)$ of H. Recall the following.

DEFINITION 2.3. *Let $D_* \subset H$ be a space of all elements $\phi \in H$ for which there exist $f(\phi) \in H$ such that $\langle A\psi, \phi\rangle = \langle \psi, f(\phi)\rangle$ for all $\psi \in D(A)$. The operator adjoint to A is defined as $A^* : D_* \to H$, such that $A^*\phi = f(\phi)$.*

One then has $D_* = D(A^*)$. An operator A is called *symmetric* if $D(A) \subset D(A^*)$ and for every $\phi, \psi \in D(A)$ holds $\langle A\psi, \phi\rangle = \langle \psi, A\phi\rangle$. A symmetric operator A is called *self-adjoint* if $D(A) = D(A^*)$.

Let $\bar{A} : D(\bar{A}) \to H_F$ be an extension of A. Assume $F \cap D(\bar{A})$ to be dense in H. Similar to the definition above, define D_* as a space of all $\phi \in F$ for which there exist $f(\phi) \in F$ such that $\langle \bar{A}\bar{\psi}, \phi\rangle = \langle \bar{\psi}, f(\phi)\rangle$ for all $\bar{\psi} \in D(\bar{A})$. Let $\bar{A}^* : D_* \to H_F$ be an operator defined by $\langle \bar{A}\bar{\psi}, \phi\rangle = \langle \bar{\psi}, \bar{A}^*\phi\rangle$ for all $\bar{\psi} \in D(\bar{A})$, $\phi \in D_*$. The F-weak closure of $\bar{A}^*$ is called *the adjoint* to $\bar{A}$ in H_F and will be also denoted by $\bar{A}^*$. As above, $\bar{A}$ is called *symmetric* if $D(\bar{A}) \subset D(\bar{A}^*)$ and A is symmetric. A symmetric operator $\bar{A}$ is called *self-adjoint* if $D(\bar{A}) = D(\bar{A}^*)$.

Example 2.1. Let M be a smooth manifold and let $H = L^2(M)$ with respect to some positive smooth density on M. Let F_k be a space of k-times continuously differentiable compactly supported functions in M. Then the F_k-weak completion of H is a space of distributions of order k in M.

Example 2.2. Let $H = l^2(\mathbb{N})$ with its standard inner product and let $A : l^2 \to l^2$ be defined by $(A\phi)_n = \sum_{k=1}^{\infty} A_{nk}\phi_k$ with suitable conditions on A_{nk}. Let $F \subset H$ be a set of all the sequences consisting of a finite number of nonzero elements. Then F is dense in H and $F \subset D(A) \subset H$. For the extension of H with respect to F one has $H_F = \mathbb{R}^{\mathbb{N}}$, the space of all the sequences with values in $\mathbb{R}$. One readily checks that the above extension of A yields a linear operator $\bar{A}$ in $\bar{l^2} = H_F$ defined by a matrix A_{nk} with $D(\bar{A}) = \{\phi : \sum_{k=1}^{\infty} |A_{nk}\phi_k|^2 < \infty, n \in \mathbb{N}\}$. An important particular case of this example is the following.

Example 2.3. Let $\mathbb{T} = (t_{\mathrm{s}}, t_{\mathrm{f}})$ be a subinterval of $\mathbb{R}$ and let $H = L^2(\mathbb{T})$ be equipped with the standard inner product. Let F be the space of L^2-integrable continuously differentiable functions on $\mathbb{T}$. Then H_F contains not only functions $f \in L^2(\mathbb{T})$ but their generalized derivatives Df as well. In the filtering problems of time series, one normally encounters a discrete set $\mathbf{t} \subset \mathbb{T}$ without accumulation points (with the possible exception of its endpoints). There is a natural correspondence between this set and the Hilbert space $l^2(\mathbf{t})$ over it. The space $L^2(\mathbb{T})$ is infinite dimensional while $l^2(\mathbf{t})$ is infinite dimensional only if $\mathbf{t}$ is infinite. If $\mathbf{t}$ is finite, the space $l^2(\mathbf{t})$ is a standard Euclidean space.

Example 2.4. In this example D denotes the generalized derivative as in Example 2.3 (no relation with the operator D in (1.3)). Let the partially observable element $z = \binom{x}{\mathrm{D}y}$ satisfy $Ez = 0$. Let $x = x(\cdot) \in \mathbb{R}^n$ be a L^2-continuous random process with realizations $x \in L^2(\mathbb{T})$. Let $\mathrm{D}y$ be a scalar random process related to x by a linear "observation scheme":

$$\mathrm{d}y(t) = C^*(t)x(t)\mathrm{d}t + \mathrm{d}w(t), \tag{2.1}$$

where $w(\cdot) \in L_2(\mathbf{T})$ is a standard Gaussian–Wiener process ($Ew(t) = 0$, $w(t_{\mathrm{s}}) = 0$, $Ew(t)w(t') = \min\{t, t'\}$), independent of x, and $C : \mathbb{T} \to \mathbb{R}^n$ is a given continuous function. The correlation operator $R_x = Exx^*$ of a random element x is an integral operator in $L_2(\mathbb{T})$,

$$(R_x\phi)(t) = \int_{t'=t_{\mathrm{s}}}^{t_{\mathrm{f}}} R_x(t, t')\phi(t')\mathrm{d}t', \tag{2.2}$$

where $R_x(\cdot,\cdot) = Ex(t)x^*(t')$ is the correlation matrix of the random process $x = \{x(t),\ t \in \mathbb{T}\}$. This matrix is continuous on $\mathbb{T}\times\mathbb{T}$ because x is L^2-continuous. The random element (process) y is now a generalized element with realizations in $L^2(\mathbb{T})_F$ (see Example 2.3). Its correlation operator $R_{\mathrm{D}}y = E\mathrm{D}y(\mathrm{D}y)^*$ is given by

$$R_{\mathrm{D}y} = C^* R_x C + I, \tag{2.3}$$

where the operator $C : L_2(\mathbb{T}) \to L_2(\mathbb{T})$ is determined by $(C\psi)(t) = C(t)\psi(t)$, $\psi(\cdot) \in L_2(\mathbb{T})$, and I is the identity operator in $L^2(\mathbb{T})$. Formula (2.3) defines the bounded operator $R_y : L^2(\mathbb{T}) \to L^2(\mathbb{T})$ which can be viewed as a singular integral operator with the kernel

$$\begin{aligned} R_{\mathrm{D}y}(t,t') &= C^*(t)R_x(t,t')C(t) + \delta(t-t') \\ &= K(t,t') + \delta(t-t'). \end{aligned} \tag{2.4}$$

The correlation operator between x and $\mathrm{D}y$ is given by

$$R_{x\mathrm{D}y} = R_x C. \tag{2.5}$$

Formulas (2.3), (2.5) allow us to represent the full correlation operator $R_z = Ezz^*$ of the partially observable element z as

$$R_z = \begin{Vmatrix} R_x & R_{x\mathrm{D}y} \\ R^*_{x\mathrm{D}y} & R_{\mathrm{D}y} \end{Vmatrix} = \begin{Vmatrix} R_x & R_x C \\ C^* R_x & C^* R_x C + I \end{Vmatrix}. \tag{2.6}$$

2.2. Generalized random elements. Let F, H, H_F be as above and let $(\Omega, \mathcal{A}, P)$ be a probability space, where P is a complete measure.

Definition 2.4. *A mapping $z : \Omega \to H_F$ is called a* random H_F-element *if for every $\phi \in F$ the following hold:*

(i) *$z^*\phi = \langle z, \phi\rangle : \Omega \to \mathbb{C}$ is a random variable.*

(ii) *$E(z^*\phi) = (Ez)^*\phi$ for some $Ez \in H_F$.*

(iii) *There exists c such that $E|(z - Ez)^*\phi|^2 \le c|\phi|^2_H$ for all $\phi \in F$.*

Without loss of generality, we will consider the centralized elements, $Ez^*\phi = 0$ for all $\phi \in F$. Then $E\overline{z^*\phi}z^*\phi = E|z^*\phi|^2 = \langle \phi, R_z\phi\rangle$ is a quadratic form in F. A linear operator R_z is called the *correlation operator* of z. Property (iii) of the definition implies that R_z is a continuous operator on F and, therefore, it can be extended to a continuous self-adjoint operator in H, $R_z \in \mathcal{L}(H)$. Thus, we have proved the following.

Proposition 2.3. *$R_z \in \mathcal{L}(H)$, $R_z^* = R_z$, and $R_z \ge 0$ in the sense of quadratic forms.*

In analogy with the classical case we will write $R_z = Ezz^*$. Note that Definition 2.4 is equivalent to the condition that $z : \Omega \to H_F$ is measurable, centralized, and has the continuous correlation operator, where Ω and H_F are equipped with σ-algebras $\mathcal{A}$ and one generated by the open sets of F-weak topology in H_F, respectively.

Example 2.5. The random processes can be interpreted as random generalized elements of a Hilbert space. Let $T > 0$ and $H = L^2(m,T)$ be a Hilbert space of m-vector functions on $[0,T]$ equipped with the standard inner product. Let $w = \{w(t), 0 \le t \le T\}$ be a Gauss process, such that almost all realizations of w are elements of $L^2(m,T)$ and $E\int_0^T |w(t)|^2 dt < \infty$. Then R_w is a nuclear operator. If $T = +\infty$, then almost all realizations of $w \notin L^2(m,T)$, but for $F = \mathcal{C}_{\mathrm{comp}}([0,T])$ the realizations of w are elements of $L^2(m,T)_F$.

Example 2.6. Now we return to Example 2.3 and assume that $\mathbb{T} = \mathbb{R}$, ($t_s = -\infty$, $t_f = \infty$). Assume that the partially observable process z is stationary. A stationary linear filter takes the form

$$\hat{x}(t) = \int_{-\infty}^{\infty} h(t - t')\mathrm{d}y(t'), \qquad t \in R. \tag{2.7}$$

Suppose that the optimal filter must minimize the quality functional given by

$$J(h) = E|x(t) - \hat{x}(t)|^2, \tag{2.8}$$

where we can observe that because we assume that the process is stationary, the quality functional $J(h)$ does not depend on t. Let $G_z(\omega)$, $\omega \in \mathbb{R}$, denote the symbol of the correlation operator R_z. Recall that the symbol is just the Fourier transform of the correlation matrix $R_z(\cdot)$. The matrix $G_z(\cdot)$ is also called the *spectral density* of the stationary process z. Then in view of (2.7) the symbol $H_{opt}(\cdot)$ of the transfer operator h_{opt} (the symbol is called the *transfer function* of the optimal filter and is equal to the Fourier transform of the weight function $h(\cdot)$) is given by

$$H_{opt}(\omega) = G_x(\omega)C[C^*G_x(\omega)C + 1]^{-1}, \tag{2.9}$$

where $C \in \mathbb{R}^n$ is the vector from (2.1). This vector C does not depend on t since z is stationary. If instead of (2.8) we take the quality functional

$$J(h) = E|D(x(t) - \hat{x}(t))|^2 \tag{2.10}$$

with a degenerate matrix D, then the transfer function of an optimal filter is not unique; see Theorem 3.1.

3. Linear filtering. Let $H = H' \times H''$, where H' and H'' are Hilbert spaces with inner products $\langle \cdot, \cdot \rangle_{H'}$ and $\langle \cdot, \cdot \rangle_{H''}$, respectively. Let $F' \subset H'$ and $F'' \subset H''$ be linear dense subsets. The elements $\phi \in H$ can be interpreted as $\phi = \left[\begin{smallmatrix} \phi' \\ \phi'' \end{smallmatrix}\right]$ with $\phi' \in H'$, $\phi'' \in H''$. Let $F = F' \times F''$. We will consider random H_F elements $z = [\begin{smallmatrix} x \\ y \end{smallmatrix}]$, with x and y random ${H'}_{F'}$- and ${H''}_{F''}$-elements, respectively. The correlation operator R_z will be assumed continuous on H, which is natural in view of Proposition 2.3, and will have the following block form:

$$R_z = \begin{bmatrix} R_x & R_{xy} \\ R_{yx} & R_y \end{bmatrix},$$

where we write $R_x = Exx^*$, $R_y = Eyy^*$, $R_{xy} = R_{yx}^* = Exy^*$.

Let $h : {H''}_{F''} \to {H'}_{F'}$ be linear. We assume now that there exists an operator $h^* : {H'}_{F'} \to {H''}_{F''}$ defined on the whole of ${H'}_{F'}$, such that for every $\phi' \in F'$, $\phi'' \in F''$ one has

$$(h\phi'')^*\phi' = (\phi'')^*(h^*\phi'). \tag{3.1}$$

Relation (3.1) defines h^* uniquely and h^* is the adjoint to the operator h.

Let x and y be the unobservable and observable components of z, respectively. We define the random ${H'}_{F'}$-element $\hat{x}$ by

$$\hat{x} = hy. \tag{3.2}$$

One readily checks that $\hat{x}$ is a random element in the sense of Definition 2.4 in view of our assumptions on h. Then $R_{\hat{x}} = hR_y h^* : H' \to H'$ is involutive in H' as the correlation operator of a random element $\hat{x}$. The element $\hat{x}$ is interpreted as a linear estimate of the nonobservable component x of a random H_F element z, based on the realizations of its observable component y. The relation (3.2) is called *a linear filter with weight operator* h.

Let a linear operator $D : {H'}_{F'} \to {H'}_{F'}$ have an adjoint D^*. We define *the quality functional* as

$$J_{\phi'}(h) = E|\langle \phi', D(x - \hat{x})\rangle|^2,\ \phi' \in F'. \tag{3.3}$$

Let $\mathbb{H}$ be a given subset of linear operators $h : {H''}_{F''} \to {H'}_{F'}$. Then the $\mathbb{H}$-*optimal filtering problem* is defined as a problem of the minimization of the functionals

$$J_{\phi'}(h) \to \inf_{h\in\mathbb{H}}, \tag{3.4}$$

defined by (3.3), (3.2) for every $\phi' \in F'$.

We will need a notion of the pseudo inversion of an operator. Let $A : H \to H$ be a Hermitian linear operator in a Hilbert space H. Let Q_A be an orthogonal projection on the image of A, $Q_A : H \to \mathrm{Im}A$. The space $Q_A H$ is invariant for A and we write $A^{-1}Q_A$ for the inverse of A in $Q_A H$. The operator

$$A^+ = Q_A \circ A^{-1}Q_A \circ Q_A$$

is called the *pseudo inverse of* A. It follows that

$$A^+A = AA^+ = Q_A. \tag{3.5}$$

One readily checks that (3.5) determines A^+ uniquely and thus we have the following.

Proposition 3.1. *Let A be a Hermitian operator. The solution of*

$$\langle Ag - f, Ag - f\rangle \to \inf_{g\in H},\ f \in H,$$

with minimal norm defines a linear functional of f which is given by $g = A^+ f$.

We will not prove this fact here since we will not use it explicitly. Assume that $\mathbb{H}$ is a space of all linear operators $h : {H''}_{F''} \to {H'}_{F'}$. Then the solution of the $\mathbb{H}$-optimal filtering problem is given by the following.

Theorem 3.1. *Let the correlation operator R_z of a random H element z be continuous in H and let R_y^+ denote the pseudo inverse operator for the correlation operator R_y of y in H''. Then the minimization problem* (3.4) *in the class $\mathbb{H}$ of all weight operators $h : {H''}_{F''} \to {H'}_{F'}$ is solvable and any solution is of the form*

$$h_{\mathrm{opt}} = R_{xy}R_y^+ + Q, \tag{3.6}$$

where $Q : {H''}_{F''} \to {H'}_{F'}$ is any linear operator satisfying $DQQ_{R_y} = 0$ and Q_{R_y} is the kernel of R_y. Moreover, one has

$$\inf_{h\in\mathbb{H}} J_{\phi'}(h) = J_{\phi'}(h_{\mathrm{opt}}) = \langle \phi', D[R_x - R_{xy}R_y^+ R_{xy}^*]D^*\phi'\rangle.$$

The proof follows the lines of the proof of Theorem 4.1, which is given in the next section. The existence of D^* ensures the decomposition (4.3), from which the statement of Theorem 3.1 follows.

4. Linear stable filtering. If $\mathbb{H}$ is a space of all *continuous* linear operators from H'' to H', then the linear filters of the form (3.2) with weight operator in $\mathbb{H}$ are called *stable* and the $\mathbb{H}$-optimal filtering problem is called the *stable filtering problem*. In this case one allows $\phi' \in H'$ in (3.3) and the minimization problem can be reformulated for scalar functionals

$$J(h) = \sup_{\phi' \in H'} \frac{E|\langle \phi', D(x - \hat{x})\rangle|^2}{|\phi'|^2_{H'}}. \tag{4.1}$$

Now we are ready to describe the solution of the linear stable filtering problem. Let us assume that R_y is continuously invertible in its image $R_y H''$, which means that there exists a neighborhood U of zero such that $\sigma(R_y) \cap U = \{0\}$, $\sigma(R_y)$ being the spectrum of R_y. We will also assume that the operator D in the quality functional (3.3) is continuous in H' and has an adjoint D^*.

THEOREM 4.1. *Let the correlation operator R_z of a random H element z be continuous in H and assume that the correlation operator R_y of y has the continuous pseudo inverse operator R_y^+ in H''. Then the minimization problem* (3.4) *in the class $\mathbb{H}$ of all continuous weight operators $h : H'' \to H'$ is solvable and any solution is of the form*

$$h_{\text{opt}} = R_{xy} R_y^+ + Q, \tag{4.2}$$

where $Q : H'' \to H'$ is any linear continuous operator satisfying $DQQ_{R_y} = 0$. Moreover, one has

$$\inf_{h \in \mathbb{H}} J_{\phi'}(h) = J_{\phi'}(h_{\text{opt}}) = \langle \phi', D[R_x - R_{xy} R_y^+ R_{xy}^*] D^* \phi' \rangle_{H'}.$$

The operators (4.2) *are also optimal in the problem with quality functional* (4.1) *and*

$$\inf_{h \in \mathbb{H}} J(h) = J(h_{\text{opt}}) = |D[R_x - R_{xy} R_y^+ R_{xy}^*] D^*|_{H'}.$$

Proof. First we rewrite the quality functionals (3.3) as

$$J_{\phi'}(h) = \langle \phi', R_{D(x - hy)} \phi' \rangle,$$

where $R_{D(x-hy)}$ is the correlation operator of $D(x - hy)$, and using the existence of D^* and h^* we have

$$\begin{aligned} R_{D(x-hy)} &= E[D(x - hy)][D(x - hy)]^* = DE(x - hy)(x - hy)D^* \\ &= D[R_x - R_{xy} h^* - h R_{yx} + h R_y h^*] D^*. \end{aligned}$$

This means

$$\begin{aligned} J_{\phi'}(h) &= \langle \phi', D[R_x - R_{xy} h^* - h R_{yx} + h R_y h^*] D^* \phi' \rangle \\ &= \langle \phi', D[R_x - R_{xy} R_y^+ R_{xy}^*] D^* \phi' \rangle \\ &+ \langle \phi', D(h - R_{xy} R_y^+) R_y (h - R_{xy} R_y^+)^* D^* \phi' \rangle. \end{aligned} \tag{4.3}$$

Here only the second term depends on h and it is a nonnegative quadratic form attaining its minimum if and only if $D(h - R_{xy} R_y^+) Q_{R_y} = 0$ in view of Proposition 2.3. The set of all continuous h satisfying this equation is precisely the set of h_{opt} in (4.2) for all linear continuous $Q : H'' \to H'$ satisfying $DQ_{R_y} = 0$. For such h_{opt} the second term in (4.3) is zero, implying the second statement of the theorem. It follows from

(4.3) that functionals (4.2) are also optimal for the problem (4.1) and one readily verifies the last statement of the theorem. The proof is complete. □

REMARK 4.1. *If the kernel of R_y is nontrivial ($Q_{R_y} \neq I_{H''}$), then one has $h = h_{\rm opt} + \tilde{h}(I_{H''} - Q_{R_y})$, where Q_{R_y} is the orthogonal projection on the image of R_y, and $J_{\phi'}(h) = J_{\phi'}(h_{\rm opt})$ for every linear continuous operator $\tilde{h} : H'' \to H'$. If R_y is bijective ($Q_{R_y} = I_{H''}$), then $R_y^+ = R_y^{-1}$ and $h_{\rm opt} = R_{xy}R_y^{-1} + Q$, $DQ = 0$. If D is bijective, then the only solution of* (3.4) *is $h_{\rm opt} = R_{xy}R_y^+$.*

REMARK 4.2. *These methods can be applied for the problems of linear estimation of the parameters of a signal based on the observations of its realizations.*

We will not discuss it here, but the reader can consult [4] and [5] for the detailed applications.

5. Linear causal filtering. In this section we will give the solution of the generalized linear causal filtering problem. However, we need some preliminary notions and results first.

5.1. Hilbert resolution spaces and causal operators. Let H be a Hilbert space, let $\mathbb{T} = (t_s, t_f)$, $-\infty \le t_s < t_f \le +\infty$, and let $\mathbb{P}_T = \{P_t, t \in \mathbb{T}\}$ be a family of commutative projectors $P_t : H \to H$, $P_t^2 = P_t$, $P_tP_s = P_sP_t$, $t, s \in \mathbb{T}$. Let $\mathbb{P}_T$ satisfy the following two properties:

(i) *Monotonicity*: $P_tP_s = P_s$ for $t \ge s$, $t, s \in \mathbb{T}$.

(ii) *Completeness*: $\lim_{t\to t_s} P_t = 0_H$, $\lim_{t\to t_f} P_t = I_H$, where the limits are taken in the strong operator topology.

Note, that condition (i) is equivalent to the fact that $P_sH \subset P_tH$, $t \ge s$. We assume the family $\mathbb{P}_T$ to be bounded uniformly in t: $\sup_{t\in\mathbb{T}} |P_t| < \infty$ and strongly continuous from the left: $\lim_{\epsilon\to 0+} P_{t-\epsilon}\phi = P_t\phi$ for every $\phi \in H$. Such family $\mathbb{P}_T$ is called a *resolution of the identity of* H and $(H, \mathbb{P}_T)$ is called a *Hilbert resolution space*. If $\mathbb{P}_T$ consists of the orthogonal projectors $P_t = P_t^*$, then it is called a *Hermitian resolution of the identity*. In this case the condition of the uniform boundedness in t is automatically satisfied since $|P_t| \le 1$.

Let $H = H' \times H''$, where $(H', \mathbb{P}'_T)$, $(H'', \mathbb{P}''_T)$ are Hilbert resolution spaces. Then H may be equipped with the canonical resolution of the identity

$$P_t = \begin{bmatrix} P'_t & 0_{12} \\ 0_{21} & P''_t \end{bmatrix}, \ t \in \mathbb{T}, \tag{5.1}$$

where $0_{12} : H'' \to H'$, $0_{21} : H' \to H''$ are zero operators.

DEFINITION 5.1. *Let $A : D(A) \to H$ be a linear densely defined operator. A is called* finite from above *if there exists a measurable, essentially bounded function $\tau : \mathbb{T} \to \mathbb{T}$, such that for almost all $t \in \mathbb{T}$ the operator P_tA is bounded in H and if $t - \tau(t) \in \mathbb{T}$, then*

$$P_tA = P_tAP_{t-\tau(t)} \tag{5.2}$$

on $D(A) \cap P_{t-\tau(t)}D(A)$. The function $\tau = \tau_+(\cdot)$ is called the upper characteristic *of A. A finite from above operator A with characteristic $\tau_+(\cdot)$ is called* τ-causal *or* τ_+-finite.

The space of all τ_+-finite operators will be denoted by $\mathbb{A}^\tau$ and $\mathbb{A}^0 = \cup_\tau \mathbb{A}^\tau$. 0-causal operators are called *causal*. For $\phi \in H$ one can consider a trajectory $\{P_t\phi, \ t \in \mathbb{T}\}$ connecting ϕ and zero in H. Then (5.2) means that a τ-causal operator A considered as a shift operator along these trajectories does not depend on a

future with respect to the resolution; namely, it follows from the completeness of $\mathbb{P}_T$ that $P_t A\phi$ is independent of $P_s\phi$ for $s > t - \tau(t)$. One also has a notion of finiteness from below, given in the following.

DEFINITION 5.2. *Let $A : D(A) \to H$ be a linear densely defined operator. A is called* finite from below *if there exists a measurable, essentially bounded function $\tau : \mathbb{T} \to \mathbb{T}$, such that for almost all $t \in \mathbb{T}$ the operator $(I_H - P_t)A$ is defined in $D(A)$ and if $t - \tau(t) \in \mathbb{T}$, then*

$$(I_H - P_t)A = (I_H - P_t)A(I_H - P_{t-\tau(t)}) \tag{5.3}$$

on $D(A) \cap (I_H - P_{t-\tau(t)})D(A)$. The function $\tau = \tau_-(\cdot)$ is called the lower characteristic *of A. A finite from below operator A with characteristic $\tau_-(\cdot)$ is called* τ-anticausal.

DEFINITION 5.3. *Let $A : D(A) \to H$ be a linear densely defined operator. A is called* finite *if it is finite from above and below.*

Note that for the Hermitian resolution of the identity, A is finite from below if and only if A^* is finite from above and in this case $\tau_-(A) = \tau_+(A^*)$. 0-anticausal operators are called *anticausal* and one writes $\mathbb{A}_\tau$ for all $\tau = \tau_-$-anticausal operators, $\mathbb{A}_0 = \cup_\tau \mathbb{A}_\tau$. If $A \in \mathbb{A}^0 \cap \mathbb{A}_0$ and $\tau_+ = \tau_- = \tau$, then A is called τ-*local*. 0-local operator is called *local*. Every operator commuting with $\mathbb{P}_T$ is local. We will need the following property.

LEMMA 5.4 (see [4]). *Let $A \in \mathbb{A}^\tau$ (resp., $\mathbb{A}_\tau$), $A' \in \mathbb{A}^{\tau'}$ (resp., $\mathbb{A}_{\tau'}$). Then $B = AA'$ (if the composition exists) is finite from above (resp., below) with characteristic $\beta(t) = \tau(t) + \tau'(t - \tau(t))$.*

Example 5.1. Let $H = L^2(\mathbb{R})$ and $(h\phi)(t) = \int_{-\infty}^{+\infty} h(t,s)\phi(s)ds$. Hilbert space L^2 becomes a resolution space when equipped with a family $\mathbb{P}_T$ defined by

$$P_t\phi(s) = \begin{cases} \phi(s) & \text{if } s \le t, \\ 0 & \text{if } s > t. \end{cases}$$

One readily sees that h is τ-causal (anticausal) if and only if $h(t,s) = 0$ for $s > \min(t, t - \tau(t))$, $(s < \max(t, t - \tau(t)))$. In particular, h is local if and only if $h = 0$.

5.2. Extended Hilbert resolution spaces. Assume now that H is infinite dimensional and $t_f = +\infty$. An element $\phi \in H$ is called *finite* if there exist $t_*(\phi) \in \mathbb{T}$, $t_* < \infty$, such that $P_t\phi = \phi$ for all $t \ge t_*$. Let F be a space of all finite elements of H. Then F is dense in H and $\bar{H} = H_F$ is called the t-*extension* or t-*completion* of H. We write $t - \lim_{n\to\infty} \phi_n = \phi$ if for every $t \in \mathbb{T}$ one has $\lim_{n\to\infty} P_t\phi_n = P_t\phi$; $\phi_n, \phi \in H$. This defines t-*convergence* in H and the associated Hausdorff topology is weaker than the canonical inner product topology of H. Note that H is not complete with respect to t-convergence. One readily checks the following simple proposition.

PROPOSITION 5.1. *The completion of H with respect to t-topology is isomorphic to H_F, the F-weak completion of H.*

A densely defined operator A in H is called t-*continuous* if for every sequence $\phi_n \in D(A)$ with $t - \lim_{n\to\infty} \phi_n = 0$ one has $t - \lim_{n\to\infty} A\phi_n = 0$. Note that a general continuous operator in H need not be t-continuous. Now we collect the further properties following [4] (see also [5], [3]).

LEMMA 5.5. *The following hold:*

(i) *Operator A is t-continuous if and only if it is finite.*

(ii) *If A is t-continuous, then by Proposition* 5.1 *it allows an extension to an operator $\bar{A}$ in $\bar{H}$: $\bar{A}|_H = A$. In particular, every $P_t \in \mathbb{P}_T$, being a local*

operator, allows an extension to $\bar{P}_t$ in $\bar{H}$. The family $\overline{\mathbb{P}_T}$ is a resolution of the identity in $\bar{H}$. One can generalize the notions of causality for $\bar{H}$, in particular, $\bar{P}_t$ are local in $\bar{H}$.

(iii) *For every $t \in \mathbb{T}$ and $\bar{\phi} \in \bar{H}$ holds $\bar{P}_t\bar{\phi} \in P_tH$.*

(iv) *A restriction of τ-causal operator $\bar{A}$ in $\bar{H}$ to H defines a τ-causal operator A in H.*

(v) *If $|\bar{\phi}|_{\bar{H}} = \sup_{t\in\mathbb{T}} |\bar{P}_t\bar{\phi}|_H$, then $|\bar{\phi}|_{\bar{H}} < \infty$ if and only if $\bar{\phi} \in H$. In this case $|\bar{\phi}|_{\bar{H}} = |\bar{\phi}|_H$.*

(vi) *Let $\bar{A} : \bar{H} \to \bar{H}$ be linear τ-causal. Then there exists an operator $\bar{A}^* : \bar{H} \to \bar{H}$ uniquely defined by*

$$(\bar{P}_t\bar{A}\bar{P}_{t-\tau(t)}\bar{\phi})^*\psi = (\bar{P}_{t-\tau(t)}\bar{\phi})^*\bar{A}^*\bar{P}_t\psi$$

for every $\psi \in F$, $\bar{\phi} \in \bar{H}$, $t \in \mathbb{T}$. The operator $\bar{A}^*$ is the adjoint to $\bar{A}$ and is $(-\tau)$-anticausal.

DEFINITION 5.6. *An operator $\bar{A} : \bar{H} \to \bar{H}$ is called* τ-bounded *for a measurable function $\tau : \mathbb{T} \to \mathbb{T}$ if*

$$\sup_{\bar{\phi}\in\bar{H}} \sup_{t\in\mathbb{T}} \frac{|\bar{P}_t\bar{A}\bar{\phi}|_H}{|\bar{P}_{t-\tau(t)}\bar{\phi}|_H} < \infty.$$

0*-bounded operators are called* stable [3].

We collect the properties of τ-bounded operators in the following.

LEMMA 5.7. *The following hold:*

(i) *Let $\bar{A} : \bar{H} \to \bar{H}$ be τ-bounded. Then H is an invariant subspace for $\bar{A}$ and the restriction $\bar{A}|_H$ is continuous.*

(ii) *Let $\bar{A} : \bar{H} \to \bar{H}$ be τ-bounded for $\tau \geq 0$. Then $\bar{A}$ is τ-causal with respect to $\overline{\mathbb{P}_T}$.*

(iii) *An operator $\bar{A} : \bar{H} \to \bar{H}$ is stable if and only if*
   (a) *$\bar{A}$ is causal,*
   (b) *H is an invariant subspace of $\bar{A}$,*
   (c) *the restriction $\bar{A}|_H$ is continuous in H.*

5.3. Linear causal filtering problem. Let $(H', \mathbb{P}'_T)$, $(H'', \mathbb{P}''_T)$ be Hermitian resolution spaces. Let $H = H' \times H''$ be equipped with the resolution defined by (5.1). We denote by $\mathbb{H}^\tau$ the space of all linear continuous τ-causal operators $h : H'' \to H'$. Let $D : H' \to H'$ be continuous with the adjoint $D^* : H' \to H'$. Then *the optimal linear causal filtering problem* is the minimization problem

$$J_{\phi'}(h) \to \inf_{h\in\mathbb{H}^\tau} \tag{5.4}$$

for every $\phi' \in H'$, where $J_{\phi'}(h)$ is defined by

$$J_{\phi'}(h) = E|\langle \phi', D(x - hy)\rangle|^2,\ h \in \mathbb{H}^\tau. \tag{5.5}$$

It turns out that the condition of the continuity of weight operators is very restrictive for the solution of the problem (5.4). In general, the optimal filtering problem in the class of continuous weight operators can be unsolvable or can be very complicated. At the same time, if we drop the continuity condition the problem can be solved. It is relatively simple to check whether the weight operator of the determined optimal filter is continuous. In this way one can obtain solutions for the original continuous problem. We will apply the methods presented in [4]; namely, first we relax the problem (5.4) allowing h to be unbounded. Analyzing the solution of the relaxed problem we derive the conditions for the solvability of (5.4).

5.4. Generalized linear causal filtering problem. Let $\bar{H}'$, $\bar{H}''$ be the t-completions of H' and H'', respectively. Let $\bar{\mathbb{H}}^\tau$ be the space of all linear τ-causal operators $\bar{h} : \bar{H}'' \to \bar{H}'$, such that for every $t \in \mathbb{T}$ the operators $\bar{P}'_t \bar{h} \bar{h}^* \bar{P}'_t : P'_t H' \to P'_t H'$ are continuous. Assume z to be a random $\bar{H}$ element, and, therefore, $R_z = Ezz^*$, $z = \left[\begin{smallmatrix} x \\ y \end{smallmatrix}\right]$, is bounded on the space F of finite elements in H and can then be continuously extended to the whole of H. The problem is to find linear estimates of a random H' element x based on the realizations of a random H'' element y of the form

$$\hat{x} = \bar{h}y, \tag{5.6}$$

minimizing for every $t \in \mathbb{T}$ the functional

$$J^{(t)}(\bar{h}) = E|D\bar{P}_t(x - \hat{x})|^2_{H'}. \tag{5.7}$$

Note that $J^{(t)}(\bar{h})$ is finite for $\bar{h} \in \bar{\mathbb{H}}^\tau$, $t \in \mathbb{T}$; therefore the problem of the minimization

$$J^{(t)}(\bar{h}) \to \inf_{\bar{h} \in \bar{\mathbb{H}}^\tau} \tag{5.8}$$

for every $t \in \mathbb{T}$ is correctly posed. Let us reformulate the problem (5.8) now. For $\phi' \in H'$ we define

$$\begin{aligned} J^{(t)}_{\phi'}(\bar{h}) &= E|\langle \phi', DP'_t(x - \hat{x})\rangle_{H'}|^2 = E|\langle \phi', DP'_t(x - \bar{h}x)\rangle_{H'}|^2 \\ &= \langle \phi', DP'_t[R_x - R_{xy}\bar{h}^* - \bar{h}R_{yx} + \bar{h}R_y\bar{h}^*]P'_t D^*\phi'\rangle_{H'}. \end{aligned} \tag{5.9}$$

Now, the problem (5.8) is equivalent to the problem

$$J^{(t)}_{\phi'}(\bar{h}) \to \inf_{\bar{h} \in \bar{\mathbb{H}}^\tau} \tag{5.10}$$

for every $\phi' \in H'$.

THEOREM 5.8. *Let $R_z = Ezz^*$ satisfy the following:*

(i) *The operators $R_z^{(t,t)} = \bar{P}_t R_z \bar{P}_t : P_t H \to P_t H$ are continuous for every $t \in \mathbb{T}$.*

(ii) *The operators $P''_t R_y P''_t : H'' \to H''$ are positive in the invariant subspace $P''_t H''$ for every $t \in \mathbb{T}$.*

Then for every $t \in \mathbb{T}$ there exist $\hat{x}_t \in P'_t H'$ such that for every $\phi \in H'$ one has

$$E|\langle \phi', D(x - \hat{x}_t)\rangle_{H'}|^2 = \inf_{h \in \bar{\mathbb{H}}^\tau} E|\langle \phi', D(x - P'_t hy)\rangle_{H'}|^2.$$

The estimates $\hat{x}_t$ are given by

$$\hat{x}_t = R_{xy}^{(t,t-\tau(t))}(R_y^{(t-\tau(t),t-\tau(t))})^{-1} P''_{t-\tau(t)} y + Q_t P''_{t-\tau(t)} y, \tag{5.11}$$

where $R_{xy}^{(t,t-\tau(t))} = P'_t R_{xy} P''_{t-\tau(t)}$, $R_y^{(t,t)} = P''_t R_y P''_t$, $(R_y^{(t,t)})^{-1}$ means the inverse of $R_y^{(t,t)}$ in the invariant subspace $P''_{t-\tau(t)} H''$, and any $Q_t : H'' \to H'$ such that $DQ_t Q_{R_y} = 0$. Moreover,

$$\begin{aligned} E|\langle \phi', D(x - \hat{x}_t)\rangle_{H'}|^2 = \langle \phi', D[&P'_t R_x P'_t \\ &- R_{xy}^{(t,t-\tau(t))}(R_y^{(t-\tau(t),t-\tau(t))})^{-1} R_{yx}^{(t-\tau(t),t)}]D^*\phi'\rangle_{H'}. \end{aligned}$$

Proof. In view of (5.11) we rewrite (5.9) as

$$J^{(t)}_{\phi'}(\bar{h}) = \langle \phi', D[P'_t R_x P'_t - R^{(t,t-\tau(t))}_{xy}\bar{h}^* - \bar{h}R^{(t-\tau(t),t)}_{yx} + \bar{h}R^{(t-\tau(t),t-\tau(t))}_{y}\bar{h}^*]D^*\phi'\rangle_{H'}. \tag{5.12}$$

The minimization problem (5.10) is now the same as the minimization of the functionals (5.12) in the invariant for $R^{(t-\tau(t),t-\tau(t))}_y$ subspace $H''_{t-\tau(t)} = P''_{t-\tau(t)}H''$. This is the minimization problem (3.4) for $H' = P'_t H'$ and $H'' = H''_{t-\tau(t)}$. Theorem 4.1 together with the invertibility of $R^{(t-\tau(t),t-\tau(t))}_y$ by the assumption (ii) of Theorem 5.8 imply the solution of the problem in the form given by (5.11) and the last formula of the theorem. □

REMARK 5.1. *In view of the discussion above it is easy to see that problem* (5.4) *is solvable if and only if the solution of* (5.8) *is a continuous* τ*-causal operator. In this case the restriction of this operator to* H'' *gives a solution to* (5.4).

The detailed discussion and the solutions of these problems for $D = I_{H'}$ can be found in [4], [5], [7]. We will treat further the spaces with the discrete resolution of the identity. In general, the problems described above can be reduced to the discrete case by a suitable approximation of $\mathbb{P}_T$ by discrete resolutions of the identity; see [4] for the details.

5.5. Discrete resolutions of the identity. We assume now that $\mathbb{P}_T$ is a piecewise constant operator valued functional on $\mathbb{T}$ with at most a countable number of discontinuity points without accumulations in $\mathbb{T}$. Let $\mathbf{t} = \{t_k, k \in \mathbb{K}\}$ be a finite or a countable ordered subset of $\mathbb{T}$ without accumulation points, $\mathbb{K} = \mathbb{Z} \cap (0, K)$, $t_0 = t_s$, $t_K = t_f$, K finite or $K = +\infty$. The *discrete resolution of the identity in* H *corresponding to* $\mathbf{t} \subset \mathbb{T}$ is the set $\mathbb{P}_{\mathbf{t}} = \{P_t, t \in \mathbf{t}\}$. The family of the orthogonal projectors $Q_k = P_{t_k} - P_{t_{k-1}}$, $k \in \mathbb{K}$, determines the resolution $\mathbb{P}_{\mathbf{t}}$ uniquely due to the relation $P_t = \sum_{k:t_k \le t} Q_k$. These projectors are mutually orthogonal: $Q_k Q_l = Q_l Q_k = 0_H$ for $k \ne l$.

DEFINITION 5.9. *A family* $\mathbb{Q}_K$ *of the mutually orthogonal projectors* Q_k *is called* the orthogonal resolution of the identity *if* $\mathbb{Q}_K$ *is complete in a sense that* $Q_k \to O_H$ *for* $k \to k_s$ *and* $\sum_{l \le k} Q_l \to I_H$ *for* $k \to k_f$. *The pair* $(H, \mathbb{Q}_K)$ *is called the* discrete resolution space.

Every linear operator $R : H \to H$ can be decomposed with respect to $\mathbb{Q}_K$ into blocks $R_{kl} = Q_k R Q_l$ and $R = \sum_{k,l \in \mathbb{K}} R_{kl}$. The definitions of finiteness, causality, and anticausality can be reformulated in terms of the discrete structure $\mathbb{Q}_K$. The function τ in Definitions 5.1, 5.2 is replaced by $\tau : \mathbf{t} \to \mathbf{t}$ with a property that $\tau(t_k) = t_l$, $k, l \in \mathbb{K}$, and the latter corresponds to a function $\kappa : \mathbb{K} \to \mathbb{K}$ such that $\tau(t_k) = t_{\kappa(k)}$. In analogy to the continuous case one has the following.

DEFINITION 5.10. *A linear operator* $R : H \to H$ *is called* κ-causal (strictly κ-causal, κ-anticausal) *if* $R_{kl} = 0_H$ *for* $l > k - \kappa(k)$ ($l \ge k - \kappa(k)$, $l < k - \kappa(k)$), *respectively. It is called* neutral *if it is causal and anticausal.*

For a linear operator $R : H \to H$ we denote its κ-causal, anticausal, and neutral components by $R_{[\kappa]} = \sum_{l \le k - \kappa(k)} R_{kl}$, $R^{[\kappa]} = \sum_{l \ge k - \kappa(k)} R_{kl}$, $R_{[[\kappa]]} = \sum_{l = k - \kappa(k)} R_{kl}$, respectively.

Now we are ready to formulate the optimal causal filtering problem for the discrete resolution space $H = H' \times H''$, H', H'' equipped with the orthogonal resolutions of the identity $\mathbb{Q}'_K$ and $\mathbb{Q}''_K$, respectively. Let $\mathbb{H}^\kappa$ denote the space of all κ-causal continuous operators $h : H'' \to H'$ and $\hat{x}_k = Q'_k \hat{x}$, $y_k = Q''_k y$, $h_{kl} = Q'_k h Q''_l$. Then the problem

is the linear estimation

$$\hat{x}_k = \sum_{l \le k-\kappa(k)} h_{kl} y_l \tag{5.13}$$

minimizing the functionals

$$J_{\phi'}(h) = E|\langle \phi', D(x - hy)\rangle|^2 \to \inf_{h \in \mathbb{H}^\kappa} \tag{5.14}$$

for every $\phi \in H'$. Note that this is the same as the minimization of

$$J_k(h) = E|D(x_k - \hat{x}_k)|^2 \to \inf_{h \in \mathbb{H}^\kappa} \tag{5.15}$$

for every $k \in \mathbb{K}$, where $x_k = Q_k' x$.

In analogy with the continuous case we will treat the relaxed problem first, replacing the condition of the continuity of h by the continuity of $h_k = Q_k' h = \sum_{l \in \mathbb{K}} h_{kl} : H'' \to H'$ for every $k \in \mathbb{K}$. The space of all linear κ-causal operators for which all the correspondent operators h_k are continuous will be denoted by $\bar{\mathbb{H}}^\kappa$. Note that because J_k are finite when R_z is bounded, the problem

$$J_k(h) = E|D(x_k - \hat{x}_k)|^2 \to \inf_{h \in \bar{\mathbb{H}}^\kappa},\ k \in \mathbb{K}, \tag{5.16}$$

is correctly posed. Note that if $\bar{H}', \bar{H}''$ are the completions of H', H'' in t-topology, then the space $\bar{\mathbb{H}}^\kappa$ is isomorphic to the space of all κ-causal operators from $\bar{H}''$ to $\bar{H}'$. The problem now becomes

$$J_k(h) = E|DQ_k'(x - \bar{h}y)|^2 \to \inf_{\bar{h} \in \bar{\mathbb{H}}^\kappa},\ k \in \mathbb{K}. \tag{5.17}$$

In analogy to Theorem 5.8 and Theorem 2.3 in [4] we have the following.

THEOREM 5.11. *Let $R_z = Ezz^*$ be continuous and R_y satisfy $P_{t_k} R_y P_{t_k} \ge \epsilon P_{t_k}$ for some $\epsilon > 0$ and for every $k \in \mathbb{K}$. Then all the solutions $\bar{h}_{\text{opt}} : \bar{H}'' \to \bar{H}'$ of the problem* (5.17) *are given by*

$$\bar{h}_{\text{opt}} = \sum_{k \in \mathbb{K}} Q_k' R_{xy} P''_{t_{k-\kappa(k)}} (P''_{t_{k-\kappa(k)}} R_y P''_{t_{k-\kappa(k)}})^{-1} P''_{t_{k-\kappa(k)}} + Q, \tag{5.18}$$

where $Q \in \bar{\mathbb{H}}^\kappa$ satisfies $DQ_k'Q = 0$. One has

$$\inf_{h \in \bar{\mathbb{H}}^\kappa} J_k(\bar{h}) = J_k(\bar{h}_{\text{opt}})$$
$$= |DP'_{t_k}[R_x - R_{xy} P''_{t_{k-\kappa(k)}} (P''_{t_{k-\kappa(k)}} R_y P''_{t_{k-\kappa(k)}})^{-1} P''_{t_{k-\kappa(k)}} R_{yx}] P'_{t_k} D^*|.$$

The proof is similar to the proof of Theorem 5.8 and is based on the calculations of $\hat{x}_k$ as an optimal estimate in the subspace of H'' spanned by $y_l' = Q_k' y$, $l \le k - \kappa(k)$. Similar to [4, Theorem 2.4] for the solution of the original problem (5.15) we have the following.

THEOREM 5.12. *Let the assumptions of Theorem* 5.11 *be satisfied. Then the problem* (5.15) *is solvable if and only if the solution $\bar{h} : \bar{H}'' \to \bar{H}'$ of* (5.17) *is κ-bounded. In this case the image of H'' under $\bar{h}$ is contained in H' and the restriction $\bar{h}|_{H''}$ is the solution of* (5.15).

Proof. Under the assumptions of Theorem 5.12, formula (5.18) defines the optimal linear filters for (5.17). If $\bar{h}_{\text{opt}}$ is κ-finite, the operator $\bar{h}_{\text{opt}}|_{H''}$ is continuous and

defines the weight operators for the solutions of (5.15). If $\bar{h}_{\rm opt}$ is not κ-bounded, the converse is also obvious from Lemma 5.7. □

Note that by taking finite partial sums in (5.18) one obtains minimizing sequences for the problem. These minimizing sequences are also available in the case when $\bar{h}$ is not κ-bounded and the problem (5.15) is not solvable in $\mathbb{H}^\kappa$.

6. Bode–Shannon representation of optimal filters. First we will briefly review the results on the spectral factorization of the operators which we need in order to discuss the application of Bode–Shannon theory (cf. [1], [4], [9], [10]) in our setting. Detailed discussion on various types of spectral factorization and separation of the operators can be found in [4].

Let $\mathbb{P}_T$ be a Hermitian resolution of the identity in H. As in the previous section we denote by $\bar{H}$ a t-completion of H and by $\mathbf{t}$ a discrete linearly ordered subset of $\mathbb{T}$. Let $\mathbb{G}_{\mathbf{t}}$ be a space of all bijective operators $\bar{G}:\bar{H}\to\bar{H}$ such that $\bar{P}_t\bar{G}\bar{P}_t$ and $\bar{P}_t\bar{G}^{-1}\bar{P}_t$ are continuous as operators from P_tH to P_tH for every $t\in\mathbf{t}$. Note that $\mathbb{G}_{\mathbf{t}}$ contains the space of all causal, causally invertible operators in H.

DEFINITION 6.1. *An operator $\bar{G}\in\mathbb{G}_{\mathbf{t}}$ is called* spectrally factorizable *if there exists a causal with respect to $\mathbb{P}_T$ operator $\bar{U}:\bar{H}\to\bar{H}$, such that the inverse of $\bar{U}$ exists and is causal in $\bar{H}$ and $\bar{G}=\bar{U}\bar{U}^*$, where $\bar{U}^*$ is the adjoint of $\bar{U}$.*

Let $\bar{G}=\bar{U}\bar{U}^*$ be a spectral factorization of $\bar{G}$. If $\bar{U},\bar{U}^{-1}$ are stable (Definition 5.6), the restrictions $\bar{U}|_H,\bar{U}^{-1}|_H$ are causal and continuous in H in view of Lemma 5.7. This implies that the restriction $G=\bar{G}|_H$ is continuous in H and we can summarize it in the following.

DEFINITION 6.2. *A continuous operator $G:H\to H$ is called* strongly spectrally factorizable *if there exists a continuous causal operator $U:H\to H$ with continuous and causal inverse, such that $G=UU^*$, where U^* is the adjoint of U.*

We call operator $\bar{G}\in\mathbb{G}_{\mathbf{t}}$ *positive* if the operators $\bar{P}_t\bar{G}\bar{P}_t$, $\bar{P}_t\bar{G}^{-1}\bar{P}_t:P_tH\to P_tH$ are nonnegative for every $t\in\mathbf{t}$. The following are Theorems 2.5 and 2.6 in [4] (see also [7], [3]).

THEOREM 6.3.

(i) *Every positive operator $\bar{G}\in\mathbb{G}_{\mathbf{t}}$ is spectrally factorizable. A causal with respect to the discrete resolution $\bar{P}_{\mathbf{t}}$ operator $\bar{U}$ factorizing $\bar{G}$ is unique up to the multiplication from the right by a neutral unitary in $\bar{H}$ operator.*

(ii) *Let $\bar{G}\in\mathbb{G}_{\mathbf{t}}$ and assume that the restriction $G=\bar{G}|_H$ is positive and continuous in H. Then G is strongly spectrally factorizable. A causal with respect to the discrete resolution $\mathbb{Q}_K$ operator U factorizing G is unique up to the multiplication from the right by a neutral unitary in H operator.*

It is convenient in the discrete case $(H,\mathbb{Q}_K)$ to denote by $\mathbb{G}_K$ the space of all bijective operators $\bar{G}:\bar{H}\to\bar{H}$ such that for every $k\in\mathbb{K}$ the operators

$$\sum_{l=0}^{k}\sum_{m=0}^{k}\bar{Q}_l\bar{G}\bar{Q}_m,\quad \sum_{l=0}^{k}\sum_{m=0}^{k}\bar{Q}_l\bar{G}^{-1}\bar{Q}_m$$

are continuous from $H^k=\bigoplus_{l=0}^{k}\bar{Q}_lH$ to H^k. $\bar{G}\in\mathbb{G}_K$ is called *positive* if the operators in the definition of $\mathbb{G}_K$ are nonnegative for every $k\in\mathbb{K}$.

Let $H=H'\times H''$ be equipped with the orthogonal resolution of the identity given by $Q_k=\left[\begin{smallmatrix}Q'_k & 0\\ 0 & Q''_k\end{smallmatrix}\right]$, $Q'_k\in\mathbb{Q}'_K$, $Q''_k\in\mathbb{Q}''_K$. The κ-causal filters are given by

$$\hat{x}_k=\sum_{l=0}^{k-\kappa(k)}h_{kl}y_l,$$

where $h_{kl} : Q_l'' H'' \to Q_k' H'$ are linear continuous. The corresponding filter $\bar{h} : \bar{H''} \to \bar{H'}$ is defined by having its blocks equal to h_{kl}.

THEOREM 6.4. *Assume that $R_z \in \mathbb{G}_K$, that $R_y : H'' \to H''$ is positive and $\kappa \geq 0$. Then all optimal linear filters for the discrete generalized linear causal filtering problem* (5.17) *are of the form*

$$\bar{h}_{\mathrm{opt}} = [R_{xy}(U^{-1})^*]_{[\kappa]} U^{-1} + Q, \tag{6.1}$$

where U is a causal operator strongly factorizing R_y, $[R_{xy}(U^{-1})^*]_{[\kappa]}$ is the κ-causal component of $R_{xy}(U^{-1})^* : \bar{H''} \to \bar{H'}$, and any $Q \in \bar{\mathbb{H}}^\kappa$ such that $DQ = 0$. One has

$$\inf_{\bar{h} \in \bar{\mathbb{H}}^\kappa} J_k(\bar{h}) = J_k(\bar{h}_{\mathrm{opt}})$$
$$= |DQ_k[R_x - R_{xy}R_y^{-1}R_{yx} + [R_{xy}(U^{-1})^*]_{[\bar{\kappa}]}([R_{xy}(U^{-1})^*]_{[\bar{\kappa}]})^*]Q_k D^*|.$$

Proof. Let $L = R_{xy}(U^{-1})^* - [R_{xy}(U^{-1})^*]_{[\kappa]}$ denote the strictly anticausal component of $R_{xy}(U^{-1})^*$. Rewriting $J_k(\bar{h})$ in analogy with (4.3) we have

$$\begin{aligned} E|\langle \phi', DQ_k'(x - \bar{h}y)\rangle|^2 &= \langle \phi', DQ_k'[R_x - R_{xy}\bar{h}^* - \bar{h}R_{yx} + \bar{h}R_y\bar{h}^*]Q_k'D^*\phi'\rangle \\ &= \langle \phi', DQ_k'[R_x - R_{xy}R_y^{-1}R_{yx}]Q_k'D^*\phi'\rangle \\ &+ \langle \phi', DQ_k'(\bar{h}U - M - L)(\bar{h}U - M - L)^*Q_k'D^*\phi'\rangle \\ &= \langle \phi', DQ_k'[R_x - R_{xy}R_y^{-1}R_{yx}]Q_k'D^*\phi'\rangle \\ &+ \langle \phi', DQ_k'(\bar{h}U - M)(\bar{h}U - M)^*Q_k'D^*\phi'\rangle \\ &+ \langle \phi', DQ_k'[(\bar{h}U - M)L^* - L(\bar{h}U - M)^*]Q_k'D^*\phi'\rangle, \end{aligned} \tag{6.2}$$

where M denotes $[R_{xy}(U^{-1})^*]_{[\kappa]}$, $R_{xy}(U^{-1})^* = L + M$. Note that in the last equality in (6.2) the first term is constant in h, the second one is quadratic in $\bar{h}U - M$, and the third is linear. Now, an application of Lemma 5.4 yields the κ-causality of $\bar{h}U - M$, and in view of strict causality of L^* again by Lemma 5.4, the operator $(\bar{h}U - M)L^*$ is strictly κ-causal. It follows that $Q_k'(\bar{h}U - M)L^*Q_k' = 0$ if $\kappa \geq 0$. This means that the linear term in (6.2) vanishes and the minimum is attained if and only if the quadratic term in (6.2) is zero. This is the case of $\bar{h}U - M = S$ for any $S \in \bar{\mathbb{H}}^\kappa$ such that $DS = 0$. Multiplication by 0-causal operators U, U^{-1} does not change κ-causality in view of Lemma 5.4 and we obtain formula (6.1) with $Q = SU^{-1}$. The last formula of the theorem follows from the substitution of $\bar{h}_{\mathrm{opt}}$ into the last expression of (6.2). □

COROLLARY 6.5. *For $D = I_{H'}$ the only operator in* (6.1) *is obtained by taking $Q = 0$. This operator is called the* Bode–Shannon weight operator *and the filter* (5.13) *is called the* Bode–Shannon filter.

COROLLARY 6.6. *If R_z is stable, R_y^{-1} exists and is continuous in H, and stable operator $R_{xy}(U^{-1})^*$ has the stable κ-causal component, then the original linear optimal causal filtering problem* (5.15) *is solvable and all optimal weight operators are the restrictions of $\bar{h}_{\mathrm{opt}}$ in* (6.1) *to H''.*

The proof is similar to the proof of Theorem 5.12 and is left as an exercise. The reader can consult [4] for the application to the finite dimensional stationary processes, where the conditions of Corollary 6.6 are reduced to the conditions in terms of analytic functions.

REMARK 6.1. *If the operator D^*D in* (1.3) *allows a spectral factorization*

$$D^*D = \tilde{D}^*\tilde{D}$$

with a causal operator $\tilde{D}$, and in addition R_y is positive and the problem (1.2) *is solvable, then an optimal weight operator must solve the equation*

$$\tilde{D}h = (\tilde{D}R_{xy}(U^{-1})^*)_{[\kappa]}U^{-1}.$$

Although in this paper we prove the Bode–Shannon formula for the optimal filtering problem with the discrete time only, its analogue is valid for the continuous time as well. We use this opportunity to demonstrate it. We return to Example 2.3.

Example 6.1. Assume that the filter takes the form

$$\hat{x}(t) = \int_0^{t-\tau} h(t,t')\,\mathrm{d}y(t'), \qquad t_\mathrm{s} \le t-\tau \le t_\mathrm{f}, \tag{6.3}$$

where $\tau \in \mathbb{R}$ is given. In this case the causality of the filter (6.3) is determined by the family P_t, $t \in (t_\mathrm{s}, t_\mathrm{f})$, of projections from Example 5.1. The quality functional is

$$J_t(h) = E|x(t) - \hat{x}(t)|^2, \qquad t_\mathrm{s} \le t-\tau \le t_\mathrm{f}. \tag{6.4}$$

The Bode–Shannon formula for the optimal τ-causal filter is (cf. Theorem 6.4 and (2.5))

$$h_{opt} = [R_x C(U^{-1})^*]^{(\tau)}U^{-1}, \tag{6.5}$$

where $I+U$ is the spectral factor for $R_{\mathrm{D}y}$. It means that $C^*R_xC+I = (I+U)(I+U)^*$ with the causal and causally invertible operator $I+U$. $[\cdot]^{(\tau)}$ stands for the τ-separation of the corresponding operator. We will not analyze formula (6.5) in detail here, but note that in the nonoperator terms it has the form

$$\hat{x}(t) = \int_{t_\mathrm{s}}^{t-\tau} g(t,t')\,\mathrm{d}v(t'), \tag{6.6}$$

where

$$g(t,t') = R_x(t,t')C(t') - \int_{t_\mathrm{s}}^{t'} R_x(t,t')C(t'')Q(t',t'')\,\mathrm{d}t'',$$

$$\mathrm{d}v(t) = \mathrm{d}y(t) - \left[\int_{t_\mathrm{s}}^{t} Q(t,t')\mathrm{d}y(t')\right]\mathrm{d}t, \tag{6.7}$$

and function $Q(\cdot,\cdot)$ solves the equation

$$Q(t,t') + \int_{t_\mathrm{s}}^{t} Q(t,t'')K(t'',t')\mathrm{d}t'' = K(t,t'), \qquad Q(t,t') = 0, \;\; t' > t, \tag{6.8}$$

and $K(\cdot,\cdot)$ is the function from (2.4). Note that $g(t,t') = 0$ when $t' > t-\tau$. See Theorem 2.1 in [5] for the details. (6.8) is a special case of the Hopf–Wiener equation.

Example 6.2. Formulas (6.6)–(6.8) for the optimal filter simplify considerably in the stationary case (see Example 2.6). The reason is that it is possible to reformulate the problem in frequency terms. Instead of the noncausal filter (2.7) the estimation is performed now with the causal filter

$$\hat{x}(t) = \int_{-\infty}^{t-\tau} h(t-t')\mathrm{d}y(t'), \qquad t \in \mathbb{R}. \tag{6.9}$$

In terms of the symbols of the involved operators, the Bode–Shannon formula (6.5) acquires the form

$$H_{opt}(\omega) = [G_{x\mathrm{D}y}(\omega)[1 + U(-\omega)]^{-1}]^{(\tau)}[1 + U(\omega)]^{-1}, \tag{6.10}$$

where $U(\cdot)$ is analytic in the lower half space and can be found from the factorization condition

$$C^* G_y(\omega) C + 1 = [1 + U(\omega)][1 + U(-\omega)] \tag{6.11}$$

(see Theorem 2.4 in [5]). τ-separation $[\cdot]^{(\tau)}$ in this case is determined by

$$\tilde{L}(\omega) = \int_{-\infty}^{\infty} e^{-i\omega t} L(t)\mathrm{d}t \to [\tilde{L}(\omega)]^{(\tau)} = \int_{\tau}^{\infty} e^{-i\omega t} L(t)\mathrm{d}t. \tag{6.12}$$

There are effective algorithms for the factorization and τ-separation in this case [5].

Example 6.3. Let us discuss finally a relation with the Kalman–Bucy filter. In many applications it is often assumed that the nonobservable component x of the partially observable process satisfies the stochastic equation

$$\mathrm{d}x(t) = A(t)x(t)\mathrm{d}t + B(t)\mathrm{d}w'(t), \tag{6.13}$$

where $A(\cdot)$, $B(\cdot)$ are continuous and bounded on $\mathbb{T}$ matrix valued functions, $w'(\cdot)$ is the standard Gaussian–Wiener process independent of the observation error $w(\cdot)$ in (2.1). If the initial state $x(0)$ is given, formulas (2.1), (6.13) uniquely define the correlation operator of the partially observable process $z = \binom{x}{\mathrm{D}y}$ (D is the generalized derivative here). This provides a principal possibility to construct the optimal τ-causal filter. However, it is sometimes problematic to use equations (6.6)–(6.8) not only because it is complicated to calculate the weight function $g(\cdot,\cdot)$, but also because one is required to store all the previously calculated values of the process. At the same time, the recursive approach of (2.1), (6.13) allows us to derive recursive formulas for the optimal filters. Such formulas do not require us the computationally complicated factorization procedure. They also do not require us to store the whole "history" of the process for further calculations. In the case of classical filtration ($\tau = 0$) these equations are the well-known Kalman–Bucy filter. We will not give these equations here since they can be found in almost any book on recursive filtering; see, for example, [5]. In the problems with $\tau \neq 0$ such formulas can also be derived and they can be viewed as a generalization of the Kalman–Bucy filter. However, this is a large independent topic and we will not pursue it here. We only note that in the derivation of such formulas the Bode–Shannon representation (6.5) of the optimal filter turns out to be very useful.

REFERENCES

[1] H. W. Bode and C. E. Shannon, *A simplified derivation of linear least square smoothing and prediction theory*, Proc. I.R.E., 38 (1950), pp. 417–425.

[2] D. K. Faddeev and V. N. Faddeeva, *Calculation Methods of Linear Algebra*, Fizmatgiz, Moscow, Leningrad, 1963 (in Russian).

[3] A. Feintuch and R. Saeks, *System Theory: A Hilbert Space Approach*, Academic Press, New York, 1982.

[4] V. N. Fomin, *Operator Methods of the Random Processes Linear Filtering Theory*, St. Petersburg University Publisher, St. Petersburg, 1995 (in Russian).

[5] V. N. Fomin, *Optimal Filtering.* Vol. 1: *Filtering of Stochastic Processes*, Kluwer, Dordrecht, The Netherlands, 1998.

[6] V. N. Fomin and M. V. Ruzhansky, *Mathematical Constructions in Optimal Linear Filtering Theory*, Tech. Rep. 948, Dept. of Math., Utrecht Univ., Utrecht, The Netherlands, 1996.

[7] O. G. Gorshkov and V. N. Fomin, *Operator approaches to time series filtering problem*, Vestnik S.-Peterburg. Univ., Ser. 1, Mat. Mekh. Astronom., 4 (1993), pp. 16–21 (in Russian).

[8] A. N. Kolmogorov, *Interpolation and extrapolation of the stationary random sequences*, Izv. AN SSSR Math., 5 (1941), pp. 3–14 (in Russian).

[9] O. A. Petrov and V. N. Fomin, *Linear Filtering of Random Processes*, LGU, Leningrad, 1991 (in Russian).

[10] M. V. Ruzhansky and V. N. Fomin, *The optimal filter construction with the quadratic quality functional of general form*, Vestnik S.-Peterburg. Univ., Ser. 1, Mat. Mekh. Astronom., 4 (1995), pp. 50–56, (in Russian; English translation in Vestnik St. Petersburg Univ. Math. 28 (4) (1996), pp. 42–46.

[11] N. Wiener, *The Extrapolation, Interpolation and Smoothing of Stationary Time Series with Engineering Applications*, Academic Press, New York, 1949.

SIAM J. CONTROL OPTIM.
Vol. 38, No. 5, pp. 1353–1368

H_2 AND H_∞ ROBUST FILTERING FOR DISCRETE-TIME LINEAR SYSTEMS*

JOSÉ C. GEROMEL†, JACQUES BERNUSSOU‡, GERMAIN GARCIA‡, AND MAURÍCIO C. DE OLIVEIRA†

Abstract. This paper investigates robust filtering design problems in H_2 and H_∞ spaces for discrete-time systems subjected to parameter uncertainty which is assumed to belong to a convex bounded polyhedral domain. It is shown that, by a suitable change of variables, both design problems can be converted into convex programming problems written in terms of linear matrix inequalities (LMI). The results generalize the ones available in the literature to date in several directions. First, all system matrices can be corrupted by parameter uncertainty and the admissible uncertainty may be structured. Then, assuming the order of the uncertain system is known, the optimal guaranteed performance H_2 and H_∞ filters are proven to be of the same order as the order of the system. Comparisons with robust filters for systems subjected to norm-bounded uncertainty are provided in both theoretical and practical settings. In particular, it is shown that under the same assumptions the results here are generally better as far as the minimization of a guaranteed cost expressed in terms of H_2 or H_∞ norms is considered. Some numerical examples illustrate the theoretical results.

Key words. linear systems, discrete-time systems, parameter uncertainty, filtering, linear matrix inequalities

AMS subject classifications. 93C05, 93C55, 93E11, 93E25

PII. S0363012997327379

1. Introduction.

State estimation has been one of the fundamental issues in the control area and there have been a lot of works following those of Kalman (in the stochastic framework [1]) and Luenberger (in the deterministic one [3, 12]). For control purpose, the will to estimate an unmeasurable linear combination of the state variables is clearly justified, for instance, by the *separation principle* well known in the case of models without uncertainty. For systems subjected to parameter uncertainty no such principle has been stated; however, there has been an increasing interest in the robust H_∞ literature and, aside from the robust control area, a large number of papers have addressed the dual problem of robust state estimation. A special issue of the Journal of Nonlinear and Robust Control was devoted to this problem in 1996 ([10]; see also [13] and the references therein).

In the late contributions to robust filtering, the problem is formulated in a Kalman stochastic-like context where the state estimation of uncertain dynamic systems in the presence of process noise is based upon noisy measurements. The filter determination is carried out by defining a suitable performance index in terms of the state estimation error variance. Fundamentally, two kinds of performance indexes have been considered according to the a priori assumptions on the input noise. In the classical H_2 filtering approach the noise characteristics are known leading to the minimization of the H_2 norm of the transfer function from the process noise to the estimation error.

*Received by the editors September 17, 1997; accepted for publication (in revised form) October 18, 1999; published electronically May 11, 2000. This research was supported by grants from "Fundação de Amparo à Pesquisa do Estado de São Paulo—FAPESP" and "Conselho Nacional de Desenvolvimento Científico e Tecnológico—CNPq," Brazil.

http://www.siam.org/journals/sicon/38-5/32737.html

†LAC-DT / School of Electrical and Computer Engineering, UNICAMP, CP 6101, 13081-970, Campinas, SP, Brazil (geromel@dt.fee.unicamp.br).

‡LAAS-CNRS / 7, Avenue du Colonel Roche, 31077, Cedex 4, Toulouse, France.

For uncertain systems, an upper bound on the previous performance index is established in terms of a single quadratic Lyapunov function, which naturally motivates the denomination of quadratic guaranteed cost state estimation [9, 13, 19]. More recently, the H_∞ filtering approach has been developed from the loose assumption of boundedness of the noise variance. In this case, the performance index to be minimized being the worst case H_∞ norm from the process noise to the estimation error [11, 16, 18]. All contributions cited rely on a particular uncertainty representation, namely, *norm-bounded uncertainty*. In this case, the mathematical model of the uncertain system exhibits explicitly a nominal model located at the center of the hyper ellipsoid of uncertainty in the parameter space. The results developed for the filtering problem are quite close to the ones derived in the robust control area due to the dual relationship of both problems. In this sense, one can classify the results and solvability conditions according to the Riccati-type equations as in [3, 20] or the linear matrix inequalities (LMI) as in [4]. Some comments concerning the two problems can be made for the mixed H_2/H_∞ filtering approach [8, 10].

Taking into account convex polytopic type uncertainty, [7] raises some difficulties linked to the fact that we are faced to structured uncertainty and the system representation does not explicitly exhibit any particular nominal model. For instance, the necessary and sufficient stabilizability conditions for dynamic output feedback control have only recently received some answer in terms of nonlinear matrix inequalities [5].

In this paper, we solve the state estimation problem for an uncertain discrete-time system under convex polytopic uncertainty addressing the H_2 as well as the H_∞ settings. It is shown that by restricting our attention to the class of linear and time-invariant filters, the filtering problem can be solved using the now classical LMI machinery [2]. Although a great effort has been made toward the synthesis of robust control for convex bounded uncertain systems [7, 5, 14, 15], in the literature to date there is no available result for linear filtering design of systems subjected to this class of parameter uncertainty. The lack of results on the robust filtering problem for convex polytopic uncertain systems appears to be due to the fact that the estimation error cannot be bounded by means of the solution of a Riccati-like equation. Here, it is considered that all matrices, namely, the state, input, and output matrices, are uncertain parameter dependent and also that structured uncertainties are allowed. A very interesting case of this new potentiality is the design of decentralized filters [6, 17]. Moreover, our approach allows to prove that even under parametric uncertainty the optimal filter order equals the order of the system supposed to be known. To ease the presentation only the stationary case is treated in detail, although the same lines can be used to deal with the nonstationary case.

The notation used throughout is as follows. Capital letters denote matrices and small letters denote vectors. For scalars we use small Greek letters. For matrices or vectors $(')$ indicates transpose. For symmetric matrices $X > 0$ (≥ 0) indicates that X is positive definite (nonnegative definite). Finally, for square matrices trace$[X]$ denotes the trace function of X being equal to the sum of its eigenvalues. For a transfer function $T(\zeta)$ analytic outside the unit circle, $\|T(\zeta)\|_2$ and $\|T(\zeta)\|_\infty$ denote the standard H_2 and H_∞ norms, respectively. Furthermore, for the sake of easing the notation of partitioned symmetric matrices the symbol $(\bullet)'$ denotes generically each of its symmetric blocks.

2. Problem statement. Let us consider the following linear time-invariant system

$$x(k+1) = Ax(k) + Bw(k), \tag{2.1}$$

$$y(k) = Cx(k) + Dw(k), \tag{2.2}$$

$$z(k) = Lx(k), \tag{2.3}$$

where $x \in R^n$ is the state, $w \in R^m$ is the zero mean, white noise input with identity power spectrum density matrix, $y \in R^r$ is the measured output, and $z \in R^s$ is the vector to be estimated. It is assumed that

1. All matrix dimensions are known.
2. Matrix $M \in R^{(n+r)\times(n+m)}$ defined as

$$M := \begin{bmatrix} A & B \\ C & D \end{bmatrix} \tag{2.4}$$

is unknown but it belongs to a given convex bounded polyhedral domain $\mathcal{D}_c$. Hence, from [7] each uncertain matrix of this set can be written as an unknown convex combination of N given extreme matrices $M_1, M_2, \ldots, M_N$, that is, $M \in \mathcal{D}_c$ if and only if

$$M = \sum_{i=1}^{N} \lambda_i M_i \tag{2.5}$$

holds for some $\lambda_1 \geq 0, \lambda_2 \geq 0, \ldots, \lambda_N \geq 0$ such that $\lambda_1 + \cdots + \lambda_N = 1$.

3. Matrix L is known.

It is important to make clear that the mathematical description of the convex polytopic domain $\mathcal{D}_c$ is sufficiently general to include, as a particular case, many uncertain systems with practical appealing. For instance, only some elements of M may be unknown and several elements of M may depend on the same unknown parameter. These two situations can easily be accommodated by a proper choice of the set of extreme matrices.

At this point, it is important to compare the above uncertainty description characterized by the convex set $\mathcal{D}_c$ with norm-bounded uncertainty as considered in [13, 18]. A system subjected to norm-bounded parameter uncertainty is such that matrix M belongs to the set $\mathcal{D}_n$ and each matrix is written in the form

$$M = M_0 + H\Omega E \tag{2.6}$$

for some matrix $\Omega \in R^{p\times q}$ such that $\Omega'\Omega \leq I$, where M_0 defines the nominal system and $H \in R^{(n+r)\times p}$ and $E \in R^{q\times(n+m)}$ are known matrices. From the Schur complement formula [2], it is seen that

$$\Omega'\Omega \leq I \Longleftrightarrow \begin{bmatrix} I & \Omega \\ (\bullet)' & I \end{bmatrix} \geq 0, \tag{2.7}$$

which makes clear that the uncertainty represented by matrix Ω belongs to a convex set and the same is true for M since it depends affinely on the matrix Ω. The conclusion is that the set $\mathcal{D}_n$ is convex but not necessarily a polyhedral set. In some important cases (see the examples in section 6) $\mathcal{D}_n$ is also polyhedral. This occurs when the uncertain matrix presents some a priori known structure, as, for instance, $\Omega'\Omega$ diagonal. In these situations it is always possible to determine a finite number of extreme matrices M_i, $i = 1, \ldots, N$ such that $\mathcal{D}_n = \mathcal{D}_c$. In the general case this equality does not hold. However, $\mathcal{D}_n$ being convex, it is again possible to determine N matrices Ω_i, $i = 1, \ldots, N$, on the boundary of the LMI (2.7) such that with $M_i = M_0 + H\Omega_i E$, $i = 1, \ldots, N$, there holds

$$\mathcal{D}_c \subset \mathcal{D}_n, \tag{2.8}$$

and $\mathcal{D}_c$ tends toward $\mathcal{D}_n$ as the number of extreme matrices N increases. It is clear that the contrary also holds since due to convexity it is always possible to determine N matrices Ω_i, $i = 1, \ldots, N$, not on the boundary of the LMI (2.7) such that $\mathcal{D}_n \subset \mathcal{D}_c$ and $\mathcal{D}_c$ tend toward $\mathcal{D}_n$ as the number of extreme matrices N increases. Hence, for N sufficiently large, the set $\mathcal{D}_n$ may be replaced by $\mathcal{D}_c$ within any desired precision. Of course, the main advantage in doing this is that $\mathcal{D}_c$ can accommodate many particular structures of the uncertain matrix which cannot be directly exploited, without introducing conservatism, with $\mathcal{D}_n$. However, it is important to point out that by adopting this approximation, the number of LMI to be handled may be very large.

To make this point clear let us consider the uncertain system of [18] which is more general than the one considered in [13]. We have $A = A_0 + H_1\Omega E$, $B = B_0$, $C = C_0 + H_2\Omega E$, and $D = D_0$ which can be written in the form (2.6), that is,

$$M = \begin{bmatrix} A_0 & B_0 \\ C_0 & D_0 \end{bmatrix} + \begin{bmatrix} H_1 \\ H_2 \end{bmatrix} \Omega \begin{bmatrix} E & 0 \end{bmatrix}. \tag{2.9}$$

However, if we want to consider a more general uncertain model defined by $A = A_0 + \bar{H}_1\Omega_1\bar{E}_1$, $B = B_0$, $C = C_0 + \bar{H}_2\Omega_2\bar{E}_2$, and $D = D_0$, then we now must write

$$M = \begin{bmatrix} A_0 & B_0 \\ C_0 & D_0 \end{bmatrix} + \begin{bmatrix} \bar{H}_1 & 0 \\ 0 & \bar{H}_2 \end{bmatrix} \begin{bmatrix} \Omega_1 & 0 \\ 0 & \Omega_2 \end{bmatrix} \begin{bmatrix} \bar{E}_1 & 0 \\ \bar{E}_2 & 0 \end{bmatrix}, \tag{2.10}$$

putting in evidence that (2.10) can be written as (2.9) with $H_1 := [\bar{H}_1\ 0]$, $H_2 := [0\ \bar{H}_2]$, and $E' := [\bar{E}_1'\ \bar{E}_2']$. However, the uncertain matrix Ω must be constrained to present the block diagonal structure

$$\Omega = \begin{bmatrix} \Omega_1 & 0 \\ 0 & \Omega_2 \end{bmatrix}. \tag{2.11}$$

As discussed before, this structure can approximately be treated with $\mathcal{D}_c$ and N sufficiently large. On the contrary, to apply directly the results of [13, 18] the a priori structure of Ω cannot be taken into account and consequently some conservatism on the robust filter design is necessarily introduced.

The problems to be dealt with in this paper are now formulated. We restrict ourselves to design an estimate $\hat{z}$ of z given by $\hat{z} = \mathcal{F} \cdot y$, where $\mathcal{F}$ is a linear, finite dimensional, and time-invariant operator producing at any time the estimation error $e := z - \hat{z}$. Defining $T_M(\zeta)$ the transfer function from the noise input w to the estimation error e, our goals are to solve the following design problems.

- H_2 *filtering problem:* Find a guaranteed estimation performance index $\rho_2(\cdot)$ such that

$$\sup_{M \in \mathcal{D}_c} \|T_M(\zeta)\|_2^2 \leq \rho_2(\mathcal{F}) \tag{2.12}$$

  and given $\mu > 0$ find all filters $\mathcal{F} \in \mathcal{C}$ such that (2.12) holds for $\rho_2(\mathcal{F}) = \mu$. Among all feasible filters, find the optimal one that minimizes $\rho_2(\mathcal{F})$ over $\mathcal{C}$.
- H_∞ *filtering problem:* Find a guaranteed estimation performance index $\rho_\infty(\cdot)$ such that

$$\sup_{M \in \mathcal{D}_c} \|T_M(\zeta)\|_\infty^2 \leq \rho_\infty(\mathcal{F}) \tag{2.13}$$

  and given $\mu > 0$ find all filters $\mathcal{F} \in \mathcal{C}$ such that (2.13) holds for $\rho_\infty(\mathcal{F}) = \mu$. Among all feasible filters, find the optimal one that minimizes $\rho_\infty(\mathcal{F})$ over $\mathcal{C}$.

In the above problems, the feasible set $\mathcal{C}$ is used to impose a particular class of linear, finite dimensional, and causal operators. In the present case, we consider $\mathcal{C}$ as the set of all linear time-invariant operators with state space realization of the form

$$\hat{x}(k+1) = A_f\hat{x}(k) + B_f y(k), \tag{2.14}$$

$$\hat{z}(k) = C_f\hat{x}(k), \tag{2.15}$$

where the matrices $A_f \in R^{n_f\times n_f}$, $B_f \in R^{n_f\times r}$, and $C_f \in R^{s\times n_f}$ and the scalar $n_f > 0$ are to be determined. In other words $\mathcal{C}$ is taken as the set of all linear time-invariant operators of any order and with strictly proper transfer functions. Moreover, it is considered that the initial condition of system (2.1) as well as the initial condition of the filter (2.14) are both zero. Connecting the filter to the system (2.1)−(2.3) we can write the transfer function $T_M(\zeta)$ as

$$T_M(\zeta) := \tilde{C}(\zeta I - \tilde{A})^{-1}\tilde{B}, \tag{2.16}$$

where matrices $\tilde{A}$, $\tilde{B}$, and $\tilde{C}$ of compatible dimensions are given by

$$\tilde{A} := \begin{bmatrix} A & 0 \\ B_fC & A_f \end{bmatrix},\ \tilde{B} := \begin{bmatrix} B \\ B_fD \end{bmatrix},\ \tilde{C} := \begin{bmatrix} L & -C_f \end{bmatrix}. \tag{2.17}$$

Some well-known results extensively used in this paper, namely, the Schur complement and partitioned LMI, can be found in [2, 14].

3. H_2 filtering. This section is devoted to H_2 filter design, as stated before. First we analyze the classical situation where all system parameters are precisely known, that is, matrix $M \in \mathcal{D}_c$ is arbitrary but fixed and $n_f = n$. In that which follows the optimal robust H_2 filter is obtained. Its dimension is proved to be equal to that of the uncertain system and a comparison with the results available in the literature to date is provided.

It is well known that for $\mu > 0$ the inequality $\|T_M(\zeta)\|_2^2 < \mu$ holds if and only if there exist symmetric matrices $\tilde{P}$ and W such that

$$\text{trace}[W] < \mu,\ \begin{bmatrix} \tilde{P} & \tilde{P}\tilde{C}' \\ (\bullet)' & W \end{bmatrix} > 0,\ \begin{bmatrix} \tilde{P} & \tilde{A}\tilde{P} & \tilde{B} \\ (\bullet)' & \tilde{P} & 0 \\ (\bullet)' & (\bullet)' & I \end{bmatrix} > 0, \tag{3.1}$$

where it is important to notice that in the above problem all constraints are taken strictly (see [3]). Hence, the minimum value of μ in (3.1) gives $\|T_M(\zeta)\|_2^2 = \mu+\epsilon$, where $\epsilon > 0$ is an amount arbitrarily small defined by the designer during the optimization process.

Our main purpose is, if possible, to convert the nonlinear matrix inequality appearing in (3.1) into LMI. If this goal is accomplished, the H_2 filter design problem turns out to be a convex programming problem which can be solved by very efficient numerical methods. To this end, let us partition $\tilde{P}$ and its inverse as

$$\tilde{P} := \begin{bmatrix} X & U \\ (\bullet)' & \hat{X} \end{bmatrix},\ \tilde{P}^{-1} := \begin{bmatrix} Y & V \\ (\bullet)' & \hat{Y} \end{bmatrix}, \tag{3.2}$$

where $X, Y \in R^{n\times n}$ and $\hat{X}, \hat{Y} \in R^{n_f\times n_f}$ are all symmetric and positive definite matrices. Multiplying the first row of $\tilde{P}$ by the first column of $\tilde{P}^{-1}$ reveals that $XY + UV' = I$, and taking into account the partition of $\tilde{P}$ and that of its inverse we

get $Y^{-1} = X - U\hat{X}^{-1}U'$. The key observation is that for any given symmetric and positive definite matrices such that $X > Y^{-1}$ and U square ($n_f = n$) and nonsingular, then from the first equality above we have V also nonsingular. Consequently, due to the second equality it is always possible to find $\hat{X} > 0$ ensuring that $\tilde{P} > 0$. From this partition of matrix $\tilde{P}$ let us introduce the following one-to-one change of variables:

$$\begin{bmatrix} A_f & B_f \\ C_f & 0 \end{bmatrix} := \begin{bmatrix} V & 0 \\ 0 & I \end{bmatrix}^{-1} \begin{bmatrix} Q & F \\ G & 0 \end{bmatrix} \begin{bmatrix} U'X^{-1} & 0 \\ 0 & I \end{bmatrix}^{-1}, \tag{3.3}$$

where the indicated inverses exist due to the fact that X is symmetric and positive definite and matrices V and U are both nonsingular. It is interesting to observe that in the above parameterization, one of matrices V or U can be freely defined by the designer without any loss of generality. This claim is easily seen from the filter transfer function

$$\begin{aligned} F_f(\zeta) &:= C_f \left(\zeta I - A_f\right)^{-1} B_f \\ &= GX \left(\zeta VU' - QX\right)^{-1} F, \end{aligned} \tag{3.4}$$

which depends only on the product of matrices V and U' being given by $VU' = I - YX$. The importance of this fact is that by choosing one of these matrices, we can get a filter with a particular state space realization. This point will be addressed in more detail in that which follows. Denoting $Z := X^{-1}$, the next theorem gives a partial solution, expressed in terms of LMI, to the H_2 filtering problem stated in the previous section for $M \in \mathcal{D}_c$ fixed and $n = n_f$.

THEOREM 1. *Let $M \in \mathcal{D}_c$ be given. All filters $\mathcal{F} \in \mathcal{C}$ such that $n = n_f$ and $\|T_M(\zeta)\|_2^2 < \mu$ are given by* (3.3)*, where U and V are full rank matrices such that $UV' = I - Z^{-1}Y$. Moreover, the scalar μ, matrices Q, G, F, and the symmetric matrices Y, Z, W satisfy the LMI*

$$\text{trace}[W] < \mu, \tag{3.5}$$

$$\begin{bmatrix} Z & Z & L' - G' \\ (\bullet)' & Y & L' \\ (\bullet)' & (\bullet)' & W \end{bmatrix} > 0, \tag{3.6}$$

$$\begin{bmatrix} Z & Z & ZA & ZA & ZB \\ (\bullet)' & Y & YA + FC + Q & YA + FC & YB + FD \\ (\bullet)' & (\bullet)' & Z & Z & 0 \\ (\bullet)' & (\bullet)' & (\bullet)' & Y & 0 \\ (\bullet)' & (\bullet)' & (\bullet)' & (\bullet)' & I \end{bmatrix} > 0. \tag{3.7}$$

Proof. From the fact that $n_f = n$ and matrices U and V are nonsingular the relation (3.3) is a one-to-one transformation. Defining the square and full rank matrix

$$\tilde{J} := \begin{bmatrix} X^{-1} & Y \\ 0 & V' \end{bmatrix} \tag{3.8}$$

it can be verified that the second inequality in (3.1) multiplied to the left by the full rank matrix $J' := \text{diag}[\tilde{J}', I]$ and to the right by J provides the LMI (3.6). Furthermore, doing the same to the third inequality in (3.1) with matrix $J := \text{diag}[\tilde{J}, \tilde{J}, I]$ we get the LMI (3.7), which together with (3.5) proves the proposed theorem. ∎

Notice that as in [7, 6] the main idea to get the above result stems from the definition of the new set of variables (3.3) which converts the problem under consideration to a convex feasibility problem expressed in terms of LMI only. If the LMI given in Theorem 1 are feasible, then a particular filter state space realization is readily obtained. For instance, the choice $V = V' = -Y$ leads to $U'Z = I - Y^{-1}Z$ in which case (3.3) provides the feasible filter

$$C_f = G\left(I - Y^{-1}Z\right)^{-1},\ A_f = -Y^{-1}Q\left(I - Y^{-1}Z\right)^{-1},\ B_f = -Y^{-1}F. \tag{3.9}$$

It is interesting to observe that the classical H_2 optimal filter can be recovered from the LMI (3.5)−(3.7) if we replace these inequalities by nonstrict ones and restrict our attention to the filters generated from $Z \to 0$. Doing so, to keep feasibility we must impose

$$G = L,\ Q = -YA - FC, \tag{3.10}$$

which together with the choice $W = LY^{-1}L'$ enables us to write

$$\operatorname{trace}[LY^{-1}L'] < \mu,\ \begin{bmatrix} Y & YA + FC & YB + FD \\ (\bullet)' & Y & 0 \\ (\bullet)' & (\bullet)' & I \end{bmatrix} > 0. \tag{3.11}$$

Considering for simplicity that $BD' = 0$ and $DD' = I$, defining $P := Y^{-1}$, and choosing $F = -YAPC'(CPC' + I)^{-1}$ we have

$$\begin{aligned} &\min_{P>0} \operatorname{trace}[LPL'], \\ &APA' - P - APC'(CPC' + I)^{-1}CPA' + BB' < 0, \end{aligned} \tag{3.12}$$

which comes from the Schur complement of inequality (3.11). From the well-known results on the monotonicity of the Riccati equation, it is clear that the optimal solution to problem (3.12) is arbitrarily close to the symmetric and nonnegative definite solution of the algebraic discrete Riccati equation

$$APA' - P - APC'(CPC' + I)^{-1}CPA' + BB' = 0. \tag{3.13}$$

In addition, from (3.9) the optimal filter state space realization is given by matrices $C_f = L$, $A_f = -Y^{-1}Q = A - B_fC$ and $B_f = -Y^{-1}F = APC'(CPC' + I)^{-1}$ which we recognize as being the Kalman filter. The above manipulations prove that the Kalman filter is on the boundary of the LMI (3.5)−(3.7) and hence is close (within any prespecified precision) to the optimal solution of the convex problem consisting of the minimization of μ under the LMI constraints provided in Theorem 1. Of course the main goal of Theorem 1 is to treat not precisely known systems but systems subjected to parameter uncertainty as will become clear in that which follows.

Referring back to the result of Theorem 1, it is interesting to notice that it is always possible to eliminate the matrix variable G since it appears only in the LMI (3.6). In fact, permuting the second and third columns and rows we get an equivalent LMI

$$\begin{bmatrix} Z & L' - G' & Z \\ (\bullet)' & W & L \\ (\bullet)' & (\bullet)' & Y \end{bmatrix} > 0, \tag{3.14}$$

which using the Schur complement can be rewritten in the equivalent form

$$\begin{bmatrix} Z - ZY^{-1}Z & L' - G' - ZY^{-1}L' \\ (\bullet)' & W - LY^{-1}L' \end{bmatrix} > 0, \tag{3.15}$$

and from [14] we are able to say that the LMI (3.6) can be replaced by

$$\begin{bmatrix} Y & L' \\ (\bullet)' & W \end{bmatrix} > 0, \quad \begin{bmatrix} Z & Z \\ (\bullet)' & Y \end{bmatrix} > 0 \tag{3.16}$$

together with $G = L(I - Y^{-1}Z)$. This is important because the number of free variables in problem (3.5)−(3.7) becomes smaller and we are sure that no conservatism is introduced. Moreover, this formula for G enables us to impose, with no loss of generality, that $C_f = L$ as already considered in [18] and the references therein. Indeed, this occurs if we set $U = U' = Z^{-1} - Y^{-1}$ since

$$\begin{aligned} C_f &= G(U'Z)^{-1} \\ &= L(I - Y^{-1}Z)Z^{-1}\left(Z^{-1} - Y^{-1}\right)^{-1} \\ &= L, \end{aligned} \tag{3.17}$$

and solving the equation $XY + UV' = I$ we get $V = V' = -Y$. These two matrices are of full rank as required to generate the optimal filter matrices from Theorem 1 variables. Hence, we have proven the following result.

THEOREM 2. *Let $M \in \mathcal{D}_c$ be given. All filters $\mathcal{F} \in \mathcal{C}$ such that $n = n_f$ and $\|T_M(\zeta)\|_2^2 < \mu$ are given by matrices*

$$C_f = L, \ A_f = -Y^{-1}Q\left(I - Y^{-1}Z\right)^{-1}, \ B_f = -Y^{-1}F, \tag{3.18}$$

where the scalar μ, matrices Q, F, and the symmetric matrices Y, Z, W satisfy the LMI (3.7) *and*

$$\text{trace}[W] < \mu, \quad \begin{bmatrix} Y & L' \\ (\bullet)' & W \end{bmatrix} > 0. \tag{3.19}$$

From the previous results it is not difficult to verify that there is no hope to determine a filter of dimension $n_f \neq n$ with smaller value of $\|T_M(\zeta)\|_2$ than those provided by Theorem 2. In fact, taking $n_f < n$, matrix $\tilde{J}$ in (3.8) is no longer square but still exhibits full row rank. From this, relaxing the strict character of the basic inequalities (3.6) and (3.7) and performing the same calculations as before we see that (3.6) and (3.7) still hold. However, it is no longer possible to solve $Y^{-1} = X - U\hat{X}^{-1}U'$ unless Y and X are such that $\text{rank}(X - Y^{-1}) = n_f$. This constraint is not convex and so convexity of the associated feasibility problem is lost. More important is the fact that if it exists the feasible reduced order filter will produce a greater or at most equal value of $\|T_M(\zeta)\|_2$. On the other hand, taking $n_f > n$ then matrix $\tilde{J}$ in (3.8) is not square but presents full column rank. So, to adopt the same reasoning as before, we need to replace it by the square and nonsingular matrix

$$\tilde{J} = \begin{bmatrix} X^{-1} & Y & 0 \\ 0 & V' & S \end{bmatrix}, \tag{3.20}$$

where S of compatible dimensions spans the null space of U. Adopting the same steps as before, we see that both inequalities (3.6) and (3.7) must hold once again,

and in the affirmative case, it is possible to construct a feasible filter which produces the same value of $\|T_M(\zeta)\|_2$ as the full order filter with $n = n_f$. Let us now turn our attention to the robust H_2 filter design. To this end we need the following result.

LEMMA 1. *Assume $\mathcal{F} \in \mathcal{C}$ is a given filter and consider the problem*

(3.21)
$$\rho_2(\mathcal{F}) := \min_{\tilde{P}} \left\{ \text{trace}[\tilde{C}\tilde{P}\tilde{C}'] \ : \ \begin{bmatrix} \tilde{P} & \tilde{A}_i\tilde{P} & \tilde{B}_i \\ (\bullet)' & \tilde{P} & 0 \\ (\bullet)' & (\bullet)' & I \end{bmatrix} > 0 \ , i = 1, 2, \ldots, N \right\},$$

where matrices $\tilde{A}_i$ and $\tilde{B}_i$ are the same as $\tilde{A}$ and $\tilde{B}$ with the submatrices A, B, C, and D of M, replaced by those of the extreme matrices

$$M_i := \begin{bmatrix} A_i & B_i \\ C_i & D_i \end{bmatrix}, \ i = 1, 2, \ldots, N. \tag{3.22}$$

If problem (3.21) *is feasible, then $\rho_2(\mathcal{F})$ is a valid upper bound to $\|T_M(\zeta)\|_2^2$ for all $M \in \mathcal{D}_c$, that is, it satisfies inequality* (2.12).

Proof. Consider that for some $\mathcal{F} \in \mathcal{C}$, problem (3.21) has been solved providing $\tilde{P} > 0$. For any $M \in \mathcal{D}_c$, there exists a convex combination such that

$$M = \sum_{i=1}^{N} \lambda_i \begin{bmatrix} A_i & B_i \\ C_i & D_i \end{bmatrix}. \tag{3.23}$$

Hence, keeping in mind the structures of matrices $\tilde{A}$ and $\tilde{B}$ given in (2.17) we get

$$\begin{bmatrix} \tilde{P} & \tilde{A}\tilde{P} & \tilde{B} \\ (\bullet)' & \tilde{P} & 0 \\ (\bullet)' & (\bullet)' & I \end{bmatrix} = \sum_{i=1}^{N} \lambda_i \begin{bmatrix} \tilde{P} & \tilde{A}_i\tilde{P} & \tilde{B}_i \\ (\bullet)' & \tilde{P} & 0 \\ (\bullet)' & (\bullet)' & I \end{bmatrix} > 0. \tag{3.24}$$

Applying to the above inequality the Schur complement formula together with (3.21), we have the upper bound to the H_2 norm of the transfer function $T_M(\zeta)$

$$\begin{aligned} \|T_M(\zeta)\|_2^2 &= \text{trace}\left[\sum_{k=0}^{\infty} \tilde{C}\tilde{A}^k\tilde{B}\tilde{B}'\tilde{A}'^k\tilde{C}'\right] \\ &< \text{trace}[\tilde{C}\tilde{P}\tilde{C}'] \\ &< \rho_2(\mathcal{F}) \ \text{ for all } \ M \in \mathcal{D}_c, \end{aligned} \tag{3.25}$$

which is the desired result. □

The upper bound to the H_2 norm of $T_M(\zeta)$ provided by Lemma 1 deserves some comments. In the case of precisely known parameters it coincides with the true value of $\|T_M(\zeta)\|_2^2$ and so does not introduce any conservatism. This case is obtained for $N = 1$ and the filter synthesis procedure of Theorem 2 is valid. For convex polytopic systems it suffices to include as many LMI as the number of extreme matrices needed to define $\mathcal{D}_c$. The other matrices in the interior of $\mathcal{D}_c$ are automatically considered. In this sense we can say that the worst situation certainly occurs in an extreme point of the uncertain domain. Finally, adopting the same reasoning as before, the following result is proven.

THEOREM 3. *The optimal H_2 robust filter which minimizes $\rho_2(\mathcal{F})$ for all $\mathcal{F} \in \mathcal{C}$ is a filter with dimension $n_f = n$ given by* (3.18)*, where matrices Q, F and the symmetric matrices Z, Y, W minimize μ under the constraints* (3.19) *and*

$$\begin{bmatrix} Z & Z & ZA_i & ZA_i & ZB_i \\ (\bullet)' & Y & YA_i + FC_i + Q & YA_i + FC_i & YB_i + FD_i \\ (\bullet)' & (\bullet)' & Z & Z & 0 \\ (\bullet)' & (\bullet)' & (\bullet)' & Y & 0 \\ (\bullet)' & (\bullet)' & (\bullet)' & (\bullet)' & I \end{bmatrix} > 0 \tag{3.26}$$

for all $i = 1, 2, \ldots, N$. Furthermore, the minimum H_2 guaranteed performance cost satisfying (2.12) *is given by $\rho_2(\mathcal{F}) = \text{trace}[W]$.*

Notice that the proof of Theorem 3 follows the same pattern as before. The structure of the constraint in (3.21) is identical to (3.1), where each submatrix of M is replaced by the corresponding submatrix of M_i, $i = 1, 2, \ldots, N$. At this point it is important to stress that the parameter uncertainty we can handle with Theorem 3 is more general than the very recent results of [13, 18].

Our purpose now is to compare the previous results of this section with the best ones available in the literature to date [13]. Given a domain $\mathcal{D}_n$ we assume the scalar N (possibly infinite) and the extreme matrices M_i, $i = 1, 2, \ldots, N$, are determined such that $\mathcal{D}_c = \mathcal{D}_n$.

The key observation is that the full order, i.e., $n_f = n$, robust filter matrices $(\bar{A}_f, \bar{B}_f, \bar{C}_f)$ provided in those references are given from the symmetric and nonnegative solutions of two algebraic discrete Riccati equations depending on the nominal model and on matrices H_1, H_2, and E in (2.9). Moreover, these solutions depend on a parameter $\varepsilon > 0$ which is optimized to get the best guaranteed performance. Hence, with the optimal value $\varepsilon^\star > 0$ a symmetric and nonnegative definite matrix $\bar{P}$ is determined such that

$$\|T_M(\zeta)\|_2^2 \le \text{trace}[\tilde{C}\bar{P}\tilde{C}'],\ \tilde{A}\bar{P}\tilde{A}' - \bar{P} + \tilde{B}\tilde{B}' \le 0 \ \text{ for all } \ M \in \mathcal{D}_n. \tag{3.27}$$

This solution presents two important properties. As expected $\bar{C}_f = L$ and $\bar{P}$ can be partitioned as

$$\bar{P} = \begin{bmatrix} P_1 & P_2 \\ P_2 & P_2 \end{bmatrix} \ge 0, \tag{3.28}$$

where P_1 and P_2 are symmetric and nonnegative definite matrices. Perturbing slightly matrices P_1 and P_2 if necessary in order that these matrices become positive definite and the inequalities in (3.27) and (3.28) become strict, choosing $M_i \in \mathcal{D}_n$ for $i = 1, 2, \ldots, N$ so as to produce $\mathcal{D}_c = \mathcal{D}_n$, and taking into account (3.27), we conclude that with

$$Z := P_1^{-1},\ U := P_2,\ Y = -V := (P_1 - P_2)^{-1} \tag{3.29}$$

there exist matrices Q and F easily determined from $(\bar{A}_f, \bar{B}_f)$ and (3.3) such that all LMI in (3.26) are feasible. In addition to $P_1 > P_2$ in (3.28), choosing $W > L(P_1 - P_2)L'$ the corresponding cost is given by

$$\text{Trace}[W] = \text{trace}[L(P_1 - P_2)L'] + \epsilon, \tag{3.30}$$

which differs from the minimum upper bound of (3.27) by an amount of order $\epsilon > 0$ which once again can be taken arbitrarily small. This proves that Theorem 3 generally

provides a robust filter with better or at least equal performance than [13]. Moreover, Theorem 3 applies to more general models subjected to parametric uncertainties and no unidimensional search (with respect to the parameter ε) is needed. These aspects will be illustrated by means of numerical examples.

4. H_∞ filtering. In this section the problem of H_∞ filtering design is addressed. The same main lines adopted to deal with H_2 filters are again used. First the classical H_∞ filter is analyzed for $M \in \mathcal{D}_c$ fixed and $n = n_f$. In that which follows the optimal guaranteed robust H_∞ filter is determined, and it is proven that it exhibits the same dimension of the uncertain plant. Finally a comparison with the very recent results of [18] is provided.

It is well known that the transfer function from the input noise to the estimation error satisfies the inequality $\|T_M(\zeta)\|_\infty^2 < \mu$ for $\mu > 0$ if and only if there exists a symmetric matrix $\tilde{P}$ such that

$$\begin{bmatrix} \tilde{P} & \tilde{A}\tilde{P} & \tilde{B} & 0 \\ (\bullet)' & \tilde{P} & 0 & \tilde{P}\tilde{C}' \\ (\bullet)' & (\bullet)' & I & 0 \\ (\bullet)' & (\bullet)' & (\bullet)' & \mu I \end{bmatrix} > 0. \tag{4.1}$$

In order to convert this nonlinear matrix inequality into an LMI we proceed by partitioning the matrix variable $\tilde{P}$ and its inverse as indicated in (3.2). Introducing again the matrix variable $Z := X^{-1}$ we have the following result.

THEOREM 4. *Let $M \in \mathcal{D}_c$ be given. All filters $\mathcal{F} \in \mathcal{C}$ such that $n = n_f$ and $\|T_M(\zeta)\|_\infty^2 < \mu$ are given by* (3.3) *where U and V are full rank matrices such that $UV' = I - Z^{-1}Y$. Moreover, the scalar μ, the matrices Q, G, F, and the symmetric matrices Y, Z satisfy the LMI*

$$\begin{bmatrix} Z & Z & ZA & ZA & ZB & 0 \\ (\bullet)' & Y & YA + FC + Q & YA + FC & YB + FD & 0 \\ (\bullet)' & (\bullet)' & Z & Z & 0 & L' - G' \\ (\bullet)' & (\bullet)' & (\bullet)' & Y & 0 & L' \\ (\bullet)' & (\bullet)' & (\bullet)' & (\bullet)' & I & 0 \\ (\bullet)' & (\bullet)' & (\bullet)' & (\bullet)' & (\bullet)' & \mu I \end{bmatrix} > 0. \tag{4.2}$$

Proof. From the fact that $n_f = n$ and matrices U and V are nonsingular the relation (3.3) is a one-to-one transformation. Defining the square and full rank matrix $\tilde{J}$ as indicated in (3.8), the multiplication of (4.1) to the left by $J' := \mathrm{diag}[\tilde{J}', \tilde{J}', I, I]$ and to the right by J provides the LMI (4.2). □

Choosing $V = V' = -Y$ we get $U'Z = I - Y^{-1}Z$ and (3.3) provides once again the filter state space realization

$$C_f = G\left(I - Y^{-1}Z\right)^{-1},\ A_f = -Y^{-1}Q\left(I - Y^{-1}Z\right)^{-1},\ B_f = -Y^{-1}F, \tag{4.3}$$

from which it is interesting to observe that the H_∞ filter design problem differs from the H_2 one in the following relevant aspect concerning matrix G. In the previous design it could be eliminated from the associated LMI constraint by applying the Schur complement together with the result of [14]. If we do the same to (4.2) it is again possible to split that inequality into two LMI. However, the value of matrix G will depend on the system matrices M and L. This solution is not valid for robust filter design since by assumption matrix M is not known.

As for H_2 filters, it is possible to recover the H_∞ central filter by allowing $Z \to 0$. Considering for ease that $BD' = 0$, $DD' = I$, and (4.2) with nonstrict inequality, the possible choice of $Z \to 0$ leads necessarily to $G = L$ and $Q = -YA - FC$. In this case the remaining variables have to satisfy the LMI

$$\begin{bmatrix} Y & YA+FC & YB+FD & 0 \\ (\bullet)' & Y & 0 & L' \\ (\bullet)' & (\bullet)' & I & 0 \\ (\bullet)' & (\bullet)' & (\bullet)' & \mu I \end{bmatrix} > 0. \tag{4.4}$$

To get the central H_∞ filter we define the new variable $P := (Y - \mu^{-1}L'L)^{-1}$ and impose $F = -YAPC'(CPC' + I)^{-1}$. Performing the Schur complement to inequality (4.4) we have

$$APA' - \left(P^{-1} + \mu^{-1}L'L\right)^{-1} - APC'(CPC' + I)^{-1}CPA' + BB' < 0. \tag{4.5}$$

Consequently, under the assumption that (4.2) is feasible, that is, it admits an interior point, there always exits a feasible solution arbitrarily close to the symmetric and positive definite solution of the algebraic discrete Riccati equation

$$APA' - \left(P^{-1} + \mu^{-1}L'L\right)^{-1} - APC'(CPC' + I)^{-1}CPA' + BB' = 0, \tag{4.6}$$

in which case the state space representation of the associated filter is given by matrices $C_f = L$, $A_f = A - B_f C$, and $B_f = APC'(CPC' + I)^{-1}$ being the central H_∞ filter which clearly satisfies $\|T_M(\zeta)\|_\infty^2 < \mu$. Furthermore, the minimum value of μ such that a solution $P > 0$ to (4.6) exists provides a central filter with optimal H_∞ performance.

Referring back to the general problem, it is important to stress that nothing is gained in terms of the performance index under consideration if we consider $n_f \neq n$. Indeed, for $n_f < n$, besides (4.2), a supplementary nonconvex constraint must be added which may cause performance deterioration. On the other hand for $n_f > n$ the same LMI (4.2) still must hold and so it is always possible to reduce the filter order to $n_f = n$ keeping the same performance level. We are now in position to analyze the robust H_∞ filter.

LEMMA 2. *Assume $\mathcal{F} \in \mathcal{C}$ is a given filter and consider the problem*

$$\rho_\infty(\mathcal{F}) := \min_{\tilde{P},\mu} \left\{ \mu \ : \ \begin{bmatrix} \tilde{P} & \tilde{A}_i\tilde{P} & \tilde{B}_i & 0 \\ (\bullet)' & \tilde{P} & 0 & \tilde{P}\tilde{C}' \\ (\bullet)' & (\bullet)' & I & 0 \\ (\bullet)' & (\bullet)' & (\bullet)' & \mu I \end{bmatrix} > 0 \ , i = 1, 2, \ldots, N \right\}, \tag{4.7}$$

where matrices $\tilde{A}_i$ and $\tilde{B}_i$ are the same as $\tilde{A}$ and $\tilde{B}$ with the submatrices A, B, C, and D of M, replaced by those of the extreme matrices $M_1, M_2, \ldots, M_N$. If problem (4.7) *is feasible, then $\rho_\infty(\mathcal{F})$ is a valid upper bound to $\|T_M(\zeta)\|_\infty^2$ for all $M \in \mathcal{D}_c$, that is, it satisfies inequality* (2.13).

The proof of Lemma 2 follows the same lines of that for Lemma 1 and is thus omitted. Notice that the index $i = 1, 2, \ldots, N$ does not appear in matrix $\tilde{C}$ due to our assumption that matrix L is fixed and known. For $N = 1$ the upper bound provided by Lemma 2 matches with $\|T_M(\zeta)\|_\infty^2$ which means that no conservatism is introduced. The next theorem is the main result of this section and states that the optimal H_∞ robust filter can be determined from a convex programming problem.

THEOREM 5. *The optimal H_∞ robust filter which minimizes $\rho_\infty(\mathcal{F})$ for all $\mathcal{F} \in \mathcal{C}$ is a filter with dimension $n_f = n$ and state space realization given by* (4.3) *where matrices Q, G, F and the symmetric matrices Z, Y minimize μ under the constraint*

$$\begin{bmatrix} Z & Z & ZA_i & ZA_i & ZB_i & 0 \\ (\bullet)' & Y & YA_i + FC_i + Q & YA_i + FC_i & YB_i + FD_i & 0 \\ (\bullet)' & (\bullet)' & Z & Z & 0 & L' - G' \\ (\bullet)' & (\bullet)' & (\bullet)' & Y & 0 & L' \\ (\bullet)' & (\bullet)' & (\bullet)' & (\bullet)' & I & 0 \\ (\bullet)' & (\bullet)' & (\bullet)' & (\bullet)' & (\bullet)' & \mu I \end{bmatrix} > 0 \tag{4.8}$$

for all $i = 1, 2, \ldots, N$. Furthermore, the minimum H_∞ guaranteed performance cost satisfying (2.13) *is given by $\rho_\infty(\mathcal{F}) = \mu$.*

Paper [18] proposes the most general method for H_∞ filter design dealing with norm-bounded uncertainty acting on matrices A and C of model (2.1)−(2.3). The procedure depends on the existence of a parameter $\varepsilon > 0$ such that two algebraic Riccati equations admit symmetric, nonnegative, and stabilizing solutions. In the affirmative case, a full order H_∞ filter defined by the triplet of matrices $(\bar{A}_f, \bar{B}_f, \bar{C}_f)$ with $\bar{C}_f = L$ is determined such that

$$\|T_M(\zeta)\|_\infty^2 < \mu,\ \tilde{A}\bar{P}\tilde{A}' - \left(\bar{P}^{-1} + \mu^{-1}\tilde{C}'\tilde{C}\right)^{-1} + \tilde{B}\tilde{B}' \leq 0 \ \text{ for all } \ M \in \mathcal{D}_n \tag{4.9}$$

hold for some matrix $\bar{P}$ exhibiting the partitioned structure (3.28) where P_1 and P_2 are symmetric and nonnegative definite matrices. Perturbing slightly matrices P_1 and P_2 if necessary and assuming that N extreme matrices M_i for all $i = 1, 2, \ldots, N$ have been determined such that $\mathcal{D}_c = \mathcal{D}_n$ then we can say that with (4.9) there exist matrices Q, G, and F easily determined from $(\bar{A}_f, \bar{B}_f, \bar{C}_f)$ and (3.3) such that all LMI in (4.8) are feasible. This means that the filter of [18] is always feasible for Theorem 5 but the contrary, of course, is not always true unless $\mathcal{D}_n$ is defined by a one-block uncertain matrix.

5. H_2 and H_∞ decentralized filtering. Another important point of the design procedure provided in the present paper concerns the H_2 and H_∞ filter structures. In signal and systems estimation when the overall system is described by a number of units coupled together by means of an interconnection network, it is of interest to know whether it is possible to connect local filters in order to estimate the local state variables [6, 17]. Assuming no parameter uncertainty exists, the model is given by (2.1)−(2.3), where B, C, D, and L but not A are block diagonal matrices. The goal is to determine a filter as (2.14)−(2.15) with A_f, B_f, and C_f block diagonal matrices of compatible dimensions. In this case the filter can be split into a set of local filters acting on each subsystem level. Recalling that the inverse of a block diagonal matrix is also a block diagonal matrix and that the product of block diagonal matrices is a block diagonal matrix, then (3.18) or more generally (4.3) reveals that our goal is accomplished provided we include in the H_2 and H_∞ filtering design problems these additional constraints: *matrices Z, Y, Q, G, and F are block diagonal.* Fortunately this corresponds to constrain to zero some entries of those matrices and convexity is preserved. Of course the same reasoning can be drawn for the design of robust filters.

6. Illustrative examples. In this section we solve the H_2 filter design problem for several systems of the form (2.1)−(2.3) to illustrate the most important aspects of the theory introduced so far. In all examples the uncertainty acts on matrix A

only but is of unstructured and structured types. The procedure given in [13], which is based on a necessary and sufficient condition for guaranteed cost state estimation, has been implemented and it is used for comparison purposes.

Consider a linear discrete-time system with

$$B = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix}, \; C = \begin{bmatrix} 1 & 0 \end{bmatrix}, \; D = \begin{bmatrix} 0 & 0 & \sqrt{2} \end{bmatrix}, \; L = \begin{bmatrix} 1 & 1 \end{bmatrix}. \tag{6.1}$$

Optimal H_2 filter: The nominal matrix $A = A_0$ is given by

$$A_0 = \begin{bmatrix} 0.9 & 0.1 \\ 0.01 & 0.9 \end{bmatrix}. \tag{6.2}$$

For $N = 1$ corresponding to the nominal system, Theorem 3 provides the Kalman optimal H_2 filter $\mathcal{F}_{kal}$ with the associated minimum cost $\rho_2(\mathcal{F}_{kal}) = 8.0759$ and state space realization

$$A_{kal} = \begin{bmatrix} 0.4427 & 0.1000 \\ -0.1615 & 0.9000 \end{bmatrix}, \; B_{kal} = \begin{bmatrix} 0.4573 \\ 0.1715 \end{bmatrix}, \; C_{kal} = \begin{bmatrix} 1 & 1 \end{bmatrix}. \tag{6.3}$$

Robust H_2 filter: Consider an uncertain system with $A = A_0 + \Delta A$ where A_0 is given by (6.2) and

$$\Delta A = \begin{bmatrix} 0 & 0.06\alpha \\ 0.05\beta & 0 \end{bmatrix} = \begin{bmatrix} 0.06 & 0 \\ 0 & 0.05 \end{bmatrix} \begin{bmatrix} \alpha & 0 \\ 0 & \beta \end{bmatrix} \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}, \tag{6.4}$$

where $|\alpha| \leq 1$ and $|\beta| \leq 1$. This is a two-block uncertainty which can be exactly described by the set $\mathcal{D}_c$ but not by the set $\mathcal{D}_n$ unless the uncertain matrix is taken as diagonal, that is, $\Omega := \text{diag}[\alpha, \beta]$. However, the results available in the literature to date [13, 18] cannot be directly applied unless the block diagonal structure of Ω is not imposed, in which case some conservatism is necessarily introduced. The design procedure of [13] with the best value for the parameter $\varepsilon = 1.5264e - 04$ provides a suboptimal guaranteed cost H_2 filter $\mathcal{F}_{sub}$ with state space realization

$$A_{sub} = \begin{bmatrix} 0.0335 & 0.1014 \\ -0.2551 & 0.9117 \end{bmatrix}, \; B_{sub} = \begin{bmatrix} 0.8667 \\ 0.2652 \end{bmatrix}, \; C_{sub} = \begin{bmatrix} 1 & 1 \end{bmatrix}. \tag{6.5}$$

On the other hand, applying Theorem 3 with $N = 4$ matrices corresponding to the extreme points of the uncertain domain we get the optimal guaranteed cost H_2 filter $\mathcal{F}_{opt}$ given by

$$A_{opt} = \begin{bmatrix} 0.0383 & 0.0982 \\ -0.3681 & 0.8986 \end{bmatrix}, \; B_{opt} = \begin{bmatrix} 0.8490 \\ 0.3779 \end{bmatrix}, \; C_{opt} = \begin{bmatrix} 1 & 1 \end{bmatrix}. \tag{6.6}$$

Table 6.1 shows, for each filter, the value of the guaranteed H_2 cost $\rho_2(\mathcal{F})$ as well as the maximum value of $\|T_M(\zeta)\|_2^2$ with respect to the matrix M which depends through (6.4) on the uncertain parameters α and β. As expected, the Kalman filter, which is optimal for the nominal system, is the worst under parametric uncertainty. The filter of [13] is suboptimal with respect to the guaranteed cost because, as we have said before, the two-block structure of the uncertainty matrix Ω cannot be considered in the design procedure proposed. It is interesting to observe that the filter determined from Theorem 3 is the better one with respect to guaranteed H_2 cost as well as the true worst case value of the H_2 estimation cost.

TABLE 6.1
H_2 filter performance: multiblock uncertainty.

Filter	$\mathcal{F}_{kal}$	$\mathcal{F}_{sub}$	$\mathcal{F}_{opt}$
$\rho_2(\mathcal{F})$	–	129.7915	100.0278
$\sup_{M\in\mathcal{D}_c}\|T_M(\zeta)\|_2^2$	49.4994	38.2183	30.0664

TABLE 6.2
H_2 filter performance: one-block uncertainty.

Filter	$\mathcal{F}_{kal}$	$N=2$	$N=4$	$N=8$	$\mathcal{F}_{opt}$
$\rho_2(\mathcal{F})$	–	9.6796	13.0219	13.0446	13.0446
$\sup_{M\in\mathcal{D}_n}\|T_M(\zeta)\|_2^2$	12.9427	–	–	11.8646	11.8646

Robust H_2 filter: Consider now an uncertain system with $A = A_0 + \Delta A$ where A_0 is given by (6.2) and

$$\Delta A = \begin{bmatrix} 0 & 0.06\alpha \\ 0 & 0.05\beta \end{bmatrix} = \begin{bmatrix} 0.06 & 0 \\ 0 & 0.05 \end{bmatrix} \begin{bmatrix} \alpha \\ \beta \end{bmatrix} \begin{bmatrix} 0 & 1 \end{bmatrix}, \tag{6.7}$$

where the uncertain parameters are such that $\alpha^2+\beta^2 \le 1$. This case of only one-block uncertainty is exactly described by the domain $\mathcal{D}_n$ with $\Omega' = [\alpha\ \beta]'$. Using [13] the following optimal guaranteed H_2 cost filter is obtained:

$$A_{opt} = \begin{bmatrix} 0.3521 & 0.1069 \\ -0.2211 & 0.9400 \end{bmatrix},\ B_{opt} = \begin{bmatrix} 0.5479 \\ 0.2311 \end{bmatrix},\ C_{opt} = \begin{bmatrix} 1 & 1 \end{bmatrix}. \tag{6.8}$$

In this case, the uncertainty domain cannot be exactly represented by the polytopic domain $\mathcal{D}_c$. However, Table 6.2 shows that with the extreme matrices calculated from

$$\begin{bmatrix} \alpha_i \\ \beta_i \end{bmatrix} = \begin{bmatrix} \cos(2\pi i/N) \\ \sin(2\pi i/N) \end{bmatrix},\ i = 1, 2, \ldots, N, \tag{6.9}$$

for $N = 8$ the problem given in Theorem 3 provides the same guaranteed cost optimal robust filter which, in addition, performs better that the nominal Kalman filter.

7. Conclusion. The robust filtering problem has been addressed in this paper in the case when convex polytopic uncertainty is present on the dynamic, input, and output matrices. Unlike Kalman-type approaches using Riccati equations that generally deal with unstructured (one-block) norm-bounded uncertainty, the uncertain type handled here may be highly structured and then encompasses most of the results available in the literature to date. The filtering problem has been solved using H_2 and H_∞ optimization formulations, both with LMI machinery. Indeed, the proposed filter parameterization enables us to define parametric optimization problems in terms of linear matrix inequalities, a feature which is now known to be very fortunate since it allows the use of very efficient numerical solvers. Some illustrative numerical examples have been developed to show the usefulness of the proposed approach and its superiority for dealing with structured (multiblock) uncertainty. The LMI tool can be exploited to solve efficiently complex problems with structured uncertainty (interval

as well as norm-bounded type) in the presence of convex constraints expressed in terms of LMI as, for instance, structural and integral quadratic constraints.

REFERENCES

[1] B. D. O. Anderson and J. B. Moore, *Optimal Filtering*, Prentice-Hall, Englewood Cliffs, NJ, 1979.

[2] S. P. Boyd, L. El Ghaoui, E. Feron, and V. Balakrishnan, *Linear Matrix Inequalities in System and Control Theory*, SIAM, Philadelphia, PA, 1994.

[3] P. Colaneri, J. C. Geromel, and A. Locatelli, *Control Theory and Design: An RH_2-RH_∞ Viewpoint*, Academic Press, New York, 1997.

[4] P. Gahinet and P. Apkarian, *A linear matrix inequality approach to H_∞ control*, Internat. J. Robust Nonlinear Control, 4 (1994), pp. 421–448.

[5] J. C. Geromel, J. Bernussou, and M. C. de Oliveira, *H_2 norm optimization with constrained dynamic output feedback controllers: Decentralized and reliable control*, IEEE Trans. Automat. Control, 44 (1999), pp. 1449–1454.

[6] J. C. Geromel, J. Bernussou, and P. L. D. Peres, *Decentralized control through parameter space optimization*, Automatica J. IFAC, 30 (1994), pp. 1565–1578.

[7] J. C. Geromel, P. L. D. Peres, and J. Bernussou, *On a convex parameter space method for linear control design of uncertain systems*, SIAM J. Control Optim., 29 (1991), pp. 381–402.

[8] J. C. Geromel, P. L. D. Peres, and S. R. de Souza, *A convex approach to the mixed H_2/H_∞ control problem for discrete time uncertain systems*, SIAM J. Control Optim., 33 (1995), pp. 1816–1833.

[9] B. N. Jain, *Guaranteed error estimation in uncertain systems*, IEEE Trans. Automat. Control, 20 (1975), pp. 230–232.

[10] P. P. Khargonekar, M. A. Rotea, and E. Bayens, *Mixed H_2/H_∞ filtering*, Internat. J. Robust Nonlinear Control, 6 (1996), pp. 313–330.

[11] K. M. Nagpal and P. P. Khargonekar, *Filtering and smoothing in an H_∞ setting*, IEEE Trans. Automat. Control, 36 (1991), pp. 152–166.

[12] J. O'Reilly, *Observers for Linear Systems*, Academic Press, New York, 1983.

[13] I. R. Petersen and D. C. MacFarlane, *Optimal guaranteed cost filtering for uncertain discrete-time systems*, IEEE Trans. Automat. Control, 6 (1996), pp. 267–280.

[14] C. Scherer, *Mixed H_2/H_∞ control*, in Trends in Control: A European Perspective, A. Isidori, ed., Springer-Verlag, Berlin, 1995, pp. 173–216.

[15] C. Scherer, P. Gahinet, and M. Chilali, *Multiobjective output-feedback control via LMI optimization*, IEEE Trans. Automat. Control, 42 (1997), pp. 896–911.

[16] U. Shaked and Y. Theodor, *H_∞ optimal estimation: A tutorial*, in Proceedings of the IEEE Conference on Decision and Control, Tucson, AZ, 1992.

[17] D. D. Siljak, *Large Scale Dynamic Systems: Stability and Structure*, North-Holland, Amsterdam, 1979.

[18] L. Xie, C. E. de Souza, and M. Fu, *H_∞ estimation for discrete-time linear uncertain systems*, Internat. J. Robust Nonlinear Control, 1 (1991), pp. 11–23.

[19] L. Xie and Y. C. Soh, *Robust Kalman filtering for uncertain systems*, Systems Control Lett., 22 (1994), pp. 123–129.

[20] K. Zhou, J. C. Doyle, and K. Glover, *Robust and Optimal Control*, Prentice-Hall, Upper Saddle River, NJ, 1996.

SIAM J. CONTROL OPTIM.
Vol. 38, No. 5, pp. 1369–1391

SECOND ORDER SUFFICIENT OPTIMALITY CONDITIONS FOR SOME STATE-CONSTRAINED CONTROL PROBLEMS OF SEMILINEAR ELLIPTIC EQUATIONS*

EDUARDO CASAS†, FREDI TRÖLTZSCH‡, AND ANDREAS UNGER‡

Abstract. This paper deals with a class of optimal control problems governed by elliptic equations with nonlinear boundary condition. The case of boundary control is studied. Pointwise constraints on the control and certain equality and set-constraints on the state are considered. Second order sufficient conditions for local optimality of controls are established.

Key words. boundary control, semilinear elliptic equations, sufficient optimality conditions, state constraints

AMS subject classifications. 49K20, 35J25

PII. S0363012997324910

1. Introduction.

In contrast to the optimal control of linear systems with a convex objective, where first order necessary optimality conditions are already sufficient for optimality, higher order conditions such as second order sufficient optimality conditions (SSC) should be employed to verify optimality for nonlinear systems. Second order sufficient optimality conditions have also proved to be useful for showing important properties of optimal control problems such as local uniqueness of optimal controls and their stability with respect to certain perturbations. Moreover, they may serve as an assumption to guarantee the convergence of numerical methods in optimal control. In this respect, we refer to the general expositions by Maurer and Zowe [15] and Maurer [14] for different aspects of second order sufficient optimality conditions. The approximation of programming problems in Banach spaces is discussed in Alt [2]. Moreover, Alt [3], [4] has established a general convergence analysis for Lagrange–Newton methods in Banach spaces.

Meanwhile, an extensive number of publications have been devoted to different aspects of second order sufficient optimality conditions for control problems governed by ordinary differential equations. The well-known *two-norm discrepancy* has in particular received a good deal of attention. We refer, for instance, to Ioffe [13] and Maurer [14].

First investigations of second order sufficient optimality conditions for control problems governed by partial differential equations have been published by Goldberg and Tröltzsch [11], [12] for the boundary control of parabolic equations with nonlinear boundary conditions. In [9], Casas, Tröltzsch, and Unger have extended these ideas to elliptic boundary control problems with pointwise constraints on the control. Moreover, they tightened the gap between second order necessary and sufficient

*Received by the editors July 21, 1997; accepted for publication (in revised form) December 4, 1998; published electronically May 11, 2000. This research was partially supported by European Union, under HCM project ERBCHRXCT940471, and by Deutsche Forschungsgemeinschaft, under project Tr 302/1-2. The research of the first author was also supported by Dirección General de Enseñanza Superior e Investigación Científica (Spain).

http://www.siam.org/journals/sicon/38-5/32491.html

†Departamento de Matemática Aplicada y Ciencias de la Computación, E.T.S.I. Industriales y de Telecomunicación, Universidad de Cantabria, 39071 Santander, Spain (casas@macc.unican.es).

‡Fakultät für Mathematik, Technische Universität Chemnitz-Zwickau, D-09107 Chemnitz, Germany (f.troeltzsch@mathematik.tu-chemnitz.de, aug@shc.tcc-chemnitz.de).

optimality conditions. This was done by the consideration of sets of strongly active constraints according to Dontchev et al. [10]. This technique is also related to first order sufficient optimality conditions introduced by Maurer and Zowe [15]. It should be mentioned that as many as four norms have to be used in this case (L^∞-norm for differentiation, L^2-norm to formulate second order sufficient optimality conditions, L^1-norm for the first order sufficient optimality condition, and certain L^p-norms to obtain optimal regularity results).

Bonnans [5] has shown that a very weak form of second order sufficient conditions can be used to verify local optimality for a particular class of semilinear elliptic control problems with constraints on the control: If the second order derivative of the Lagrange function is a Legendre form, then it suffices to have its positivity in all critical directions.

In our paper, the results of [9] will be extended to additional constraints on the state. In this way, we are continuing the investigations by Casas and Tröltzsch [8] on second order *necessary conditions*. We also rely on general ideas of Maurer and Zowe [15], combining their approach with a detailed splitting technique.

At the beginning, we aimed to establish second order sufficient optimality conditions for boundary control problems governed by semilinear elliptic equations in domains of arbitrary dimension with general pointwise constraints on the control and the state. However, we soon recognized that *pointwise* state-constraints lead to essential and somewhat surprising difficulties. To establish second order sufficient optimality conditions for problems with pointwise state-constraints given on the whole domain, we had to restrict ourselves to two-dimensional domains with controls appearing linearly in the boundary condition. These obstacles might indicate some limits for the "traditional" type of second order sufficient optimality conditions for control problems governed by PDEs.

If pointwise state-constraints are imposed on compact subsets of the domain, while the other quantities are sufficiently smooth, then arbitrary dimensions can be treated without restrictions on the nonlinearities. In this case the adjoint state belongs to $L^\infty(\Gamma)$. Moreover, we are able to avoid the assumption of linearity of the boundary condition with respect to the control by introducing some extended form of second order optimality conditions.

2. The optimal control problem. We consider the problem: *Minimize* the functional

$$F_0(y,u) = \int_\Omega f(x,y(x))\,dx + \int_\Gamma g(x,y(x),u(x))\,dS(x) \tag{2.1}$$

subject to the *equation of state*

$$\begin{cases} -\Delta y(x) + y(x) = 0 & \text{in } \Omega, \\ \partial_\nu y(x) = b(x,y(x),u(x)) & \text{on } \Gamma, \end{cases} \tag{2.2}$$

to the constraints on the *state* y

$$F_i(y) = 0, \quad i = 1,\dots,m, \tag{2.3}$$

$$E(y) \in K, \tag{2.4}$$

and to the constraints on the *control* u

$$u_a(x) \le u(x) \le u_b(x) \quad \text{almost everywhere (a.e.) on } \Gamma. \tag{2.5}$$

In this setting, $\Omega \subset \mathbb{R}^n$ is a bounded domain with a Lipschitz boundary Γ according to the definition by Nečas [17]. Moreover, sufficiently smooth functions $f : \Omega \times \mathbb{R} \to \mathbb{R}$ and $g, b : \Gamma \times \mathbb{R}^2 \to \mathbb{R}$ are given. The symbol ∂_ν is used for the derivative in the direction of the unit outward normal ν on Γ. The functionals $F_i : C(\overline{\Omega}) \to \mathbb{R}$, $i = 1, \dots, m$, are supposed to be twice continuously Fréchet differentiable, that is, to be of class C^2. By E we denote a mapping of class C^2 from $C(\overline{\Omega})$ into a real Banach space Z. $K \subset Z$ is a nonempty convex closed set, and u_a, $u_b : \Gamma \to \mathbb{R}$ are functions of $L^\infty(\Gamma)$ satisfying $u_a(x) \le u_b(x)$ on Γ.

The *control* u is looked for in the control space $\mathcal{U} = L^\infty(\Gamma)$, while the *state* y is defined as *a weak solution* of (2.2) in the state space $C(\overline{\Omega}) \cap H^1(\Omega) = Y$, that is,

$$\int_\Omega (\nabla y \nabla v + yv)\, dx = \int_\Gamma b(\cdot, y, u) v \, dS \quad \forall v \in H^1(\Omega). \tag{2.6}$$

We endow Y with the norm $\|y\|_Y = \|y\|_{C(\overline{\Omega})} + \|y\|_{H^1(\Omega)}$. The following assumptions are imposed on the given quantities.

(A1) For each fixed $x \in \Omega$ or Γ, respectively, the functions $f = f(x, y)$, $g = g(x, y, u)$, and $b = b(x, y, u)$ are of class C^2 with respect to (y, u). For fixed (y, u), they are Lebesgue measurable with respect to $x \in \Omega$ or $x \in \Gamma$, respectively.

Throughout the paper, partial derivatives are indicated by associated subscripts. For instance, b_{yu} stands for $\partial^2 b / \partial y \partial u$. By $b'(x, y, u)$ and $b''(x, y, u)$ we denote the gradient and the Hessian matrix of b with respect to (y, u):

$$b'(x, y, u) = \begin{pmatrix} b_y(x, y, u) \\ b_u(x, y, u) \end{pmatrix}, \quad b''(x, y, u) = \begin{pmatrix} b_{yy}(x, y, u) & b_{yu}(x, y, u) \\ b_{uy}(x, y, u) & b_{uu}(x, y, u) \end{pmatrix};$$

$|b'|$ and $|b''|$ are defined by adding the absolute values of all entries.

In the next assumption, fixed parameters $p > n - 1$ and s, r are used, which depend on n. For the possible (minimal) choice of s and r we refer to the discussion of regularity in (3.13). Roughly speaking, we have $y|_\Gamma \in L^s(\Gamma)$ and $y \in L^r(\Omega)$ in the linearized system (2.2) if $u \in L^2(\Gamma)$. As usual, s' and r' denote conjugate numbers. For instance, s' is defined by $1/s' + 1/s = 1$.

(A2) For all $M > 0$ there are constants $C_M > 0$, functions $\Psi_f^M \in L^{(r/2)'}(\Omega)$, $\Psi_g^{M,1} \in L^{(s/2)'}(\Gamma)$, $\Psi_g^{M,2} \in L^{2(s/2)'}(\Gamma)$, $\Psi_g^{M,3} \in L^\infty(\Gamma)$, and a continuous, monotone increasing function $\eta \in C(\mathbb{R}^+ \cup \{0\})$ with $\eta(0) = 0$ such that

(i)

$$b_y(x, y, u) \le 0 \qquad \text{a.e. } x \in \Gamma, \ \forall (y, u) \in \mathbb{R}^2, \tag{2.7}$$

$b(\cdot, 0, 0) \in L^p(\Gamma)$, for a $p > n - 1$,
$|b'(x, y, u)| + |b''(x, y, u)| \le C_M$,
$|b''(x, y_1, u_1) - b''(x, y_2, u_2)| \le C_M \, \eta(|y_1 - y_2| + |u_1 - u_2|)$
for almost all $x \in \Gamma$ and all $|y|, |u|, |y_i|, |u_i| \le M$, $i = 1, 2$;

(ii) $f(\cdot, 0) \in L^1(\Omega)$, $f_y(\cdot, 0) \in L^{r'}(\Omega)$, $f_{yy}(\cdot, 0) \in L^{(r/2)'}(\Omega)$
$|f_{yy}(x, y_1) - f_{yy}(x, y_2)| \le \Psi_f^M(x) \, \eta(|y_1 - y_2|)$
$\forall \, x \in \Omega$, $|y_i| \le M$, $i = 1, 2$;

(iii) $g(\cdot,0,0) \in L^1(\Gamma)$, $g_y(\cdot,0,0) \in L^{s'}(\Gamma)$, $g_u(\cdot,0,0) \in L^2(\Gamma)$,
$g_{yy}(\cdot,0,0) \in L^{(s/2)'}(\Gamma)$, $g_{yu}(\cdot,0,0) \in L^{2(s/2)'}(\Gamma)$, $g_{uu}(\cdot,0,0) \in L^\infty(\Gamma)$
(here, $\cdot$ stands for x)

$$|g_{yy}(x,y_1,u_1) - g_{yy}(x,y_2,u_2)| \le \Psi_g^{M,1}(x)\eta(|y_1-y_2|+|u_1-u_2|),$$
$$|g_{yu}(x,y_1,u_1) - g_{yu}(x,y_2,u_2)| \le \Psi_g^{M,2}(x)\eta(|y_1-y_2|+|u_1-u_2|),$$
$$|g_{uu}(x,y_1,u_1) - g_{uu}(x,y_2,u_2)| \le \Psi_g^{M,3}(x)\eta(|y_1-y_2|+|u_1-u_2|),$$
for almost all $x \in \Gamma$ and all $|y_i| \le M$, $|u_i| \le M$.

REMARK 2.1. *Notice that the estimates in* (i)–(iii) *imply boundedness and Lipschitz properties of* b, f, g, b', f', g' *in several L-spaces. We omit them, because they follow from the mean value theorem.*

(A3) (i) Let us define the norm

$$\|y\|_2 = \|y\|_{C(A)} + \|y\|_{L^r(\Omega)} + \|y\|_{L^s(\Gamma)}$$

for $y \in C(\overline{\Omega})$, where $A \subset \overline{\Omega}$ is a certain measurable compact subset. Here A stands for a set, where we know $y \in C(A)$ for Neumann boundary data given in $L^2(\Gamma)$. In the case $n = 2$ we may take $A = \overline{\Omega}$, while $A \subset \Omega$ is needed for $n > 2$. For $A = \emptyset$ we put $\|y\|_{C(A)} = 0$.
We assume at a fixed reference state $\overline{y} \in C(\overline{\Omega})$ that

$$|F_i'(\overline{y})y| \le C_F\|y\|_2 \quad \forall y \in C(\overline{\Omega}),$$
$$|F_i''(\overline{y})[y_1,y_2]| \le C_F\|y_1\|_2\|y_2\|_2 \quad \forall y_1, y_2 \in C(\overline{\Omega})$$

holds with some $C_F > 0$. Moreover, we require with a $C_M > 0$

$$|F_i'(y_1)y - F_i'(y_2)y| \le C_M\|y_1-y_2\|_2\|y\|_2,$$
$$|(F_i''(y_1) - F_i''(y_2))[y,v]| \le C_M\,\eta(\|y_1-y_2\|_{C(\overline{\Omega})})\|y\|_2\|v\|_2$$

$\forall\, y_j$ with $\|y_j\|_{C(\overline{\Omega})} \le M$, $j = 1,2$, $\forall\, y$, v from $C(\overline{\Omega})$, and $\forall\, i = 1,\dots,m$.

(ii) Analogous assumptions are imposed on $E : C(\overline{\Omega}) \to Z$, where $\|\cdot\|_Z$ is to be substituted for $|\cdot|$. For instance,

$$\|E'(\overline{y})y\|_Z \le C_E\|y\|_2 \quad \forall y \in C(\overline{\Omega})$$

is supposed.

We shall explain the main constructions of our paper by the following canonical example (P) that fits in the general setting.

Example (P). Minimize

$$\frac{1}{2}\int_\Omega (y-y_d)^2dx + \frac{\alpha}{2}\int_\Gamma u^2dS$$

subject to

$$\begin{aligned} -\triangle y + y &= 0 && \text{in } \Omega,\\ \partial_\nu y &= u - y^3 && \text{on } \Gamma, \end{aligned}$$

and

$$|u| \le 1, \qquad y(0) \le y_0$$

in the open unit ball $\Omega \subset \mathbb{R}^3$ around zero, where $\alpha > 0$, $y_0 \in \mathbb{R}$, and $y_d \in L^\infty(\Omega)$ are given. Here, we have $Z = \mathbb{R}$, $K = \mathbb{R}^-$, $A = \{0\}$, $E(y) = y(0) - y_0$, and we need $y \in C(\Omega)$ to make E well defined.

3. The state equation and first order necessary optimality conditions. It can be shown that the equation (2.2) admits for each $u \in \mathcal{U}^{ad}$ a unique weak solution $y = y(u) \in Y$, where $\mathcal{U}^{ad} = \{u \in L^\infty(\Gamma) \,|\, u_a(x) \le u(x) \le u_b(x) \text{ a.e. on } \Gamma\}$. Moreover, there is a constant M such that

$$\|y(u)\|_Y \le M \quad \forall u \in \mathcal{U}^{ad}. \tag{3.1}$$

In particular, it holds that $\|y\|_{C(\overline{\Omega})} \le M$. Casas and Tröltzsch [8] have proved that the mapping $u \mapsto y(u)$ from $L^\infty(\Gamma)$ into Y is of class C^2. Furthermore, the Lipschitz property

$$\|y(u_1) - y(u_2)\|_2 \le C_2 \|u_1 - u_2\|_{L^2(\Gamma)}$$

holds for all u_1, $u_2 \in \mathcal{U}^{ad}$, where C_2 is a positive constant and $\|\cdot\|_2$ is defined in (A3). For fixed $u \in \mathcal{U}^{ad}$ we have $b(\cdot, y, u) \in L^p(\Gamma)$, hence the weak solution $y \in Y$ of (2.2) belongs to the space

$$Y_{q,p} = \{y \in H^1(\Omega) \,|\, -\Delta y + y \in L^q(\Omega), \partial_\nu y \in L^p(\Gamma)\},$$

which is known to be continuously embedded into $Y = C(\overline{\Omega}) \cap H^1(\Omega)$ for each $q > n/2$ and each $p > n - 1$.

In all of what follows we assume that a *reference pair* $(\overline{y}, \overline{u}) \in Y \times \mathcal{U}^{ad}$ is given, satisfying, together with an associated *adjoint state* $\overline{\varphi} \in W^{1,\sigma}(\Omega)$ $\forall \sigma < n/(n-1)$, and with *Lagrange multipliers*

$$\overline{\lambda} = (\overline{\lambda}_1, \dots, \overline{\lambda}_m)^T \in \mathbb{R}^m, \qquad \overline{z}^* \in Z^*,$$

the associated standard *first order necessary optimality conditions*. We will just assume them. They can be proved following Casas [7], Bonnans and Casas [6], or Zowe and Kurcyusz [23]. The first order optimality system to be satisfied by $(\overline{y}, \overline{u})$ consists of the state equations (2.2), the constraint $\overline{u} \in \mathcal{U}^{ad}$, the *adjoint equations*

$$-\Delta\overline{\varphi} + \overline{\varphi} = f_y(\cdot, \overline{y}) + \sum_{i=1}^m \overline{\lambda}_i F_i'(\overline{y})|_\Omega + (E'\overline{y})^* \overline{z}^*|_\Omega \qquad \text{in } \Omega, \tag{3.2}$$

$$\partial_\nu \overline{\varphi} = b_y(\cdot, \overline{y}, \overline{u})\overline{\varphi} + g_y(\cdot, \overline{y}, \overline{u}) + \sum_{i=1}^m \overline{\lambda}_i F_i'(\overline{y})|_\Gamma + (E'\overline{y})^* \overline{z}^*|_\Gamma \quad \text{on } \Gamma \tag{3.3}$$

for the adjoint state $\overline{\varphi}$, the *complementary slackness condition*

$$\langle \overline{z}^*, \kappa - E(\overline{y}) \rangle \le 0 \quad \forall \kappa \in K, \tag{3.4}$$

and the *variational inequality*

$$\int_\Gamma (g_u(x, \overline{y}(x), \overline{u}(x)) + \overline{\varphi}(x) b_u(x, \overline{y}(x), \overline{u}(x)))(u(x) - \overline{u}(x))\, dS(x) \ge 0 \tag{3.5}$$

$\forall\, u \in \mathcal{U}^{ad}$. We have $F_i'(\overline{y}) \in C(\overline{\Omega})^*$, $i = 1, \dots, m$, and $E'(\overline{y})^* \overline{z}^* \in C(\overline{\Omega})^*$; hence these quantities can be identified with real Borel measures on $\overline{\Omega}$. Let a nonnegative function $\beta \in L^\infty(\Gamma)$ and real Borel measures μ_Ω and μ_Γ concentrated on Ω and Γ, respectively, be given. Then the problem

$$\begin{cases} -\Delta\varphi + \varphi = \mu_\Omega & \text{in } \Omega, \\ \partial_\nu \varphi + \beta\varphi = \mu_\Gamma & \text{on } \Gamma \end{cases} \tag{3.6}$$

admits a unique solution $\varphi \in W^{1,\sigma}(\Omega)\ \forall\ \sigma < n/(n-1)$ (see Casas [7]; the reader is also referred to Stampacchia [20] for the Dirichlet case). In view of this, we may write

$$\overline{\varphi} = \varphi_0 + \sum_{i=1}^{m} \lambda_i \varphi_i + \varphi_E,$$

where φ_0, φ_i, and φ_E solve (3.6) for $\mu_\Omega = f_y$, $F_i'(\overline{y})|_\Omega$, $E'(\overline{y})^*\overline{z}^*|_\Omega$, and $\mu_\Gamma = g_y$, $F_i'(\overline{y})|_\Gamma$, $E'(\overline{y})^*\overline{z}^*|_\Gamma$, respectively. We have at least φ_0, φ_i, and φ_E in $W^{1,\sigma}(\Omega)$. Moreover, $\overline{\varphi}$ satisfies the formula of integration by parts

$$\int_\Omega (-\Delta y + y)\varphi\, dx + \int_\Gamma (\partial_\nu y + \beta y)\varphi\, dS(x) = \int_\Omega y\, d\mu_\Omega + \int_\Gamma y\, d\mu_\Gamma \tag{3.7}$$

$\forall\ y \in Y_{q,p}$, where $q > n/2$, $p > n-1$. It is easy to verify that the optimality conditions can be expressed by the *Lagrange function*

$$\begin{aligned}\mathcal{L}(y,u,\varphi,\lambda,z^*) = F_0(y,u) - \int_\Omega (-\Delta y + y)\,\varphi\, dx - \int_\Gamma (\partial_\nu y - b(\cdot,y,u))\varphi\, dS \\ + \sum_{j=1}^{m} \lambda_j F_j(y) + \langle z^*, E(y)\rangle,\end{aligned} \tag{3.8}$$

$\mathcal{L} : Y_{q,p} \times \mathcal{U} \times W^{1,\sigma}(\Omega) \times \mathbb{R}^m \times Z^* \to \mathbb{R}$. The regularity of y and φ fits together, as $\varphi \in W^{1,\sigma}(\Omega)\ \forall\ \sigma < n/(n-1)$ ensures $\varphi \in L^s(\Omega)\ \forall\ s < n/(n-2)$ (Nečas [17, Thm. 3.4, p. 69]), and $\varphi|_\Gamma \in L^r(\Gamma)$ holds $\forall\ r < 1 + 1/(n-2)$ [17, Thm. 4.2, p. 84]). Therefore, this definition makes sense. In (3.8), $\langle\cdot,\cdot\rangle$ denotes the duality pairing between Z and its dual space Z^*. The Lagrange function $\mathcal{L}$ is of class C^2 with respect to (y,u) for fixed φ, λ, and z^*.

Thanks to (3.7), the optimality system can be rewritten in terms of $\mathcal{L}$. Then it is expressed by (2.6), the constraints on the state (2.3), (2.4), the constraints on the control $u \in \mathcal{U}^{ad}$, and

$$\mathcal{L}_y(\overline{y},\overline{u},\overline{\varphi},\overline{\lambda},\overline{z}^*)y = 0 \quad \forall y \in Y, \tag{3.9}$$

$$\mathcal{L}_u(\overline{y},\overline{u},\overline{\varphi},\overline{\lambda},\overline{z}^*)(u-\overline{u}) \geq 0 \quad \forall u \in \mathcal{U}^{ad}, \tag{3.10}$$

$$\langle \overline{z}^*, \kappa - E(\overline{y})\rangle \leq 0 \quad \forall \kappa \in K. \tag{3.11}$$

This form is more convenient for our later evaluations.

Example. In (P), adjoint equation and variational inequality are given by

$$\begin{aligned} -\triangle\overline{\varphi} + \overline{\varphi} = \overline{y} - y_d + \overline{z}^* \circ \delta(0), \qquad & \partial_\nu\overline{\varphi} + 3\,\overline{y}^2\overline{\varphi} = 0, \\ \int_\Gamma (\alpha\overline{u} + \overline{\varphi})(u-\overline{u})dS \geq 0 \qquad & \forall |u| \leq 1, \end{aligned}$$

where $\delta(0)$ is the Dirac measure.

To shorten our notation, derivatives taken at $(\overline{y},\overline{u},\overline{\varphi},\overline{\lambda},\overline{z}^*)$ will be indicated by a bar. For instance, $\overline{\mathcal{L}}_y y$, $\overline{\mathcal{L}}_u(u-\overline{u})$ stand for the derivatives in (3.9) and (3.10), respectively. $\overline{\mathcal{L}}_{yy}[y_1,y_2]$ denotes the second order derivative of $\mathcal{L}$ in the directions y_1, y_2 taken at $(\overline{y},\overline{u},\overline{\varphi},\overline{\lambda},\overline{z}^*)$. Moreover, $\overline{\mathcal{L}}_{ww}[w_1,w_2]$ is the second order derivative of $\mathcal{L}$ in the directions $w_1 = (y_1,u_1)$, $w_2 = (y_2,u_2)$. If $w_1 = w_2 = w$, then we write for short $\overline{\mathcal{L}}_{ww}[w,w] = \overline{\mathcal{L}}_{ww}[w]^2$.

Next we provide some useful results on linearized versions of the state equation. Regard first the linear system

$$\begin{cases} -\Delta y + y = f & \text{in } \Omega, \\ \partial_\nu y + \beta y = g & \text{on } \Gamma, \end{cases} \tag{3.12}$$

where $\beta \in L^\infty(\Gamma)$ is nonnegative. For each pair $(f,g) \in L^1(\Omega) \times L^1(\Gamma)$, this system admits a unique solution $y \in W^{1,\sigma}(\Omega)$, where $\sigma < n/(n-1)$; see Casas [7]. (Notice that a function of L^1 can be considered as a Borel measure.) On the other hand, the solution y of (3.12) belongs to $H^1(\Omega) \cap C(\overline{\Omega})$ if $(f,g) \in L^q(\Omega) \times L^p(\Gamma)$. This regularity result is well known for domains with C^1-boundary. Moreover, it remains true for domains with Lipschitz boundary in the sense of Nečas [17] (see Stampacchia [19] and Murthy and Stampacchia [16]). On account of this, the mapping $D : (f,g) \mapsto (y, y|_\Gamma)$ is continuous from $L^1(\Omega) \times L^1(\Gamma)$ into $L^s(\Omega) \times L^t(\Gamma)$ for $s < n/(n-2)$ and $t < (n-1)/(n-2)$. D is continuous also from $L^q(\Omega) \times L^p(\Gamma)$ into $L^\infty(\Omega) \times L^\infty(\Gamma)$. We obtain these spaces by embedding results for $W^{1,\sigma}(\Omega)$ [1], [17], [20]. In both cases, this mapping is linear and continuous. Interpolation theory applies to show the following results for D considered as a mapping defined on $L^2(\Omega) \times L^2(\Gamma)$:

$$y \in \begin{cases} C(\overline{\Omega}), & n = 2, \\ L^r(\Omega)\ \forall r < \infty, & n = 3, \\ L^r(\Omega)\ \forall r < \dfrac{2n}{n-3}, & n \geq 4, \end{cases} \quad y|_\Gamma \in \begin{cases} C(\Gamma), & n = 2, \\ L^s(\Gamma)\ \forall s < \infty, & n = 3, \\ L^s(\Gamma)\ \forall s < \dfrac{2(n-1)}{n-3}, & n \geq 4. \end{cases} \tag{3.13}$$

4. Regularity condition and linearization theorem. Let us recall that we consider a fixed reference pair $(\overline{y}, \overline{u})$ satisfying, together with $(\overline{\varphi}, \overline{\lambda}, \overline{z}^*)$, the first order necessary conditions (3.9)–(3.11).

The *linearized cone* of $\mathcal{U}^{ad}$ at $\overline{u}$ is the set $\mathcal{C}(\overline{u}) = \{v \in L^\infty(\Gamma) \,|\, v = \varrho(u - \overline{u}),\ \varrho \geq 0,\ u \in \mathcal{U}^{ad}\}$. Let $F = F(y)$ denote the mapping $y \mapsto (F_1(y), \dots, F_m(y))^T$ from Y to $\mathbb{R}^m$. For convenience, we introduce the set of all feasible pairs

$$\mathcal{M} = \{w = (y,u) \in Y \times \mathcal{U}^{ad} \,|\, y = G(u) \text{ and } y \text{ satisfies the state-constraints}\}$$

(notice that G is the nonlinear control-state-mapping). Following Maurer and Zowe [15], the *linearized cone* $L(\mathcal{M}, \overline{w})$ at $\overline{w} = (\overline{y}, \overline{u})$ is defined by

$$L(\mathcal{M}, \overline{w}) = \{w \,|\, w = (y,u),\ u \in \mathcal{C}(\overline{u}) \text{ and } (y,u) \text{ satisfies (4.1)–(4.3)}\},$$

where

$$\begin{cases} -\Delta y + y &= 0 & \text{in } \Omega, \\ \partial_\nu y &= b_y(\cdot, \overline{y}, \overline{u})y + b_u(\cdot, \overline{y}, \overline{u})u & \text{on } \Gamma, \end{cases} \tag{4.1}$$

$$F'(\overline{y})y = 0, \tag{4.2}$$

$$E'(\overline{y})y \in K(E(\overline{y})). \tag{4.3}$$

Here, $K(E(\overline{y})) = \{z \in Z \mid z = \varrho(\kappa - E(\overline{y})),\ \varrho \geq 0,\ \kappa \in K\}$ is the conical hull of $K - E(\overline{y})$.

REMARK 4.1. *The choice* $Z = \mathbb{R}^k$, $E(y) = (E_1(y), \dots, E_k(y))^T$, $K = (\mathbb{R}^k)^-$ *for* $E : Y \to Z$ *is of particular interest. Then* (4.3) *reduces to*

$$E_i'(\overline{y})y \leq 0$$

for all active $i \in \{1, \dots, k\}$, *that is for all* i, *where* $E_i(\overline{y}) = 0$ *holds.*

Example. The linearized cone for (P) is the set of the following pairs (y,u): They satisfy $u \in \mathcal{C}(\overline{u})$ and

$$-\triangle y + y = 0, \qquad \partial_\nu y + 3\overline{y}^2 y = u, \tag{4.4}$$

$$y(0) \le 0, \tag{4.5}$$

if $\overline{y}(0) = y_0$ (active state constraint). If the state constraint is not active, then (4.5) disappears.

The following *regularity assumption* (R) is basic for our further analysis: To formulate (R) we combine the two state constraints to one general constraint. We therefore take $\boldsymbol{Z} = \mathbb{R}^m \times Z$, $\boldsymbol{K} = \{0\} \times K$, define $T : Y \to \boldsymbol{Z}$ by $T(y) = (F(y), E(y))$, and put $\boldsymbol{K}(T(\overline{y})) = \{0\} \times K(E(\overline{y}))$. The regularity condition was introduced by Zowe and Kurcyusz [23] and requires

$$T'(\overline{y})G'(\overline{u})\mathcal{C}(\overline{u}) - \boldsymbol{K}(T(\overline{y})) = \boldsymbol{Z}. \tag{R}$$

This condition is sufficient for the existence of a (nondegenerate) Lagrange multiplier associated to the state-constraint $E(y) \in K$; see [23]. We should emphasize that (R) does not rely on the condition $\operatorname{int} K \neq \emptyset$. In Appendix 7.1 we shall present some sufficient conditions for (R) which, however, require $\operatorname{int} K \neq \emptyset$. (R) is discussed for the canonical example (P) there. For $Z = \mathbb{R}^k$, $K = (\mathbb{R}^k)^-$, the condition (R) is equivalent to the well-known *Mangasarian–Fromowitz condition.*

THEOREM 4.2. *Suppose that (R) is satisfied. Then for all pairs $(\hat{y},\hat{u}) \in \mathcal{M}$ there is a pair $(y,u) \in L(\mathcal{M},\overline{w})$ such that the difference $r = (r^y, r^u) = (\hat{y},\hat{u}) - (\overline{y},\overline{u}) - (y,u)$ can be estimated by*

$$\|r\|_{Y\times L^\infty(\Gamma)} \le C_{L,p}\|\hat{u}-\overline{u}\|_{L^\infty(\Gamma)}\|\hat{u}-\overline{u}\|_{L^p(\Gamma)} \quad \forall p > n-1, \tag{4.6}$$

$$\|r\| \le C_{L,2}\|\hat{u}-\overline{u}\|_{L^\infty(\Gamma)}\|\hat{u}-\overline{u}\|_{L^2(\Gamma)}, \tag{4.7}$$

where $\|r\| = \|r^y\|_2 + \|r^u\|_{L^2(\Gamma)}$. In the particular case $b(x,y,u) = b_1(x,y) + b_2(x)u$ we have

$$\|r\|_{Y\times L^\infty(\Gamma)} \le C_{L,p}\|\hat{u}-\overline{u}\|^2_{L^p(\Gamma)} \quad \forall p > n-1. \tag{4.8}$$

This theorem is proved in Appendix 7.2. Let us conclude this section by considering some useful estimates for $\mathcal{L}''$ and for certain remainder terms. First, we evaluate

$$\overline{\mathcal{L}}''[(y_1,u_1),(y_2,u_2)] = \mathcal{L}''(\overline{y},\overline{u},\overline{\varphi},\overline{\lambda},\overline{z}^*)[(y_1,u_1),(y_2,u_2)],$$

where $\mathcal{L}''$ denotes the second order derivative of $\mathcal{L}$ with respect to (y,u). We have

$$\begin{aligned}\overline{\mathcal{L}}''[(y_1,u_1),(y_2,u_2)] &= \int_\Omega f_{yy}(\cdot,\overline{y})y_1y_2\,dx + \int_\Gamma (y_1,u_1)g''(\cdot,\overline{y},\overline{u})(y_2,u_2)^T\,dS \\ &\quad + \int_\Gamma \overline{\varphi}\cdot(y_1,u_1)b''(\cdot,\overline{y},\overline{u})(y_2,u_2)^T\,dS \\ &\quad + \sum_{i=1}^m \overline{\lambda}_i F_i''(\overline{y})[y_1,y_2] + \langle \overline{z}^*, E''(\overline{y})[y_1,y_2]\rangle.\end{aligned} \tag{4.9}$$

Example. In the case of **(P)**, $\overline{\mathcal{L}}''$ admits the form

$$\overline{\mathcal{L}}''[(y_1,u_1),(y_2,u_2)] = \int_\Omega y_1y_2dx + \int_\Gamma (6\overline{\varphi}\,\overline{y}\,y_1y_2 + \alpha u_1u_2)dS.$$

The term connected with $\overline{\varphi}$ causes trouble, more precisely,

$$I = \int_\Gamma \overline{\varphi}\,(b_{yy}(\cdot,\overline{y},\overline{u})y_1y_2 + b_{yu}(\cdot,\overline{y},\overline{u})(y_1u_2+y_2u_1) + b_{uu}(\cdot,\overline{y},\overline{u})u_1u_2)\,dS. \tag{4.10}$$

An estimate of I is needed with respect to the norm $\|y\|_2 + \|u\|_{L^2(\Gamma)}$ (cf. (4.19)). We therefore have to require at least $\overline{\varphi} \in L^2(\Gamma)$ in the second item and $\overline{\varphi} \in L^\infty(\Gamma)$ in the third one. On the other hand, only $\overline{\varphi} \in L^r(\Gamma)$ follows from $\overline{\varphi} \in W^{1,\sigma}(\Omega)$ for $r < (n-1)/(n-2)$; see Nečas [17, p. 84]. For $n = 2$ we obtain $\overline{\varphi} \in L^r(\Gamma)$ for all $r < \infty$, while $n = 3$ yields the regularity $\overline{\varphi} \in L^r(\Gamma)$ for all $r < 2$. On account of this, the following additional assumption is crucial for our analysis.

(A4) Let one of the following statements be true:
- (i) $\overline{\varphi} \in L^\infty(\Gamma)$.
- (ii) $b_{uu}(x,y,u) = 0$ on $\Gamma \times \mathbb{R}^2$ and, if $n \geq 3$, then $\overline{\varphi} \in L^r(\Gamma)$ for some $r > n-1$.
- (iii) $b_{uu}(x,y,u) = b_{yu}(x,y,u) = 0$ on $\Gamma \times \mathbb{R}^2$ and, if $n \geq 4$, then $\overline{\varphi} \in L^r(\Gamma)$ for some $r > (n-1)/2$.
- (iv) $b''(\cdot,y,u) = 0$.

We briefly comment on the consequences of these assumptions: (i) is true if $\overline{f}_y \in L^q(\Omega)$, $\overline{g}_y \in L^p(\Gamma)$ and if the restrictions of F_i', $i = 1,\dots,m$, and $E'(\overline{y})^*\overline{z}^*$ to Ω and Γ, respectively, belong to $L^q(\Omega)$, $L^p(\Gamma)$, as well. Moreover, (i) holds for functionals F_i', $i = 1,\dots,m$, and $E'(\overline{y})^*\overline{z}^*$ of $C(\overline{\Omega})^*$, where the associated real Borel measures are concentrated on the set $A \subset \Omega$.

In addition to some assumptions on the regularity of $\overline{\varphi}$ for $n \geq 3,4$, (ii) requires linearity of b with respect to u, that is $b(x,y,u) = b_0(x,y) + b_1(x,y)u$, (iii) means that $b(x,y,u) = b_1(x,y) + b_2(x)u$, while (iv) is true only for an affine-linear boundary condition (but yet also true for a nonlinear functional F_0).

(A4) is obviously satisfied in the example (P).

As a consequence of (A3) and (A4), *pointwise state-constraints* on the whole set $\overline{\Omega}$ can only be handled by the standard part of our theory if u appears linearly in the boundary condition and $n = 2$. In the considerations below, we denote by r_i^T the remainder terms associated with the ith order Taylor expansion of a mapping T. For instance, the following first and second order expansions of $b(x,y,u)$ are used at triplets (x,y,u) and $(x,\overline{y},\overline{u}) \in \mathbb{R}^{n+2}$:

$$b(x,y,u) - b(x,\overline{y},\overline{u}) = b'(x,\overline{y},\overline{u})(y-\overline{y},u-\overline{u}) + r_1^b, \tag{4.11}$$

where

$$r_1^b = (b_y^\vartheta - \overline{b}_y)(y-\overline{y}) + (b_u^\vartheta - \overline{b}_u)(u-\overline{u}) \tag{4.12}$$

and $b_y^\vartheta, b_u^\vartheta, \overline{b}_y, \overline{b}_u$ denote b_y, b_u taken at $(x,\overline{y}+\vartheta(y-\overline{y}),\overline{u}+\vartheta(u-\overline{u}))$ and $(x,\overline{y},\overline{u})$, respectively, with some $\vartheta \in (0,1)$. Expanding the same expression up to the order two, we have

$$\begin{aligned} b(x,y,u) - b(x,\overline{y},\overline{u}) &= b'(x,\overline{y},\overline{u})(y-\overline{y},u-\overline{u}) \\ &\quad + \frac{1}{2}(y-\overline{y},u-\overline{u})b''(x,\overline{y},\overline{u})\begin{pmatrix} y-\overline{y} \\ u-\overline{u}\end{pmatrix} + r_2^b, \end{aligned} \tag{4.13}$$

with the second order remainder term

$$r_2^b = \frac{1}{2}(y-\overline{y}, u-\overline{u})[b''^{,\vartheta} - \overline{b}''](y-\overline{y}, u-\overline{u})^T. \tag{4.14}$$

Here, $b''^{,\vartheta}$, $\overline{b}''$ denote the Hessian matrix of b with respect to (y,u) taken at the same triplets as above. Due to our assumptions on b' and b'', the estimates

$$|r_1^b| \le C_M(|y-\overline{y}|^2 + |u-\overline{u}|^2), \tag{4.15}$$

$$|r_2^b| \le C_M\,\eta(|y-\overline{y}| + |u-\overline{u}|)(|y-\overline{y}|^2 + |u-\overline{u}|^2) \tag{4.16}$$

are valid $\forall\ |y|, |\overline{y}|, |u|, |\overline{u}| \le M$. We continue with the discussion of the remainders $r_1^{\mathcal{L}}$ and $r_2^{\mathcal{L}}$. A Taylor expansion of $\mathcal{L}$ gives

$$\begin{aligned}
&\mathcal{L}(y,u,\overline{\varphi},\overline{\lambda},\overline{z}^*) - \mathcal{L}(\overline{y},\overline{u},\overline{\varphi},\overline{\lambda},\overline{z}^*)\\
&= \overline{\mathcal{L}}_y(y-\overline{y}) + \overline{\mathcal{L}}_u(u-\overline{u}) + r_1^{\mathcal{L}}\\
&= \overline{\mathcal{L}}_y(y-\overline{y}) + \overline{\mathcal{L}}_u(u-\overline{u}) + \frac{1}{2}\left(\overline{\mathcal{L}}_{yy}[y-\overline{y}]^2 + 2\overline{\mathcal{L}}_{yu}[y-\overline{y}, u-\overline{u}] + \overline{\mathcal{L}}_{uu}[u-\overline{u}]^2\right) + r_2^{\mathcal{L}},
\end{aligned}$$

where $\overline{\mathcal{L}}$ indicates that $\mathcal{L}$ and its derivatives are taken at $(\overline{y},\overline{u},\overline{\varphi},\overline{\lambda},\overline{z}^*)$. We have

$$\begin{aligned}
r_1^{\mathcal{L}} &= (\mathcal{L}_y^\vartheta - \overline{\mathcal{L}}_y)(y-\overline{y}) + (\mathcal{L}_u^\vartheta - \overline{\mathcal{L}}_u)(u-\overline{u})\\
r_2^{\mathcal{L}} &= \frac{1}{2}\big((\mathcal{L}_{yy}^\vartheta - \overline{\mathcal{L}}_{yy})[y-\overline{y}]^2 + 2(\mathcal{L}_{yu}^\vartheta - \overline{\mathcal{L}}_{yu})[y-\overline{y}, u-\overline{u}] + (\mathcal{L}_{uu}^\vartheta - \overline{\mathcal{L}}_{uu})[u-\overline{u}]^2\big).
\end{aligned}$$

$\mathcal{L}^\vartheta$ indicates that $(\overline{y}+\vartheta(y-\overline{y}), \overline{u}+\vartheta(u-\overline{u}), \overline{\varphi}, \overline{\lambda}, \overline{z}^*)$ is substituted for $(y,u,\overline{\varphi},\overline{\lambda},\overline{z}^*)$ in $\mathcal{L}'$ and $\mathcal{L}''$ with some $\vartheta \in (0,1)$. On account of the assumptions (A1)–(A4), we are able to verify

$$|r_1^{\mathcal{L}}| \le C_{\mathcal{L}}(\|y-\overline{y}\|_2^2 + \|u-\overline{u}\|_{L^2(\Gamma)}^2), \tag{4.17}$$

$$|r_2^{\mathcal{L}}| \le C_{\mathcal{L}}\,\eta(\|y-\overline{y}\|_{C(\overline{\Omega})} + \|u-\overline{u}\|_{L^\infty(\Gamma)})\cdot(\|y-\overline{y}\|_2^2 + \|u-\overline{u}\|_{L^2(\Gamma)}^2), \tag{4.18}$$

and

$$|\overline{\mathcal{L}}''[(y_1,u_1),(y_2,u_2)]| \le C_{\mathcal{L}}(\|y_1\|_2 + \|u_1\|_{L^2(\Gamma)})(\|y_2\|_2 + \|u_2\|_{L^2(\Gamma)}). \tag{4.19}$$

The constant $C_{\mathcal{L}} > 0$ depends in particular on $\overline{\varphi}$. For the definition of η we refer to the assumption (A2). The analysis of (4.17)–(4.19) is performed in Appendix 7.3.

5. Standard second order sufficient optimality condition. Our main aim is to establish sufficient optimality conditions close to the necessary ones derived in Casas and Tröltzsch [8]. Therefore, we include also certain *first order sufficient optimality conditions*. We shall combine an approach going back to Zowe and Maurer [15] with a splitting technique introduced by Dontchev et al. [10]. The method of [10] was focused on the optimal control of ordinary differential equations. It was extended later by the authors in [9] to the case of elliptic equations without state-constraints.

In [15], Maurer and Zowe introduced first order sufficient optimality conditions for differentiable optimization problems subject to a general constraint $g(w) \le 0$. For our problem, the application of their approach in its full generality is rather technical. Therefore, in an initial step we incorporate the first order sufficient optimality condition only for the constraints on the control. Later, we shall deal in the same way with additional state-constraints.

The role of first order sufficient conditions can be explained most easily by the minimization problem $\{\min f(x) \,|\, x_a \le x \le x_b\}$, where $f : \mathbb{R}^n \to \mathbb{R}$ is of class C^2. Let $\bar{x}$ satisfy the first order necessary conditions (variational inequality). If $n = 1$, then $f'(\bar{x}) \neq 0$ implies that $\bar{x}$ is a local minimizer (even for concave f). Therefore, the second order sufficient optimality condition $f''(\bar{x}) > 0$ is needed only in the case $f'(\bar{x}) = 0$, where the first order necessary condition is not sufficient. The situation is similar for $n > 1$: The positive definiteness of $f''(\bar{x})$ has to be required only on the subspace $\{x \in \mathbb{R}^n \,|\, x^i = 0 \text{ if } D^i f(\bar{x}) \neq 0\}$.

Define for fixed $\tau > 0$ (arbitrarily small) the set

$$\Gamma_\tau = \{x \in \Gamma \,|\, |g_u(x, \overline{y}(x), \overline{u}(x)) + \overline{\varphi}(x) b_u(x, \overline{y}(x), \overline{u}(x))| \ge \tau\}.$$

Γ_τ is a subset of "strongly active" control constraints (cf. (3.5)). In other words, $\Gamma_\tau = \{x \in \Gamma| \quad |\mathcal{L}_u(\bar{y}, \bar{u}, \bar{\varphi}, \bar{\lambda}, \bar{z}^*)(x)| \ge \tau\}$ is the set, where the gradient of the objective (expressed as a function of the control) is sufficiently steep. In the example above, τ can be chosen as the minimal value of all nonvanishing $|D^i f(\bar{x})|$.

We mention at this point the relation

$$\langle \overline{z}^*, E'(\overline{y}) y \rangle \le 0 \tag{5.1}$$

$\forall\ (y, u) \in L(\mathcal{M}, \overline{w})$, which follows from $\langle \overline{z}^*, E'(\overline{y}) y \rangle = \varrho \langle \overline{z}^*, \kappa - E(\overline{y}) \rangle \le 0$ in view of (3.4).

Let $\mathcal{P}_\tau : L^\infty(\Gamma) \to L^\infty(\Gamma)$ denote the projection operator $u \mapsto \chi_{\Gamma \setminus \Gamma_\tau} u = \mathcal{P}_\tau u$. In other words, $(\mathcal{P}_\tau u)(x) = u(x)$ holds on $\Gamma \setminus \Gamma_\tau$, while $(\mathcal{P}_\tau u)(x) = 0$ holds on Γ_τ. We begin with our first and at the same time simplest *second order sufficient optimality condition*.

(SSC) There exist positive numbers τ and δ such that

$$\mathcal{L}''(\overline{y}, \overline{u}, \overline{\varphi}, \overline{\lambda}, \overline{z}^*)[w_2, w_2] \ge \delta \|u_2\|^2_{L^2(\Gamma)} \tag{5.2}$$

holds for all pairs $w_2 = (y_2, u_2)$ constructed in the following way: For every $w = (y, u) \in L(\mathcal{M}, \overline{w})$, we split up the control part u in $u_1 = (u - \mathcal{P}_\tau u)$ and $u_2 = \mathcal{P}_\tau u$. The solutions of the linearized state equation

$$\begin{cases} -\Delta y_i + y_i = 0 & \text{in } \Omega, \\ \partial_\nu y_i = b_y(\cdot, \overline{y}, \overline{u}) y_i + b_u(\cdot, \overline{y}, \overline{u}) u_i & \text{on } \Gamma \end{cases} \tag{5.3}$$

associated with u_i are denoted by y_i, $i = 1, 2$. By this construction, we get the representation $w = w_1 + w_2 = (y_1, u_1) + (y_2, u_2)$.

REMARK 5.1. *The coercitivity condition* (5.2) *of* (SSC) *is required on the whole set* $L(\mathcal{M}, \overline{w})$ *if* Γ_τ *is empty. This rather strong second order condition is obtained by the formal setting* $\tau = \infty$.

THEOREM 5.2. *Let the feasible pair* $\overline{w} = (\overline{y}, \overline{u})$ *satisfy the regularity condition* (R)*, the first order necessary optimality conditions* (3.9)–(3.11)*, and the second order sufficient optimality condition* (SSC)*. Suppose further that the general assumptions* (A1)–(A4) *are satisfied. Then there are constants* $\varrho > 0$ *and* $\delta' > 0$ *such that*

$$F_0(\hat{y}, \hat{u}) \ge F_0(\overline{y}, \overline{u}) + \delta' \|\hat{u} - \overline{u}\|^2_{L^2(\Gamma)} \tag{5.4}$$

holds for all feasible pairs $\hat{w} = (\hat{y}, \hat{u})$ *such that*

$$\|\hat{u} - \overline{u}\|_{L^\infty(\Gamma)} < \varrho. \tag{5.5}$$

Proof. We denote by $\bar l = (\overline{\varphi}, \overline{\lambda}, \overline{z}^*)$ the triplet of Lagrange multipliers appearing in the first order necessary optimality conditions. Let an arbitrary feasible pair $\hat w = (\hat y, \hat u)$ be given. Then

$$F_0(\hat w) - F_0(\overline{w}) = \mathcal{L}(\hat w, \bar l) - \mathcal{L}(\overline{w}, \bar l) - \langle \overline{z}^*, E(\hat y) - E(\bar y)\rangle \tag{5.6}$$

follows from $F(\hat w) = F(\overline{w}) = 0$. The complementary slackness condition implies

$$-\langle \overline{z}^*, E(\hat y) - E(\bar y))\rangle \geq 0.$$

Hence we can neglect this term, and a second order Taylor expansion yields

$$\begin{aligned} F_0(\hat w) - F_0(\overline{w}) &\geq \mathcal{L}(\hat w, \bar l) - \mathcal{L}(\overline{w}, \bar l) \\ &\geq \int_\Gamma l_u\,(\hat u - \overline{u})\,dS + \frac{1}{2}\mathcal{L}''(\overline{w}, \bar l)[\hat w - \overline{w}]^2 + r_2^{\mathcal{L}}(\overline{w}, \hat w - \overline{w}), \end{aligned}$$

where $l_u(x) = g_u(x, \overline{y}(x), \overline{u}(x)) + \overline{\varphi}(x) b_u(x, \overline{y}(x), \overline{u}(x))$. Using the variational inequality, we find

$$F_0(\hat w) - F(\overline{w}) \geq \tau \int_{\Gamma_\tau} |\hat u - \overline{u}|\,dS + \frac{1}{2}\mathcal{L}''(\overline{w}, \bar l)[\hat w - \overline{w}]^2 + r_2^{\mathcal{L}}(\overline{w}, \hat w - \overline{w}). \tag{5.7}$$

Let us introduce for convenience the bilinear form $B = \mathcal{L}''(\overline{w}, \bar l)$. Next we approximate $\hat w - \overline{w}$ by $w = (y, u) \in L(\mathcal{M}, \overline{w})$, according to Theorem 4.2. In this way we get a remainder $r = (r^y, r^u) = \hat w - \overline{w} - w$ satisfying the estimate

$$\|r\| \leq C_L \|\hat u - \overline{u}\|_{L^\infty(\Gamma)} \|\hat u - \overline{u}\|_{L^2(\Gamma)}. \tag{5.8}$$

It follows that $B[\hat w - \overline{w}]^2 = B[w]^2 + 2B[r, w] + B[r]^2$. We have $w \in L(\mathcal{M}, \overline{w})$; hence (SSC) applies to $B[w]^2$. Now we substitute in $B[w]^2$ the representation $w = w_1 + w_2$ described in (SSC) and deduce

$$\begin{aligned} B[w]^2 &= B[w_2]^2 + 2B[w_1, w_2] + B[w_1]^2 \\ &\geq \delta \|u_2\|^2_{L^2(\Gamma)} - 2C_{\mathcal{L}}(\|y_1\|_2 + \|u_1\|_{L^2(\Gamma)})(\|y_2\|_2 + \|u_2\|_{L^2(\Gamma)}) \\ &\quad - C_{\mathcal{L}}(\|y_1\|_2 + \|u_1\|_{L^2(\Gamma)})^2 \end{aligned}$$

from (SSC) and (4.19). In the following, c will denote a generic constant. Suppose that $\varrho < 1$ is given and assume $\|\hat u - \overline{u}\|_{L^\infty(\Gamma)} < \varrho$. Then $\|y_i\|_2 \leq c\|u_i\|_{L^2(\Gamma)}$ and Young's inequality together yield

$$\begin{aligned} B[w]^2 &\geq \delta \|u_2\|^2_{L^2(\Gamma)} - \frac{\delta}{2}\|u_2\|^2_{L^2(\Gamma)} - c\|u_1\|^2_{L^2(\Gamma)} \\ &\geq \frac{\delta}{2}\int_{\Gamma\setminus\Gamma_\tau} u^2\,dS - c\int_{\Gamma_\tau} u^2\,dS \\ &\geq \frac{\delta}{2}\int_{\Gamma\setminus\Gamma_\tau} |\hat u - \overline{u}|^2\,dS - c\int_{\Gamma\setminus\Gamma_\tau} |\hat u - \overline{u}|\,|r^u|\,dS - c\int_{\Gamma_\tau} |\hat u - \overline{u}|^2\,dS \\ &\quad - c\int_{\Gamma_\tau} |\hat u - \overline{u}|\,|r^u|\,dS - c\int_{\Gamma_\tau} |r^u|^2\,dS. \end{aligned} \tag{5.9}$$

The expression under the third integral is estimated by $\|\hat u - \overline{u}\|_{L^\infty(\Gamma)}|\hat u - \overline{u}|$. In the other integrals (except the first) we insert (5.8) and derive

$$B[w]^2 \geq \frac{\delta}{2}\int_{\Gamma\setminus\Gamma_\tau} |\hat u - \overline{u}|^2\,dS - c\varrho\int_{\Gamma_\tau} |\hat u - \overline{u}|\,dS - c\varrho\|\hat u - \overline{u}\|^2_{L^2(\Gamma)}. \tag{5.10}$$

The treatment of $B[r,w]$ and $B[r]^2$ is simpler. We find

$$\begin{aligned}|B[r,w]| &\le c\|r\|\|u\|_{L^2(\Gamma)} = c\|r\|\|\hat u-\overline u+r^u\|_{L^2(\Gamma)}\\ &\le c\varrho\|\hat u-\overline u\|^2_{L^2(\Gamma)}.\end{aligned}$$

The same type of estimate applies to $B[r]^2$. Altogether,

$$B[\hat w-\overline w]^2 \ge \frac{\delta}{2}\int_{\Gamma\setminus\Gamma_\tau}|\hat u-\overline u|^2\,dS - c\varrho\int_{\Gamma_\tau}|\hat u-\overline u|\,dS - c\varrho\|\hat u-\overline u\|^2_{L^2(\Gamma)} \tag{5.11}$$

is obtained. By substituting (5.11) in (5.7), we get

$$\begin{aligned}F_0(\hat w)-F_0(\overline w) &\ge (\tau-c\varrho)\int_{\Gamma_\tau}|\hat u-\overline u|\,dS+\frac{\delta}{2}\int_{\Gamma\setminus\Gamma_\tau}|\hat u-\overline u|^2\,dS - c\varrho\|\hat u-\overline u\|^2_{L^2(\Gamma)}\\ &\quad -|r_2^{\mathcal L}(\overline w,\hat w-\overline w)|\\ &\ge \frac{\tau}{2}\int_{\Gamma_\tau}|\hat u-\overline u|\,dS+\frac{\delta}{2}\int_{\Gamma\setminus\Gamma_\tau}|\hat u-\overline u|^2\,dS - c\varrho\|\hat u-\overline u\|^2_{L^2(\Gamma)}\\ &\quad -|r_2^{\mathcal L}(\overline w,\hat w-\overline w)|.\end{aligned}$$

Since $\|\hat u-\overline u\|_{L^\infty(\Gamma)}\le 1$ was assumed, $|\hat u-\overline u|\ge|\hat u-\overline u|^2$ holds a.e. Using this in the first integral, setting $\delta'=\min\{\tau/2,\delta/2\}$, and substituting the estimate (4.18) for $r_2^{\mathcal L}$, we complete our estimation by

$$\begin{aligned}F_0(\hat w)-F_0(\overline w) &\ge \|\hat u-\overline u\|^2_{L^2(\Gamma)}(\delta'-c\varrho-\eta(c\|\hat u-\overline u\|_{L^\infty(\Gamma)}))\\ &\ge \frac{\delta'}{2}\|\hat u-\overline u\|^2_{L^2(\Gamma)}\end{aligned}$$

for sufficiently small $\varrho>0$. □

Our condition (SSC) does not have the form expected from a comparison with second order conditions in finite dimensional spaces. In particular, the pair (y_2,u_2) constructed in (SSC) does not in general belong to $L(\mathcal M,\overline w)$. To overcome this difficulty, we introduce another regularity condition $(\mathrm R)_\tau$ that is stronger than (R). This new constraint qualification is similar to that one used in Casas and Tröltzsch [8] to derive second order *necessary* conditions.

Let $\mathcal C_\tau(\overline u)$ denote the set of controls $u\in\mathcal C(\overline u)$ having the property $u(x)=0$ if $x\in\Gamma_\tau$. We strengthen (R) to

$$T'(\overline y)G'(\overline u)\mathcal C_\tau(\overline u)-\boldsymbol K(T(\overline y))=\boldsymbol Z. \tag{$\mathrm R)_\tau$}$$

On using $(\mathrm R)_\tau$, we are able to show that the following *second order sufficient optimality condition* implies (5.4) as well.

$(\mathrm{SSC})_\tau$ There exist positive numbers τ and δ such that

$$\mathcal L''(\overline y,\overline u,\overline\varphi,\overline\lambda,\overline z^*)[w,w]\ge\delta\|u\|^2_{L^2(\Gamma)} \tag{5.12}$$

holds for all pairs $w=(y,u)$ of $L(\mathcal M,\overline w)$ with the property $u(x)=0$ for almost every $x\in\Gamma_\tau$.

THEOREM 5.3. *Let the assumptions of Theorem* 5.2 *be fulfilled, where* (R) *and* (SSC) *are replaced by* $(\mathrm R)_\tau$ *and* $(\mathrm{SSC})_\tau$. *Then the assertion of Theorem* 5.2 *remains true.*

Proof. The proof is almost identical to that of Theorem 5.2. The only difference consists in a more detailed splitting. In the first part of the proof we repeat the steps up to the splitting $w = w_1 + w_2$ after (5.8). Define $\Phi = T \circ G$. Then we have

$$\Phi'(\overline{u})(u_1 + u_2) \in K(\Phi(\overline{u})),$$

as $w_1 + w_2 \in L(\mathcal{M}, \overline{w})$. Therefore,

$$\Phi'(\overline{u})u_2 \in K(\Phi(\overline{u})) - \Phi'(\overline{u})u_1$$

holds so that $w_2 = (y_2, u_2)$ does not in general belong to the linearized cone. Thanks to the regularity condition $(\mathrm{R})_\tau$, the linear version of the Robinson–Ursescu theorem (see Robinson [18]) implies the existence of u_H in $\mathcal{C}_\tau(\overline{u})$ with the following properties: The inclusion

$$\Phi'(\overline{u})u_H \in K(\Phi(\overline{u}))$$

holds, and

$$\|u_2 - u_H\|_{L^2(\Gamma)} \le c\|u_1\|_{L^2(\Gamma)} \tag{5.13}$$

is satisfied (see the proof of Theorem 4.2 in the appendix). In other words, we find a pair $w_H = (y_H, u_H)$ in $L(\mathcal{M}, \overline{w})$ with $u_H = 0$ on Γ_τ. Hence, (SSC) applies to $B[w_H]^2$. Moreover, the control u_H is sufficiently close to u_2.

Now we define $\tilde{u}_2 = u_H$ and $\tilde{u}_1 = u_1 + (u_2 - u_H)$. Further, let $\tilde{y}_i = G'(\overline{u})\tilde{u}_i$ denote the corresponding solution of the linearized state equation. Then $\tilde{w}_i = (\tilde{y}_i, \tilde{u}_i)$ is substituted for $w_i = (y_i, u_i)$, $i = 1, 2$. The only difference between the proofs of Theorem 5.2 and 5.3 appears between the formulas and (5.8) and (5.9): We use the splitting $w = \tilde{w}_1 + \tilde{w}_2$ instead of $w = w_1 + w_2$. Moreover, the first line of the estimate (5.9) is changed as follows:

$$\begin{aligned} B[w]^2 &\ge \delta \|\tilde{u}_2\|^2_{L^2(\Gamma)} - \frac{\delta}{4}\|\tilde{u}_2\|^2_{L^2(\Gamma)} - c\,\|\tilde{u}_1\|^2_{L^2(\Gamma)} \\ &\ge \frac{3\delta}{4}\,\|u_2 + (u_H - u_2)\|^2_{L^2(\Gamma)} - c\,\|u_1 + (u_2 - u_H)\|^2_{L^2(\Gamma)} \\ &\ge \frac{\delta}{2}\,\|u_2\|^2_{L^2(\Gamma)} - c\,\|u_1\|^2_{L^2(\Gamma)}, \end{aligned}$$

where we have used the estimate (5.13). Then we proceed word for word as in the proof of Theorem 5.2. □

Example. Let us briefly comment on (SSC) in the case of (P) for an active state constraint $\overline{y}(0) = y_0$. Then $L(\mathcal{M}, \overline{w})$ is expressed through (4.4), (4.5), and a quite strong second order condition is formulated by

$$\overline{\mathcal{L}}''(\overline{y}, \overline{u}, \overline{\varphi}, \overline{z}^*)[w, w] \ge \delta\|u\|^2_{L^2(\Gamma)} \tag{5.14}$$

for *all* $w = (y, u) \in L(\mathcal{M}, \overline{w})$. In this way, we would not take advantage of strongly active control constraints. These constraints appear on $\Gamma_\tau = \{x \in \Gamma \,|\, |\alpha\overline{u}(x) + \overline{\varphi}(x)| \ge \tau|\}$. We split $(y, u) = (y_1, u_1) + (y_2, u_2)$, where $u_2 = 0$ on Γ_τ and $u_1 = 0$ on $\Gamma \setminus \Gamma_\tau$. (SSC) requires the coercitivity condition (5.14) only for (y_2, u_2), while (y_1, u_1) is handled by first order sufficient optimality conditions. Notice that y_2 might violate the state-constraint $y(x_0) \le 0$. We avoid this by $(\mathrm{SSC})_\tau$: It requires the coercitivity

condition for the following $u \in \mathcal{C}(\overline{u})$: They vanish on Γ_τ and satisfy together with the associated solution y of the linearized partial differential equation the state-constraint $y(x_0) \le 0$.

The paper [15] shows that "strongly active" state-constraints may also contribute terms to the first order sufficient optimality conditions. However, this leads to a rather technical construction and more restrictive assumptions. We have to suppose that the function b is linear with respect to the control u and $n = 2$. The corresponding theorem is stated below. Define for fixed $\beta > 0$ and $\tau > 0$ the following subset of $L(\mathcal{M}, \overline{w})$:

$$L_{\beta,\tau}(\mathcal{M}, \overline{w}) = \{w \,|\, w = (y, u) \in L(\mathcal{M}, \overline{w}) \text{ and } w \text{ satisfies (5.15) below}\}.$$

The decisive inequality characterizing $L_{\beta,\tau}$ is

$$\langle \overline{z}^*, E'(\overline{y})y\rangle \ge -\beta \int_{\Gamma\setminus\Gamma_\tau} |u(x)|\, dS(x). \tag{5.15}$$

$L_{\beta,\tau}(\mathcal{M}, \overline{w})$ is the subset of $L(\mathcal{M}, \overline{w})$, where the term $\langle \overline{z}^*, E(y)\rangle$ does not much contribute to the first order sufficient optimality condition. It is only this set $L_{\beta,\tau}(\mathcal{M}, \bar{w})$ where we have to require second order conditions, namely, the following condition.

(SSC′) There exist positive numbers β, τ, and δ such that

$$\mathcal{L}''(\overline{y}, \overline{u}, \overline{\varphi}, \overline{\lambda}, \overline{z}^*)[w_2, w_2] \ge \delta \|u_2\|^2_{L^2(\Gamma)} \tag{5.16}$$

holds for all $w_2 = (y_2, u_2)$ obtained in the same way introduced in (SSC) by elements w taken from the smaller set $L_{\beta,\tau}(\mathcal{M}, \overline{w})$.

Using this condition, we formulate the following.

THEOREM 5.4. *Let the feasible pair $\overline{w} = (\overline{y}, \overline{u})$ satisfy the regularity condition* (R)*, the first order necessary optimality conditions* (3.9)–(3.11)*, and the second order sufficient optimality condition* (SSC′)*. Suppose further that the general assumptions* (A1)–(A4) *are satisfied. Moreover, assume that $n = 2$ and $b(x, y, u) = b_1(x, y) + b_2(x)u$. Then there are constants $\varrho > 0$ and $\delta' > 0$ such that*

$$F_0(\hat{y}, \hat{u}) \ge F_0(\overline{y}, \overline{u}) + \delta' \|\hat{u} - \overline{u}\|^2_{L^2(\Gamma)} \tag{5.17}$$

holds for all feasible pairs $\hat{w} = (\hat{y}, \hat{u})$ satisfying

$$\|\hat{u} - \overline{u}\|_{L^\infty(\Gamma)} < \varrho. \tag{5.18}$$

Proof. We begin in the way we have shown Theorem 5.2 by

$$F_0(\hat{w}) - F_0(\overline{w}) = \mathcal{L}(\hat{w}, \bar{l}) - \mathcal{L}(\overline{w}, \bar{l}) - \langle \overline{z}^*, E(\hat{y}) - E(\bar{y})\rangle. \tag{5.19}$$

Once again, the representation $\hat{w} - \overline{w} = w + r$ is obtained. Now we distinguish between two cases.

Case I: $w = (y, u) \in L(\mathcal{M}, \overline{w})\setminus L_{\beta,\tau}(\mathcal{M}, \overline{w})$. This is the case where we deduce (5.17) from first order sufficiency. Here, the inequality

$$-\langle \overline{z}^*, E'(\overline{y})y\rangle > \beta \int_{\Gamma\setminus\Gamma_\tau} |u(x)|\, dS(x) \tag{5.20}$$

is fulfilled. We transform (5.19) as follows:

$$\begin{aligned} F_0(\hat w) - F_0(\overline w) &= \mathcal{L}'(\overline w, \bar l)(\hat w - \overline w) + r_1^{\mathcal{L}}(\overline w, \hat w - \overline w) - \langle \overline z^*, E(\hat y) - E(\overline y)\rangle \\ &= \mathcal{L}_y(\overline w, \bar l)(\hat y - \overline y) + \mathcal{L}_u(\overline w, \bar l)(\hat u - \overline u) - \langle \overline z^*, E'(\overline y)(\hat y - \overline y)\rangle \\ &\quad + r_1^{\mathcal{L}}(\overline w, \hat w - \overline w) - \langle \overline z^*, r_1^E(\overline y, \hat y - \overline y)\rangle \\ &= 0 + \int_\Gamma l_u(x)(\hat u(x) - \overline u(x))\, dS(x) - \langle \overline z^*, E'(\overline y) y\rangle \\ &\quad + r_1^{\mathcal{L}}(\overline w, \hat w - \overline w) - \langle \overline z^*, E'(\overline y) r^y + r_1^E(\overline y, \hat y - \overline y)\rangle, \end{aligned} \tag{5.21}$$

where $l_u(x)$ stands for $g_u(x, \overline y(x), \overline u(x)) + \overline\varphi(x) b_u(x, \overline y(x), \overline u(x))$.

Owing to $n = 2$ and $b(x, y, u) = b_1(x, y) + b_2(x)u$, we are able to apply the strong estimate (4.8) with $p = 2$. This yields

$$\|r\|_{Y \times L^\infty(\Gamma)} \le C_{L,2} \|\hat u - \overline u\|^2_{L^2(\Gamma)}. \tag{5.22}$$

By Theorem 4.2, (5.22), (4.17), and (A3 (ii)) we have

$$\max\{\|r^y\|_2, |r_1^{\mathcal{L}}|, \|r_1^E\|_Z\} \le c(\|\hat y - \overline y\|_2^2 + \|\hat u - \overline u\|^2_{L^2(\Gamma)}).$$

Now the Lipschitz property of the mapping $u \mapsto y(u) = G(u)$ from $L^2(\Gamma)$ into $C(\overline\Omega)$ (note that $n = 2$) permits us to estimate the last three items of (5.21) by $c\,\|\hat u - \overline u\|^2_{L^2(\Gamma)}$. (5.20) is applied to the second one, while the first one is treated by Γ_τ: We know that

$$l_u(x)(\hat u(x) - \overline u(x)) \ge 0 \quad \text{a.e. on } \Gamma;$$

hence

$$\int_\Gamma l_u\, (\hat u - \overline u)\, dS \ge \int_{\Gamma_\tau} l_u(\hat u - \overline u)\, dS = \int_{\Gamma_\tau} |l_u|\, |\hat u - \overline u|\, dS \ge \tau \int_{\Gamma_\tau} |\hat u - \overline u|\, dS.$$

Inserting this in (5.21) we continue with

$$\begin{aligned} F_0(\hat w) - F_0(\overline w) &\ge \tau \int_{\Gamma_\tau} |\hat u - \overline u|\, dS + \beta \int_{\Gamma \setminus \Gamma_\tau} |u|\, dS - c\|\hat u - \overline u\|^2_{L^2(\Gamma)} \\ &\ge \tau \int_{\Gamma_\tau} |\hat u - \overline u|\, dS + \beta \int_{\Gamma \setminus \Gamma_\tau} |\hat u - \overline u|\, dS - c\|\hat u - \overline u\|^2_{L^2(\Gamma)} \end{aligned}$$

in view of $\|r^u\|_{L^\infty(\Gamma)} \le c\|\hat u - \overline u\|^2_{L^2(\Gamma)}$. Proceeding with the estimation, we deduce

$$\begin{aligned} F_0(\hat w) - F_0(\overline w) &\ge \min\{\beta, \tau\} \|\hat u - \overline u\|_{L^1(\Gamma)} - c\varrho \|\hat u - \overline u\|_{L^1(\Gamma)} \\ &\ge \beta' \|\hat u - \overline u\|_{L^1(\Gamma)} \end{aligned}$$

with some $\beta' > 0$, provided that $\|\hat u - \overline u\|_{L^\infty(\Gamma)} \le \varrho \le \varrho_1$ is fulfilled and ϱ_1 is sufficiently small. Assume additionally that $\varrho_1 \le 1$. Then $|\hat u - \overline u|^2 \le |\hat u - \overline u|$ holds a.e.; hence

$$F_0(\hat w) - F_0(\overline w) \ge \beta' \|\hat u - \overline u\|^2_{L^2(\Gamma)} \tag{5.23}$$

follows for $\|\hat u - \overline u\|_{L^\infty(\Gamma)} \le \varrho_1$.

Case II: $w \in L_{\beta,\tau}(\mathcal{M}, \overline w)$ (partial use of first order sufficient optimality conditions). Here, we neglect the term $\langle \overline z^*, E(\hat y) - E(\overline y)\rangle$ and proceed word for word as in the proof of Theorem 5.2, using $L_{\beta,\tau}$ instead of L. ∎

REMARK 5.5. *Unfortunately, the definition of* $L_{\beta,\tau}(\mathcal{M},\overline{w})$ *is not constructive. It is difficult to describe in an explicit way which* $(y,u)\in L(\mathcal{M},\overline{w})$ *belong to the different cases* I *or* II. *Therefore, this type of first order sufficient condition is only of limited value (see, for instance, the next example).*

Example. To illustrate (SSC′) for (P) in comparison with (SSC), let us assume for simplicity $\overline{u}\in \operatorname{int}\mathcal{U}^{ad}$, hence $\Gamma_\tau=\emptyset$. Then (SSC) requires the coercitivity condition (5.14) on the whole set $L(\mathcal{M},\overline{w})$. If $\overline{y}(0)=y_0$ and $\overline{z}^*>0$ (strong complementarity), then (SSC′) is weaker than (SSC): (5.14) is not needed for all $(y,u)\in L(\mathcal{M},\overline{w})$ satisfying

$$-\overline{z}^*y(0)\ge\beta\int_\Gamma|u(x)|dS. \tag{5.24}$$

Assume that y can be represented by a positive Green's function $G=G(x,\xi)$,

$$y(0)=\int_\Gamma G(0,\xi)u(\xi)dS(\xi),$$

such that $G(0,\xi)\ge\gamma>0$ on Γ. Then (5.24) is fulfilled with $\beta=\overline{z}^*\gamma\ \forall\ u\le 0$. Moreover, all $u\ge 0$, $u\ne 0$ do not contribute to $L(\mathcal{M},\overline{w})$. Therefore, the coercitivity condition (5.14) is needed only for all u having positive and negative parts U_+ and U_-, where U_+ dominates U_-. However, this information does not essentially improve (SSC).

REMARK 5.6. *Theorem* 5.2 *follows from Theorem* 5.4 *by setting* $\beta=0$, *where we can avoid the restrictions* $n=2$ *and* $b(x,y,u)=b_1(x,y)+b_2(x)u$.

The cone $\mathcal{C}(\overline{u})$ *is defined by* $\mathcal{C}(\overline{u})=\{\rho(u-\overline{u})\,|\,u\in\mathcal{U}^{ad},\rho\ge 0\}$. *Its closure in* $L^2(\Gamma)$ *is*

$$\operatorname{cl}\mathcal{C}(\overline{u})=\{v\in L^2(\Gamma)\,|\,v(x)\le 0\ \textit{if}\ \overline{u}(x)=u_b(x),\ v(x)\ge 0,\ \textit{if}\ \overline{u}(x)=u_a(x)\}.$$

Let us redefine $L(\mathcal{M},\overline{w})$ *by substituting* $\operatorname{cl}\mathcal{C}(\overline{u})$ *for* $\mathcal{C}(\overline{u})$ *and require* (SSC) *in this form. Then* (SSC) *appears to be stronger, and Theorem* 5.2 *holds as well, since* $\operatorname{cl}\mathcal{C}(\overline{u})\supset\mathcal{C}(\overline{u})$. *However, it can be proved by* (R) *and the generalized open mapping theorem that* (SSC) *based on* $\operatorname{cl}\mathcal{C}(\overline{u})$ *is in fact equivalent to* (SSC) *established with* $\mathcal{C}(\overline{u})$. *This follows by continuity arguments.*

6. Extended second order conditions. A study of the preceding sections reveals that (SSC) is sufficient for local optimality in any dimension of Ω without restrictions on the form of the nonlinear function b, whenever (A3) is satisfied and $\overline{\varphi}\in L^\infty(\Gamma)$. $\overline{\varphi}$ is bounded and measurable if pointwise state-constraints are given only in compact subsets of Ω with the other quantities being sufficiently smooth. In two-dimensional domains, pointwise state-constraints can be imposed on $\bar\Omega$, if $b(x,y,u)$ is linear with respect to u. An extension to $\overline{\varphi}\in L^r(\Gamma)$ requires stronger assumptions on b. However, we shall briefly sketch in this section that some extended form of (SSC) may partially improve the results for $n\le 3$.

Let us assume $\overline{\varphi}\notin L^\infty(\Gamma)$. Then it seems to be natural to introduce in $L^\infty(\Gamma)$ another norm

$$\|u\|_\varphi=\left(\int_\Gamma(1+|\overline{\varphi}(x)|)u^2(x)\,dS(x)\right)^{1/2}.$$

This definition is justified, as $u\in L^\infty(\Gamma)$ and $y\in C(\overline{\Omega})$ holds in all parts of our paper. For $\overline{\varphi}\in L^\infty(\Gamma)$, the new norm is equivalent to $\|u\|_{L^2(\Gamma)}$. To get rid of the restrictions

imposed on b in (A4) we redefine the set of strongly active control constraints Γ_τ by

$$\Gamma_{\tau,\varphi} = \{x \in \Gamma \,|\, |g_u(x,\overline{y}(x),\overline{u}(x)) + \overline{\varphi}(x) b_u(x,\overline{y}(x),\overline{u}(x))| \geq \tau(1+|\overline{\varphi}(x)|)\}. \tag{6.1}$$

Moreover, we substitute the condition

$$\mathcal{L}''(\overline{y},\overline{u},\overline{\varphi},\overline{\lambda},\overline{z}^*)[w_2,w_2] \geq \delta \|u_2\|_\varphi^2 \tag{6.2}$$

for (5.2). If $\overline{\varphi} \notin L^\infty(\Gamma)$, then (6.2) is stronger than (5.2). On the other hand, the term $\int_\Gamma \overline{\varphi} b_{uu} u_2^2 dS$ contributes to $\mathcal{L}''$. (SSC) implies (at least) the nonnegativity of $\overline{\varphi} b_{uu}$; hence

$$\int_\Gamma \overline{\varphi} b_{uu} u_2^2 dS = \int_\Gamma |\overline{\varphi}|\, |b_{uu}| u_2^2 dS \geq \kappa \int_\Gamma |\overline{\varphi}| u_2^2 dS$$

holds, provided that $|b_{uu}| \geq \kappa$. In view of this, (6.2) appears quite natural.

Now Theorem 5.2 *remains true for* $n \leq 3$ *without assumption* (A4).

This statement is easy to verify. Apart from the estimates (4.17)–(4.19), our theory is not influenced by introducing $\|u\|_\varphi$. The discussion of (4.17)–(4.19) is the decisive point. We are able to replace $\|\cdot\|_{L^2(\Gamma)}$ by $\|\cdot\|_\varphi$ there, as the basic inequalities (7.14)–(7.16) (Appendix 7.3) can be slightly reformulated: (7.14) is nothing more than

$$\int_\Gamma |\overline{\varphi}| u^2 dS \leq \|u\|_\varphi^2, \tag{6.3}$$

while (7.16) remains unchanged ($n = 2, 3$). Only (7.15) has to be substituted by

$$\begin{aligned} \int_\Gamma |\overline{\varphi}|\, |y|\, |u|\, dS &= \int_\Gamma |\overline{\varphi}|^{1/2} |y|\, |\overline{\varphi}|^{1/2} |u|\, dS \leq \|u\|_\varphi \left(\int_\Gamma |\overline{\varphi}| y^2 dS \right)^{1/2} \\ &\leq \|\overline{\varphi}\|^{1/2}_{L^{s/(s-2)}(\Gamma)} \|y\|_{L^s(\Gamma)} \|u\|_\varphi. \end{aligned} \tag{6.4}$$

Here we have invoked (7.15) for sufficiently large s ($n = 2, 3$). Now a careful study of the proof of Theorem 5.2 shows that (A4) can be removed on using (6.3) and (6.4). Assuming (6.2), we arrive at the estimate (5.4) with $\|\hat{u} - \overline{u}\|_\varphi^2$ instead of $\|\hat{u} - \overline{u}\|_{L^2(\Gamma)}^2$. Then (5.4) follows from $\|u\|_\varphi \geq \|u\|_{L^2(\Gamma)}$. The same arguments apply to the first order sufficient conditions in Theorem 5.4 for $n = 2$ if we redefine $L_{\beta,\tau}(\mathcal{M},\overline{w})$ by substituting for (5.15) the inequality

$$\langle \overline{z}^*, E'(\overline{y}) y \rangle \geq -\beta \int_{\Gamma \setminus \Gamma_\tau} (1 + |\overline{\varphi}|) |u|\, dS. \tag{6.5}$$

7. Appendix.

7.1. On the regularity condition. Regard the state equation (4.1) linearized at $(\overline{y},\overline{u})$. Let $\hat{Y} \subset H^1(\Omega)$ be the set of all solutions of this equation associated to $u \in L^\infty(\Gamma)$. In other words, we have $\hat{Y} = G'(\overline{u}) L^\infty(\Gamma)$. (R) is satisfied in the following particular cases.

(a) $K = Z$ (no inequality constraints). Then (R) means $F'(\overline{y}) G'(\overline{u}) \mathcal{C}(\overline{u}) = \mathbb{R}^m$. This condition is satisfied if, in addition to the surjectivity property $F'(\overline{y})\hat{Y} = \mathbb{R}^m$, the following holds: There is a $\tilde{u} \in \operatorname{int}_{L^\infty(\Gamma)} \mathcal{U}^{ad}$ with $F'(\overline{y})\tilde{y} = 0$. Here, $\tilde{y}$ denotes the solution of the linearized state equation (4.1) associated with $\tilde{u} - \overline{u}$, that is,

$\tilde{y} = G'(\overline{u})(\tilde{u} - \overline{u})$. The proof follows from [22, Lemma 1.2.2].

(b) $F = 0$ (no equality constraints). In this case, (R) is read as $E'(\overline{y})G'(\overline{u})\mathcal{C}(\overline{u}) - K(E(\overline{u})) = Z$. Once again, (R) is implied by two separate conditions: We assume $E'(\overline{y})\hat{Y} - K(E(\overline{y})) = Z$ and require the existence of an $\tilde{u} \in \operatorname{int}_{L^\infty(\Gamma)} \mathcal{U}^{ad}$ with the property that $E'(\overline{y})\tilde{y} \in K(E(\overline{y}))$ holds at $\tilde{y} = G'(\overline{u})(\tilde{u} - \overline{u})$ [22, Lemma 1.2.2]. It should be mentioned that case (a) follows from (b).

Example. (P) is worth discussing in this context. If the state constraint $y(0) \le y_0$ is not active at $\overline{y}$, then (R) is obviously satisfied. Therefore, we assume $\overline{y}(0) = y_0$ and get $K(E(\overline{y})) = \{z \in \mathbb{R} \,|\, z \le 0\} = \mathbb{R}^-$. Then $E'(\overline{y})\hat{Y} - K(E(\overline{y})) = Z$ reduces to the following requirement: For every $z \in \mathbb{R}^-$ there exists a function $u \in L^\infty(\Gamma)$ such that the equation $y(0) = z$ is satisfied by the corresponding solution y of the linearized equation (4.4). This property is fulfilled, since we may find at least one $u \in L^\infty(\Gamma)$ such that $y(0) \ne 0$. Hence, (R) is implied by the following conditions: There are $\tilde{u} \in L^\infty(\Gamma)$ and $\varepsilon > 0$ such that $|\tilde{u}| \le 1 - \varepsilon$ holds and that the solution $\tilde{y}$ of (4.4) corresponding to $\tilde{u} - \overline{u}$ satisfies $\tilde{y}(0) \le 0$.

(c) General case. Let us assume $\operatorname{int}_Z K \ne \emptyset$ and $\operatorname{int}_{L^\infty(\Gamma)} \mathcal{U}^{ad} \ne \emptyset$. We require the surjectivity property

$$F'(\overline{y})\hat{Y} = \mathbb{R}^m. \tag{7.1}$$

Moreover, assume the existence of a $\tilde{u} \in \operatorname{int}_{L^\infty(\Gamma)} \mathcal{U}^{ad}$ such that

$$E(\overline{y}) + E'(\overline{y})\tilde{y} \in \operatorname{int}_Z K, \tag{7.2}$$

$$F'(\overline{y})\tilde{y} = 0 \tag{7.3}$$

holds for $\tilde{y} = G'(\overline{u})(\tilde{u} - \overline{u})$. Then (R) is fulfilled. To show this, we first mention the simple fact that $\tilde{z} \in \operatorname{int}_Z K$ implies $\tilde{z} + z/\varrho \in K$ for arbitrary $z \in Z$ if ϱ is sufficiently large. We have to verify that the system

$$F'(\overline{y})y = z_1, \tag{7.4}$$

$$E'(\overline{y})y - \varrho(k - E(\overline{y})) = z_2 \tag{7.5}$$

is solvable $\forall\ z_1 \in \mathbb{R}^m$, $z_2 \in Z$ by some $y \in G'(\overline{u})\mathcal{C}(\overline{u})$, $k \in K$, and $\varrho \ge 0$: From (7.1) we find $u_1 \in L^\infty(\Gamma)$ such that $y_1 = G'(\overline{u})u_1$ solves the equation

$$F'(\overline{y})y_1 = z_1.$$

Now we add to y_1 a multiple of $\tilde{y}$. Then

$$F'(\overline{y})(y_1 + \varrho\tilde{y}) = F'(\overline{y})y_1 = z_1$$

is obtained from (7.3). Consequently, (7.4) holds for $y = y_1 + \varrho\tilde{y}$. Moreover, we deduce from (7.2) for sufficiently large ϱ that

$$E(\overline{y}) + E'(\overline{y})\tilde{y} - \frac{1}{\varrho}(z_2 - E'(\overline{y})y_1) = k \in K.$$

This relation is equivalent to

$$E'(\overline{y})(y_1 + \varrho\tilde{y}) - \varrho(k - E(\overline{y})) = z_2.$$

Therefore, (7.5) is satisfied by $y = y_1 + \varrho\tilde{y}$. Furthermore, $u_1 + \varrho(\tilde{u} - \overline{u}) = \varrho(\tilde{u} + (1/\varrho)u_1 - \overline{u}) \in \mathcal{C}(\overline{u})$ holds for sufficiently large ϱ. This is true, since $\tilde{u} + (1/\varrho)u_1 \in \mathcal{U}^{ad}$ for ϱ large enough (notice that $\tilde{u} \in \operatorname{int}_{L^\infty(\Gamma)} \mathcal{U}^{ad}$). Thus we have also shown $y \in G'(\overline{u})\mathcal{C}(\overline{u})$.

7.2. Proof of the linearization theorem. To prove Theorem 4.2 we need the following auxiliary result.

LEMMA 7.1. *Let $\overline{u}, \hat{u} \in \mathcal{U}^{ad}$ be given with associated states $\overline{y}, \hat{y}$ defined by* (2.2). *Introduce $y \in Y$ as the solution of the linearized state equation*

$$\begin{cases} -\Delta y + y = 0 & \text{in } \Omega, \\ \partial_\nu y = b_y(\cdot, \overline{y}, \overline{u})y + b_u(\cdot, \overline{y}, \overline{u})(\hat{u} - \overline{u}) & \text{on } \Gamma. \end{cases} \tag{7.6}$$

Then the estimates

$$\|\hat{y} - \overline{y} - y\|_Y \le C_p \|\hat{u} - \overline{u}\|_{L^\infty(\Gamma)} \|\hat{u} - \overline{u}\|_{L^p(\Gamma)} \quad \forall p > n - 1, \tag{7.7}$$

$$\|\hat{y} - \overline{y} - y\|_2 \le C_2 \|\hat{u} - \overline{u}\|_{L^\infty(\Gamma)} \|\hat{u} - \overline{u}\|_{L^2(\Gamma)} \tag{7.8}$$

are satisfied with certain constants C_p, C_2. If $b_u(x, y, u)$ does not depend on y and u, then we have

$$\|\hat{y} - \overline{y} - y\|_Y \le C_p \|\hat{u} - \overline{u}\|^2_{L^p(\Gamma)} \quad \forall p > n - 1. \tag{7.9}$$

Proof. We use the first order expansion of b at $(x, \overline{y}, \overline{u})$ and obtain from (2.2), (7.6), and (4.11) the system

$$\begin{aligned} -\Delta(\hat{y} - \overline{y} - y) + (\hat{y} - \overline{y} - y) &= 0 \quad \text{in } \Omega, \\ \partial_\nu(\hat{y} - \overline{y} - y) - b_y(\cdot, \overline{y}, \overline{u})(\hat{y} - \overline{y} - y) &= r_1^b \quad \text{on } \Gamma, \end{aligned}$$

where

$$|r_1^b(x)| \le C_M(|\hat{y}(x) - \overline{y}(x)|^2 + |\hat{u}(x) - \overline{u}(x)|^2)$$

and M depends on $\mathcal{U}^{ad}$ (notice that the boundedness of $\mathcal{U}^{ad}$ implies a uniform bound on all admissible states). Therefore, the discussion of (3.12) yields for $p > n - 1$

$$\begin{aligned} \|\hat{y} - \overline{y} - y\|_Y &\le c\|r_1^b\|_{L^p(\Gamma)} \\ &\le c\left(\left(\int_\Gamma |\hat{y} - \overline{y}|^{2p} dS\right)^{\frac{1}{p}} + \left(\int_\Gamma |\hat{u} - \overline{u}|^{2p} dS\right)^{\frac{1}{p}}\right). \end{aligned}$$

The mapping $u \mapsto y = G(u)$ is Lipschitz from $L^p(\Gamma)$ to $C(\overline{\Omega})$ for $p > n - 1$. If $p = 2$, then the Lipschitz property holds in the norm $\|y\|_2$ for y. For $p > n - 1$, we continue by

$$\|\hat{y} - \overline{y} - y\|_Y \le c\left(\|\hat{u} - \overline{u}\|^2_{L^p(\Gamma)} + \|\hat{u} - \overline{u}\|_{L^\infty(\Gamma)} \|\hat{u} - \overline{u}\|_{L^p(\Gamma)}\right),$$

while $p = 2$ yields only

$$\|\hat{y} - \overline{y} - y\|_2 \le c\|\hat{u} - \overline{u}\|_{L^\infty(\Gamma)} \|\hat{u} - \overline{u}\|_{L^2(\Gamma)}.$$

We have shown (7.7) and (7.8). If b_u does not depend on (y, u), then $b_u(\cdot, \overline{y} + \vartheta(\hat{y} - \overline{y}), \overline{u} + \vartheta(\hat{u} - \overline{u})) = b_u(\cdot, \overline{y}, \overline{u})$; hence

$$|r_1^b| = |(b_y^\vartheta - \bar{b}_y)(\hat{y} - \overline{y})| \le c|\hat{y} - \overline{y}|^2.$$

This yields

$$\|r_1^b\|_{L^p(\Gamma)} \le c\left(\|\hat{y} - \overline{y}\|_{C(\Gamma)} \|\hat{y} - \overline{y}\|_{L^p(\Gamma)}\right) \le c\|\hat{u} - \overline{u}\|^2_{L^p(\Gamma)},$$

that is, (7.9). □

Proof of Theorem 4.2. Define $v = \hat{u} - \overline{u}$ and let $\tilde{y}$ denote the solution of the linear system (4.1) associated to $u := v$. We have $\tilde{y} = G'(\overline{u})v$, where $G : L^\infty(\Gamma) \to Y$ is the control-state mapping $u \mapsto y = G(u)$ for the nonlinear system (2.2). By Lemma 7.1,

$$\|\hat{y} - \overline{y} - \tilde{y}\|_Y \le e(v), \tag{7.10}$$

where $e(v)$ denotes the right-hand side of the estimates (7.7) and (7.9), respectively, depending on the assumptions on b. Let us introduce the mapping $\Phi(u) = T(G(u))$. Its derivative is $\Phi'(\overline{u})v = T'(\overline{y})G'(\overline{u})v$, and the regularity condition (R) can be rewritten as

$$\Phi'(\overline{u})\mathcal{C}(\overline{u}) - \boldsymbol{K}(\Phi(\overline{u})) = \boldsymbol{Z}.$$

We know that $\Phi(\hat{u}) \in K$, hence a Taylor expansion yields

$$\Phi(\hat{u}) = \Phi(\overline{u}) + \Phi'(\overline{u})(\hat{u} - \overline{u}) + r_1^\Phi, \tag{7.11}$$

where the norm of r_1^Φ can be estimated by

$$\|r_1^\Phi\|_Z \le c\, e(v). \tag{7.12}$$

Since $\Phi(\hat{u})$ and $k = \Phi(\overline{u})$ belong to K, (7.11) implies $\Phi'(\overline{u})(\hat{u} - \overline{u}) + k + r_1^\Phi \in K$; thus also

$$\Phi'(\overline{u})(\hat{u} - \overline{u}) \in -r_1^\Phi + K(\Phi(\overline{u})). \tag{7.13}$$

In other words, we have $\hat{u} - \overline{u} \in \mathcal{C}(\overline{u})$ and $\Phi'(\overline{u})(\hat{u} - \overline{u}) \le_{K(\Phi(\overline{u}))} -r_1^\Phi$, where $z \ge_{K(\Phi(\overline{u}))} 0$ is defined by $z \in K(\Phi(\overline{u}))$. Owing to (R), this inequality is regular in the sense of Robinson [18]. Therefore, we are able to apply the linear version of the Robinson–Ursescu theorem (see [18]): It implies the existence of a constant $C_R > 0$ and a $u \in \mathcal{C}(\overline{u})$ satisfying $\|u - (\hat{u} - \overline{u})\|_{L^\infty(\Gamma)} \le C_R \|r_1^\Phi\|_Z$ together with

$$\Phi'(\overline{u})u \in K(\Phi(\overline{u})).$$

Consequently, for $y = G'(\overline{u})u$, we have $(y, u) \in L(\mathcal{M}, \overline{w})$ and

$$\|u - (\hat{u} - \overline{u})\|_{L^\infty(\Gamma)} \le \tilde{c} e(v).$$

The estimates stated in (4.6) and (4.8) follow immediately.

(4.7) is proved completely analogous. Here, $e(v)$ is defined by (7.8), $\|\cdot\|_Y$ is to be replaced by $\|\cdot\|_2$, and $\|\cdot\|_{L^2(\Gamma)}$ is to be substituted for $\|\cdot\|_{L^\infty(\Gamma)}$. We rely on the continuity of $\Phi'(\overline{y})$ in the L^2-norm. □

7.3. Estimates of the Lagrange function. In this subsection we derive the estimates (4.17)–(4.19) for $r_1^{\mathcal{L}}$, $r_2^{\mathcal{L}}$, and $\mathcal{L}''$. They depend mainly on the estimation of I defined in (4.10), which is performed by the discussion of the following integrals:

$$\int_\Gamma |\overline{\varphi}|\, u^2\, dS \le c\|u\|^2_{L^2(\Gamma)}, \tag{7.14}$$

provided that assumption (A4 (i)) is fulfilled, and

$$\int_\Gamma |\overline{\varphi}|\,|y|\,|u|\,dS \le c\|\overline{\varphi}y\|_{L^2(\Gamma)}\|u\|_{L^2(\Gamma)} \le c\|\overline{\varphi}^2\|^{1/2}_{L^{(s/2)'}(\Gamma)}\|y^2\|^{1/2}_{L^{s/2}(\Gamma)}\|u\|_{L^2(\Gamma)}$$
$$\le c\|\overline{\varphi}\|_{L^{2s/(s-2)}(\Gamma)}\|y\|_{L^s(\Gamma)}\|u\|_{L^2(\Gamma)}. \tag{7.15}$$

These estimates are justified by (A4 (ii)): For $n = 2$ we know $y \in C(\Gamma)$ and $\varphi \in L^r(\Gamma)$ $\forall r < \infty$. If $n \ge 3$, then $y \in L^s(\Gamma)$ holds $\forall\ s < 2(n-1)/(n-3)$ (including $s < \infty$ for $n = 3$). The function $2s/(s-2) = 2/(1-1/s)$ is monotone decreasing. Therefore, $s \uparrow 2(n-1)/(n-3)$ implies $2s/(s-2) \downarrow n-1$, so that $\overline{\varphi} \in L^r(\Gamma)$ for some $r > n-1$ justifies (7.15) with a sufficiently large s. Finally,

$$\int_\Gamma |\overline{\varphi}|y^2\,dS \le \|\overline{\varphi}\|_{L^{(s/2)'}(\Gamma)}\|y^2\|_{L^{s/2}(\Gamma)} = \|\overline{\varphi}\|_{L^{s/(s-2)}(\Gamma)}\|y\|^2_{L^s(\Gamma)} \tag{7.16}$$

is estimated by (A4 (iii)): In the case $n = 2$ we can take $s = \infty$, as $y \in C(\Gamma)$ and $\varphi \in L^1(\Gamma)$ is true without any additional assumption. For $n = 3$ we know $y \in L^s(\Gamma)\ \forall\ s < \infty$. If $s \uparrow \infty$, then $s/(s-2) \downarrow 1 < n/(n-1)$. Since $\varphi \in L^r(\Gamma)$ holds $\forall\ r < n/(n-1)$, (7.16) is true for sufficiently large s. In the case $n \ge 4$ we repeat the analysis of the case $n \ge 3$. This leads to the additional assumption $\overline{\varphi} \in L^r(\Gamma)$ for some $r > \frac{n-1}{2}$. Now it is easy to derive the estimates (4.17)–(4.19) for $\mathcal{L}''$, $r_1^{\mathcal{L}}$, and $r_2^{\mathcal{L}}$: For instance, I in (4.10) is handled by (7.14)–(7.16), and

$$|I| \le \int_\Gamma |\overline{\varphi}|(|\overline{b}_{yy}|\,|y_1y_2| + |\overline{b}_{yu}|(|y_1u_2| + |y_2u_1|) + |\overline{b}_{uu}|\,|u_1u_2|\,dS$$
$$\le c(\|y_1\|_2 + \|u_1\|_{L^2(\Gamma)})(\|y_2\|_2 + \|u_2\|_{L^2(\Gamma)}),$$

as $\overline{b}_{yy}$, $\overline{b}_{yu}$, and $\overline{b}_{uu}$ belong to $L^\infty(\Gamma)$. The other parts of $\mathcal{L}''$ are discussed by means of (A1)–(A3). This yields (4.19) after easy evaluations. In the same way, the remainder terms are investigated. Here, the quantities in I are the most difficult ones again. For instance, (7.14)–(7.16) applies to discussing

$$|r_2^I| = \int_\Gamma |\overline{\varphi}|\ \big\{|b^\vartheta_{yy} - \overline{b}_{yy}|\,|y-\overline{y}|^2 + 2|b^\vartheta_{yu} - \overline{b}_{yu}|\,|y-\overline{y}|\,|u-\overline{u}|$$
$$+ |b^\vartheta_{uu} - \overline{b}_{uu}|\,|u-\overline{u}|^2\big\}\,dS$$
$$\le c\eta(\|y-\overline{y}\|_{C(\Gamma)} + \|u-\overline{u}\|_{L^\infty(\Gamma)})(\|y-\overline{y}\|_2^2 + \|u-\overline{u}\|^2_{L^2(\Gamma)}),$$

which contributes to $r_2^{\mathcal{L}}$. The other terms of $r_2^{\mathcal{L}}$ are handled by the estimates for second order derivatives in (A1)–(A3) in a direct way. Simple evaluations of this type verify (4.17)–(4.18). We leave the details to the reader.

REFERENCES

[1] R. A. Adams, *Sobolev Spaces*, Academic Press, New York, NY, 1978.

[2] W. Alt, *On the approximation of infinite optimization problems with an application to optimal control problems*, Appl. Math. Optim., 12 (1984), pp. 15–27.

[3] W. Alt, *The Lagrange-Newton method for infinite-dimensional optimization problems*, Numer. Funct. Anal. Optim., 11 (1990), pp. 201–224.

[4] W. Alt, *Discretization and mesh-independence of Newton's method for generalized equations*, in Mathematical Programming with Data Perturbations, Lecture Notes in Pure Appl. Math. 195, A. K. Fiacco, ed., Dekker, New York, 1998, pp. 1–30.

[5] F. Bonnans *Second order analysis for control constrained optimal control problems of semi-linear elliptic systems*, Appl. Math. Optim., 38 (1998), pp. 303–325.

[6] F. Bonnans, and F. Casas, *Contrôle de systémes elliptiques semilinéaires comportant des contraintes sur l' état*, Nonlinear Partial Differential Equations and Their Applications, H. Brezis and J. L. Lions, eds., College de France Seminar Vol. VIII; Pitman Res. Notes Math. Ser. 166, Longman, Harlow, UK, 1988, pp. 69–86.

[7] E. Casas, *Boundary control of semilinear elliptic equations with pointwise state constraints*, SIAM J. Control Optim., 31 (1993), pp. 993–1006.

[8] E. Casas, and F. Tröltzsch, *Second order necessary optimality conditions for some state-constrained control problems of semilinear elliptic equations*, Appl. Math. Optim., 39 (1999), pp. 211–228.

[9] E. Casas, F. Tröltzsch, and A. Unger, *Second order sufficient optimality conditions for a nonlinear elliptic control problem*, Z. Anal. Anwendungen, 15 (1996), pp. 687–707.

[10] A. L. Dontchev, W. W. Hager, A. B. Poore, and B. Yang, *Optimality, stability, and convergence in nonlinear control*, Appl. Math. Optim., 31 (1995), pp. 297–326.

[11] H. Goldberg and F. Tröltzsch, *Second order optimality conditions for a class of control problems governed by non–linear integral equations with applications to parabolic boundary control*, Optim., 20 (1989), pp. 687–698.

[12] H. Goldberg and F. Tröltzsch, *Second-order sufficient optimality conditions for a class of nonlinear parabolic boundary control problems*, SIAM J. Control Optim., 31 (1993), pp. 1007–1025.

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

[14] H. Maurer, *First and second order sufficient optimality conditions in mathematical programming and optimal control*, Math. Programming Study, 14 (1981), pp. 163–177.

[15] H. Maurer and J. Zowe, *First– and second–order conditions in infinite–dimensional programming problems*, Math. Programming, 16 (1979), pp. 98–110.

[16] M. K. V. Murthy and G. Stampacchia, *A variational inequality with mixed boundary conditions*, Israel J. Math., 13 (1972), pp. 188–224.

[17] J. Nečas, *Les Méthodes Directes en Théorie des Equations Elliptiques*, Editeurs Academia, Prague, 1967.

[18] S. M. Robinson, *Stability theory for systems of inequalities.* I. *Linear systems*, SIAM J. Numer. Anal., 12 (1975), pp. 754–769.

[19] G. Stampacchia, *Problemi al contorno ellittici con dati discontinui dotati di soluzioni Hölderiane*, Ann. Math. Pura Appl., 51 (1960), pp. 1–38.

[20] G. Stampacchia, *Equations Elliptiques du Second Ordre à Coefficients Discontinus*, Les Presses de l'Université de Montreal, Montreal, 1966.

[21] H. Triebel, *Interpolation Theory, Function Spaces, Differential Operators*, 2nd ed., J. A. Barth Verlag, Heidelberg, 1995.

[22] F. Tröltzsch, *Optimality Conditions for Parabolic Control Problems and Applications*, Teubner-Texte Math. 62, B. G. Teubner Verlagsgesellschaft, Leipzig, 1984.

[23] J. Zowe and S. Kurcyusz, *Regularity and stability for the mathematical programming problem in Banach spaces*, Appl. Math. Optim., 5 (1979), pp. 49–62.

SIAM J. CONTROL OPTIM.
Vol. 38, No. 5, pp. 1392–1407

RELATIONSHIP BETWEEN BACKWARD STOCHASTIC DIFFERENTIAL EQUATIONS AND STOCHASTIC CONTROLS: A LINEAR-QUADRATIC APPROACH*

MICHAEL KOHLMANN† AND XUN YU ZHOU‡

Abstract. It is well known that backward stochastic differential equations (BSDEs) stem from the study on the Pontryagin type maximum principle for optimal stochastic controls. A solution of a BSDE hits a given *terminal* value (which is a random variable) by virtue of an *additional* martingale term and an *indefinite* initial state. This paper attempts to explore the relationship between BSDEs and stochastic controls by interpreting BSDEs as some stochastic optimal control problems. More specifically, associated with a BSDE, a new stochastic control problem is introduced with the *same* dynamics but a *definite* given initial state. The martingale term in the original BSDE is regarded as the control, and the objective is to minimize the second moment of the difference between the terminal state and the terminal value given in the BSDE. This problem is solved in a closed form by the stochastic linear-quadratic (LQ) theory developed recently. The general result is then applied to the Black–Scholes model, where an optimal mean-variance hedging portfolio is obtained explicitly in terms of the option price. Finally, a modified model is investigated, where the difference between the state and the expectation of the given terminal value at *any* time is taken into account.

Key words. BSDE, stochastic control, LQ control, stochastic Riccati equation (SRE), Black–Scholes model

AMS subject classifications. 60H10, 49N10, 90A09

PII. S036301299834973X

1. Introduction.

Backward stochastic differential equation (BSDE) theory and applications have remained very active in recent years. Consider the linear BSDE

$$\begin{cases} dp(t) = [A(t)p(t) + B(t)q(t) + f(t)]dt + q(t)dW(t), \\ p(T) = \xi, \end{cases} \tag{1.1}$$

where ξ is a random variable that will become certain only at the terminal time T. As is well known the equation was initially introduced by Bismut [2, 3] when he was studying the adjoint (dual) equations associated with a stochastic maximum principle for stochastic optimal controls. Basically, (1.1) tells how to price the *marginal value* of the resource represented by the state variable in a random environment. The solution of (1.1) has two components, $p(\cdot)$ and $q(\cdot)$, the former being the price while the latter signifies the uncertainty between the present and terminal times. The linear BSDEs were later extended to nonlinear ones by Pardoux and Peng [21], motivated by stochastic control problems, and independently by Duffie and Epstein [6] in their study of recursive utility in finance. The BSDE theory has found wide applications

*Received by the editors December 23, 1998; accepted for publication (in revised form) October 25, 1999; published electronically May 11, 2000.
http://www.siam.org/journals/sicon/38-5/34973.html

†Fakultät für Mathematik und Informatik, Universität Konstanz, Postfach 5560, D-78457, Konstanz, Germany (michael.kohlmann@uni-konstanz.de). The research of this author was supported by the Center of Finance and Econometrics, Universität Konstanz.

‡Department of Systems Engineering and Engineering Management, The Chinese University of Hong Kong, Shatin, Hong Kong (xyzhou@se.cuhk.edu.hk). The research of this author was supported by the RGC Earmarked Grants CUHK 4125/97E, CUHK 4054/98E, and CUHK4435/99E and by the Deutscher Akademischer Austauschdienst (DAAD). This research was done while this author was visiting Prof. Dr. Michael Kohlmann at Konstanz.

in partial differential equations, stochastic controls, and, especially, mathematical finance. For the most updated accounts of the BSDE theory and applications see the books by Yong and Zhou [24, Chapter 7] and Ma and Yong [19].

While BSDEs originally arise from stochastic control problems, the purpose of this paper is to further investigate the relationship between BSDEs and stochastic controls from a different angle by interpreting a BSDE as a stochastic linear-quadratic (LQ) control problem (so it is a reverse of the original "birth process" of BSDEs!) which can be solved explicitly. To be precise, note that in (1.1) the terminal value is specified while the initial value is left open. But if the equation has a solution, then the initial value cannot be chosen arbitrarily; rather it is uniquely determined by the solution and is hence *part* of the solution. Therefore, solving (1.1) amounts to the following statement: start with a *proper* initial condition and choose an *appropriate* diffusion term to hit the given value ξ at the terminal.

Then it will be very natural to modify the above statement and consider the following stochastic optimal control problem. For the *same* dynamics of (1.1), starting with a *given* initial state x, choose a control $q(\cdot)$ so that the terminal state $p(T)$ stays as *close* to the given terminal value ξ as possible. Note that since now the initial value x is given a priori, one in general cannot expect that $p(T)$ will hit ξ exactly by choosing certain $q(\cdot)$. Hence it is reasonable to require that the *difference* between the two is minimized. Here, the "difference" may be measured by, say, the second moment of the algebraic difference between the two random variables. More interestingly, if we regard the initial state also as a decision variable, then the optimal state-control pair of the problem $(p(\cdot), q(\cdot))$ is exactly the solution of the original BSDE!

It turns out that the control problem formulated above is a stochastic optimal LQ problem that can be solved *analytically* via a stochastic Riccati equation (SRE), employing the similar technique as developed recently in [4, 5].

We then apply the general results obtained to the Black–Scholes model. Taking advantage of the fact that in this particular case the state (wealth) is a scalar, one can solve the SRE explicitly and hence the corresponding optimal mean-variance hedging problem. It turns out that an optimal portfolio consists of the replicating strategy for the claim and the Merton portfolio for a quadratic terminal utility.

Finally we consider a modified model where the difference between the state and the expected terminal value must be kept small at *any* time. Again explicit optimal control is derived via an SRE which is shown to be always solvable. The result is then applied to the Black–Scholes model which gives rise to an auxiliary process that dynamically corrects any large deviation of the price from the expected value of the claim.

It should be noted that the idea of regarding the martingale term $q(t)$ as a control variable was first employed by Ma and Yong [17, 18] to prove the solvability of certain class of nonlinear forward-backward stochastic differential equations (FBSDEs). Moreover, Yong [23], using the so-called four-step-scheme developed by Ma, Protter, and Yong [16], discussed the solvability of a linear FBSDE from a stochastic controllability point of view. This paper, however, concentrates on the *relation* between the BSDEs and stochastic controls instead of the solvability of BSDEs, as we believe that this relation is a fundamental issue in BSDE and control theory. To elaborate, a linear BSDE can be interpreted as a dual of the stochastic control problem, where the martingale term $q(t)$ is exactly the dual variable corresponding to the controlled diffusion term. Now, it turns out that this dual variable can also be regarded as a control variable by which the BSDE naturally leads to a stochastic control problem.

This reveals a certain "symmetric" duality relation between BSDEs and stochastic control problems.

The rest of the paper is organized as follows. In section 2 we formulate the model. Section 3 presents the optimal solution to the problem. Section 4 is concerned with the solvability of the SRE necessary for the optimal control derived in section 3. In section 5 we have a special case, namely, the Black–Scholes model is considered and an optimal hedging portfolio is derived explicitly based on the results of the previous sections. Section 6 is devoted to a modified model. Finally, section 7 concludes the paper.

2. Problem formulation. Throughout this paper $(\Omega, \mathcal{F}, P, \{\mathcal{F}_t\}_{t\geq 0})$ is a fixed filtered complete probability space on which is defined a standard $\mathcal{F}_t$-adapted m-dimensional Brownian motion $W(t) \equiv (W^1(t), \dots, W^m(t))'$ with $W(0) = 0$. It is assumed that $\mathcal{F}_t = \sigma\{W(s) : s \leq t\}$. We denote by $L^2_{\mathcal{F}}(0, T; R^d)$ the set of all R^d-valued, measurable stochastic processes $\psi(t)$ adapted to $\mathcal{F}_t$, such that $E\int_0^T |\psi(t)|^2 dt < +\infty$.

Notation. We make the following additional notation.

M' : the transpose of any vector or matrix M;

$|M| := \sqrt{\sum_{i,j} m_{ij}^2}$ for any matrix or vector $M = (m_{ij})$;

S^n : the space of all $n \times n$ symmetric matrices;

S^n_+ : the subspace of all nonnegative definite matrices of S^n;

$\hat{S}^n_+$: the subspace of all positive definite matrices of S^n;

$C([0, T]; X)$: the Banach space of X-valued continuous functions on $[0, T]$ endowed with the maximum norm $\|\cdot\|$ for a given Hilbert space X;

$L^2(0, T; X)$: the Hilbert space of X-valued integrable functions on $[0, T]$ endowed with the norm $\left(\int_0^T \| f(t) \|_X^2 \, dt\right)^{\frac{1}{2}}$ for a given Hilbert space X;

$L^\infty(0, T; X)$: the Banach space of X-valued essentially bounded functions on $[0, T]$ endowed with the norm $\sup_{0\leq t\leq T} \| f(t) \|_X$ for a given Hilbert space X.

Consider the controlled system

$$\begin{cases} dx(t) = \Big[A(t)x(t) + \sum_{j=1}^m B_j(t)u_j(t) + f(t)\Big] dt \\ \qquad\qquad + \sum_{j=1}^m u_j(t) dW^j(t), \ t \in [0, T], \\ x(0) = x, \end{cases} \tag{2.1}$$

where $x(t), x, u_j(t), f(t) \in R^n$, and $A(t), B_j(t) \in R^{n\times n}$. Throughout this paper we assume that $A(t), B_j(t)$ are bounded deterministic functions and $f \in L^2_{\mathcal{F}}(0, T; R^n)$. For a given $\mathcal{F}_T$-measurable square integrable random variable ξ (i.e., $E|\xi|^2 < +\infty$), the problem is to select an ($\mathcal{F}_t$-adapted) control process $u(\cdot) \equiv (u_1(\cdot), \dots, u_m(\cdot)) \in L^2_{\mathcal{F}}(0, T; R^{mn})$ so as to minimize the cost functional

$$J(x, u(\cdot)) = E\frac{1}{2}|x(T) - \xi|^2. \tag{2.2}$$

To simplify the cost functional, it is natural to define

$$y(t) = x(t) - E(\xi|\mathcal{F}_t). \tag{2.3}$$

Since $E(\xi|\mathcal{F}_t)$ is an $\mathcal{F}_t$-martingale and $\mathcal{F}_t$ is generated by the standard Brownian motion $W(t)$, by the martingale representation theorem ([14]) there is $z(\cdot) \equiv (z_1(\cdot), \dots, z_m(\cdot)) \in L^2_{\mathcal{F}}(0,T;R^{mn})$ so that

$$E(\xi|\mathcal{F}_t) = E\xi + \sum_{j=1}^{m} \int_0^t z_j(s) dW^j(s). \tag{2.4}$$

By (2.1), (2.3), and (2.4), with the new state variable $y(\cdot)$ the controlled system becomes

$$\begin{cases} dy(t) = \Big[A(t)y(t) + \sum_{j=1}^m B_j(t)u_j(t) + g(t)\Big]\, dt \\ \qquad\quad + \sum_{j=1}^m [u_j(t) - z_j(t)] dW^j(t),\ t \in [0,T], \\ y(0) = x - E\xi \equiv y, \end{cases} \tag{2.5}$$

where

$$g(t) = f(t) + A(t)E(\xi|\mathcal{F}_t), \tag{2.6}$$

and the cost functional reduces to

$$J(y, u(\cdot)) = E\frac{1}{2}|y(T)|^2. \tag{2.7}$$

Notice that the above problem is a stochastic LQ control problem with *random* nonhomogeneous terms in both drift and diffusion coefficients.

3. Solutions. In this section we solve the problem (2.1)–(2.2) or (2.5)–(2.7) by LQ techniques. The main idea is simply the *completion of squares*. It should be noted that the problem under consideration is a *singular* LQ problem in that the running cost is identically zero and therefore it cannot be solved by the conventional approach as developed by Wonham [22] and others. Indeed, study on the general (possibly singular) stochastic LQ problem is interesting in its own right and has recently been developed extensively (see [4, 5]). For a systematic treatment of stochastic LQ problems, see also [24, Chapter 6].

In the rest of this paper, we may write X for a (deterministic or stochastic) process $X(t)$, omitting the variable t, whenever no confusion arises. Under this convention, when $X \in C([0,T];S^n)$, $X \geq (>)0$ means $X(t) \geq (>)0,\ \forall t \in [0,T]$.

We introduce the following SRE

$$\begin{cases} \dot{P} + PA + A'P - \sum_{j=1}^m PB_jP^{-1}B_j'P = 0, \\ P(T) = I, \\ P(t) > 0 \quad \forall\, t \in [0,T], \end{cases} \tag{3.1}$$

along with a BSDE

$$\begin{cases} d\phi(t) = & -\Big[(A' - \sum_{j=1}^m PB_jP^{-1}B_j')\phi - \sum_{j=1}^m PB_jP^{-1}\beta_j \\ & +P(g + \sum_{j=1}^m B_jz_j)\Big] \\ & \quad (t)dt + \sum_{j=1}^m \beta_j(t)dW^j(t), \\ \phi(T) = & 0. \end{cases} \tag{3.2}$$

Note that the SRE (3.1) is fundamentally different from the *conventional* Riccati equation[1] in that (3.1) involves the *inverse* of the unknown. In addition, the third constraint of (3.1) must also be satisfied by any solution. In general, such an equation does *not* automatically admit a solution. (The solvability of this equation is interesting on its own; see section 4 below.) On the other hand, if (3.1) has a solution $P(\cdot) \in C([0,T]; R^{n\times n})$, then (3.2) must admit an $\mathcal{F}_t$-adapted solution $(\phi(\cdot), \beta_j(\cdot), j = 1,\dots,m)$ as (3.2) is a linear BSDE with bounded coefficients and square integrable nonhomogeneous terms; see [2, 3, 7, 21, 19] or [24, Chapter 7] for more details about BSDEs.

THEOREM 3.1. *If* (3.1) *and* (3.2) *admit solutions* $P \in C([0,T]; \hat{S}^n_+)$ *and* $(\phi(\cdot), \beta_j(\cdot), j = 1,\dots,m) \in L^2_{\mathcal{F}}(0,T;R^n) \times L^2_{\mathcal{F}}(0,T;R^{nm})$, *respectively, then the problem* (2.5)–(2.7) *has an optimal feedback control* $u^*(\cdot) \equiv (u_1^*(\cdot),\dots,u_m^*(\cdot))$, *where*

$$u_j^*(t) = -P(t)^{-1}B_j(t)'\big[P(t)y^*(t) + \phi(t)\big] - P(t)^{-1}\beta_j(t) + z_j(t),\ j = 1,\dots,m. \tag{3.3}$$

Moreover, the optimal cost value under the above control is

$$\begin{aligned} J^*(y) = {} & \tfrac{1}{2}y'P(0)y + y'\phi(0) + \tfrac{1}{2}E\int_0^T \Big[2\phi' g - 2\textstyle\sum_{j=1}^m \beta_j z_j + \sum_{j=1}^m z_j' P z_j \\ & - \textstyle\sum_{j=1}^m (P^{-1}B_j'\phi + P^{-1}\beta_j - z_j)'P(P^{-1}B_j'\phi + P^{-1}\beta_j - z_j)\Big](t)dt. \end{aligned} \tag{3.4}$$

Proof. For any control $u(\cdot)$ and the corresponding state trajectory $y(\cdot)$, applying Ito's formula, we get

$$\begin{aligned} \tfrac{1}{2}d[y(t)'P(t)y(t)] = {} & \tfrac{1}{2}\Big[\textstyle\sum_{j=1}^m (u_j - z_j)'P(u_j - z_j) + 2y'Pg \\ & + \textstyle\sum_{j=1}^m (y'PB_jP^{-1}B_j'Py + 2u_j'B_j'Py)\Big](t)dt + \tfrac{1}{2}\{\cdots\}dW(t), \end{aligned} \tag{3.5}$$

and

$$\begin{aligned} d[\phi(t)'y(t)] = {} & \Big[-\phi'(A - \textstyle\sum_{j=1}^m B_jP^{-1}B_j'P)y + \sum_{j=1}^m \beta_j'P^{-1}B_j'Py \\ & - \Big(g' + \textstyle\sum_{j=1}^m z_j'B_j'\Big)Py + \phi'\Big(Ay + \sum_{j=1}^m B_ju_j + g\Big) \\ & + \textstyle\sum_{j=1}^m \beta_j'(u_j - z_j)\Big](t)dt + \{\cdots\}dW(t). \end{aligned} \tag{3.6}$$

Then we integrate both (3.5) and (3.6) from 0 to T, take expectations, and add them together. Trying to complete a square and going through a fairly tedious manipulation, we end up with

$$\begin{aligned} & J(y,u(\cdot)) \\ & = \tfrac{1}{2}E\int_0^T \Big[2\phi' g - 2\textstyle\sum_{j=1}^m \beta_j z_j + \sum_{j=1}^m z_j' P z_j \\ & + \textstyle\sum_{j=1}^m (u_j + P^{-1}B_j'Py + P^{-1}B_j'\phi + P^{-1}\beta_j - z_j)'P(u_j + P^{-1}B_j'Py \\ & \qquad + P^{-1}B_j'\phi + P^{-1}\beta_j - z_j) \\ & - \textstyle\sum_{j=1}^m (P^{-1}B_j'\phi + P^{-1}\beta_j - z_j)'P(P^{-1}B_j'\phi + P^{-1}\beta_j - z_j)\Big](t)dt \\ & + \tfrac{1}{2}y'P(0)y + y'\phi(0). \end{aligned} \tag{3.7}$$

It follows immediately that the optimal feedback control is given by (3.3) and the optimal value is given by (3.4) *provided* that the corresponding equation (2.5) under (3.3) has a solution. But under the linear feedback (3.3), the system (2.5) is a

[1]By a conventional Riccati equation we mean one associated with the *deterministic* LQ problem (see [1]) as opposed to that associated with the *stochastic* LQ problem (see [22, 4]).

nonhomogeneous linear SDE with bounded coefficients and square integrable nonhomogeneous terms. Hence it must admit one and only one solution by standard SDE theory. This completes the proof. □

Now we would like to derive the optimal feedback control in terms of the original variable $x(t)$. Interestingly, the optimal control can be obtained via the original BSDE that motivated the optimal control problem (2.1)–(2.2).

THEOREM 3.2. *Under the same assumptions of Theorem* 3.1, *the control problem* (2.1)–(2.2) *has an optimal feedback control* $u^*(\cdot) \equiv (u_1^*(\cdot), \dots, u_m^*(\cdot))$, *where*

$$u_j^*(t) = -P(t)^{-1}B_j(t)'P(t)\big[x^*(t) - p(t)\big] + q_j(t),\ j = 1, \dots, m, \tag{3.8}$$

where $(p(\cdot), q_j(\cdot), j = 1, \dots, m) \in L^2_{\mathcal{F}}(0,T;R^n) \times L^2_{\mathcal{F}}(0,T;R^{nm})$ *is the unique* $\mathcal{F}_t$*-adapted solution of the following BSDE:*

$$\begin{cases} dp(t) = \Big[A(t)p(t) + \sum_{j=1}^m B_j(t)q_j(t) + f(t)\Big]\,dt \\ \qquad\quad + \sum_{j=1}^m q_j(t)dW^j(t),\ t \in [0,T], \\ p(T) = \xi. \end{cases} \tag{3.9}$$

Proof. First of all, consider the matrix-valued ODE

$$\begin{cases} \dot{Q} - AQ - QA' + \sum_{j=1}^m B_jQB_j' = 0, \\ Q(T) = I, \end{cases} \tag{3.10}$$

which must admit a unique solution $Q(\cdot)$ since it is linear with bounded coefficients. Denote $S = PQ$, and then by the differential chain rule it is easy to verify that S satisfies

$$\begin{cases} \dot{S} = SA' - A'S + \sum_{j=1}^m PB_jP^{-1}B_j'S - \sum_{j=1}^m PB_jP^{-1}SB_j', \\ S(T) = I. \end{cases} \tag{3.11}$$

This is a linear equation, and hence it has a unique solution $S \equiv I$. It leads to $Q(t) = P(t)^{-1}$.

Now, noting (2.3), the feedback control (3.3) can be written as

$$\begin{aligned} u_j^*(t) = \quad & -P(t)^{-1}B_j(t)'\big[P(t)x^*(t) - P(t)E(\xi|\mathcal{F}_t) + \phi(t)\big] \\ & -P(t)^{-1}\beta_j(t) + z_j(t), \quad j = 1, \dots, m. \end{aligned} \tag{3.12}$$

Denote

$$\begin{aligned} &p(t) = E(\xi|\mathcal{F}_t) - P(t)^{-1}\phi(t) \equiv E(\xi|\mathcal{F}_t) - Q(t)\phi(t), \\ &q_j(t) = z_j(t) - P(t)^{-1}\beta_j(t) \equiv z_j(t) - Q(t)\beta_j(t),\ j = 1, \dots, m. \end{aligned} \tag{3.13}$$

Since $Q(t)$ is bounded and $(p(\cdot), q_j(\cdot), j = 1, \dots, m) \in L^2_{\mathcal{F}}(0,T;R^n) \times L^2_{\mathcal{F}}(0,T;R^{nm})$, applying Ito's formula to (2.4), (3.10), and (3.2), we can verify that $(p(\cdot), q_j(\cdot), j = 1, \dots, m)$ satisfies the BSDE (3.9). Therefore the desired result follows by virtue of the uniqueness of solutions to (3.9). □

Remark 3.1. Equation (3.4) also gives the optimal cost functional value as a function of the initial value $y \equiv x - E\xi$, which turns out to be quadratic. If the controller has the choice of selecting the initial value y so as to minimize $J^*(y)$, then

the "best" initial value would be obtained by setting $\frac{d}{dy}J^*(y)|_{y=y^*} = 0$. This yields $y^* = -P(0)^{-1}\phi(0)$. Returning to the original variable x, we get that the best initial value for $x(\cdot)$ will be

$$x^* = y^* + E\xi = -P(0)^{-1}\phi(0) + E\xi = p(0), \tag{3.14}$$

where the last equality is due to (3.13). This certainly makes perfect sense, as it implies that one should choose the initial value $p(0)$ so as to minimize the difference between the terminal state value and the given value ξ. (Of course, in this case the minimum difference is zero since starting with $p(0)$ one can hit ξ *exactly* at the end, by the BSDE theory.) Therefore, the solution pair $(p(\cdot), q(\cdot))$ of the BSDE (3.9) may be regarded as the optimal state-control pair of minimizing $J(x, u(\cdot))$ (as given by (2.2)) over $(x, u(\cdot))$ subject to the dynamics of (2.1). This gives an interpretation of $(p(\cdot), q(\cdot))$ via a stochastic control problem. In this perspective, if a BSDE does not have an adapted solution (e.g., when the underlying filtration is *not* generated by the Brownian motion involved), we may still define a "pseudosolution" via the corresponding stochastic control problem.

Remark 3.2. Under the optimal feedback (3.8), the optimal trajectory $x^*(\cdot)$ evolves as

$$\begin{cases} dx^*(t) = & \left\{ \left[A - \sum_{j=1}^m B_j P^{-1} B_j' P\right] x^* + \sum_{j=1}^m B_j P^{-1} B_j' P p + f \right. \\ & \left. + \sum_{j=1}^m B_j q_j \right\}(t)dt \\ & + \sum_{j=1}^m [-P^{-1} B_j' P x^* + P^{-1} B_j' P p + q_j](t) dW^j(t), \\ x^*(0) = & x. \end{cases} \tag{3.15}$$

Moreover, the difference $\Delta(t) = x^*(t) - p(t)$ satisfies

$$\begin{cases} d\Delta(t) = & \left[A - \sum_{j=1}^m B_j P^{-1} B_j' P\right](t)\Delta(t)dt \\ & - \sum_{j=1}^m [P^{-1} B_j' P](t)\Delta(t) dW^j(t), \\ \Delta(0) = & x - p(0). \end{cases} \tag{3.16}$$

Notice that $\Delta(\cdot)$ satisfies a homogeneous linear SDE, and hence must be identically zero if the initial is zero, namely, if $x = p(0)$. In this case, by (3.8), the optimal control is $u_j^*(t) = q_j(t)$. This is exactly in line with the observation in Remark 3.1.

Remark 3.3. Applying Ito's formula, we can show that $\Lambda(t) \equiv \Delta(t)\Delta(t)'$ satisfies

$$\begin{cases} d\Lambda(t) = & \left[\left(A - \sum_{j=1}^m B_j P^{-1} B_j' P\right)\Lambda + \Lambda\left(A - \sum_{j=1}^m B_j P^{-1} B_j' P\right)'\right](t)dt \\ & + \sum_{j=1}^m [P^{-1} B_j' P \Lambda P B_j P^{-1}](t)dt + \sum_{j=1}^m \{\cdots\} dW^j(t), \\ \Lambda(0) = & (x - p(0))(x - p(0))'. \end{cases}$$

Taking expectations, we conclude that $\Gamma(t) \equiv E[\Delta(t)\Delta(t)']$ follows

$$\begin{cases} \dot{\Gamma} - \bar{A}\Gamma - \Gamma\bar{A}' - \sum_{j=1}^m P^{-1} B_j' P \Gamma P B_j P^{-1} = 0, \\ \Gamma(0) = (x - p(0))(x - p(0))', \end{cases} \tag{3.17}$$

where $\bar{A} = A - \sum_{j=1}^m B_j P^{-1} B_j' P$. This is a linear ODE. Solving it gives the optimal cost functional value to be $\frac{1}{2}\mathrm{tr}\,\Gamma(T)$.

Remark 3.4. The results obtained above are based on the LQ approach. LQ models constitute an extremely important class of optimal control problems and their optimal solutions can be obtained explicitly via the Riccati equations, due to the nice underlying structures (see [1, 4, 5, 10, 12, 22, 24]). The general SRE is introduced in [4] as a BSDE of the Pardoux–Peng type ([21]) for the case where all the coefficients are *random*. It reduces to (3.1) for the present case. Consequently, the results in this section can be extended to the case where the coefficients A, B_j are adapted random processes.

4. Solvability of SRE. In the previous section we derived explicitly an optimal control (in a feedback form) of the problem. However, there is one gap remaining, namely the results depend on the SRE (3.1) being solvable. The solvability of the SRE is by no means trivial and is interesting in its own right. In [4], a necessary and sufficient condition for the solvability of SREs more general than (3.1) is derived. However the condition there is rather implicit. This section gives an explicit condition which ensures that (3.1) admits a unique solution.

To this end, we first consider a *conventional* Riccati equation

$$\begin{cases} \dot{P} + PA + A'P - \sum_{j=1}^{m} PB_j K^{-1} B_j' P = 0, \\ P(T) = I, \end{cases} \tag{4.1}$$

where $K > 0$ is given a priori (compare (4.1) and (3.1)). Note that the above equation is a bit different from the standard one arising from *deterministic* control problems (see [1]) where $m = 1$. But the case $m > 1$ can be treated in the same way without any difficulty. In particular, (4.1) is associated with the deterministic LQ problem

$$\text{Minimize} \quad J(s, y; u(\cdot)) = \int_s^T \frac{1}{2} \sum_{j=1}^{m} u_j(t)' K(t) u_j(t) dt + \frac{1}{2} |x(T)|^2,$$

$$\text{Subject to} \quad \begin{cases} \dot{x}(t) = A(t)x(t) + \sum_{j=1}^{m} B_j(t) u_j(t), \\ x(s) = y, \end{cases}$$

where $(s, y) \in [0, T] \times R^n$. Namely, the value function of the above LQ problem is $\frac{1}{2} y' P(s) y$, where P is the solution to (4.1). Denote $\mathcal{K} = \{K \in L^\infty(0, T; \hat{S}_+^m) \mid K^{-1} \in L^\infty(0, T; \hat{S}_+^m)\}$. It can be checked that $C([0, T]; \hat{S}_+^m) \subset \mathcal{K}$. For each $K \in \mathcal{K}$, we know from the classical Riccati theory (as well as the remark above) that (4.1) admits a unique solution $P \in C([0, T]; S_+^n)$. Thus we can define a mapping $\Psi : \mathcal{K} \to C([0, T]; S^n)$ as $\Psi(K) = P$.

THEOREM 4.1. *The SRE* (3.1) *admits a unique solution if and only if there exists* $K \in C([0, T]; \hat{S}_+^m)$ *such that*

$$\Psi(K) \geq K. \tag{4.2}$$

Proof. This is a special case of [4, Theorem 4.2]. □

THEOREM 4.2. *If*

$$A(t) + A(t)' \geq \sum_{j=1}^{m} B_j(t) B_j(t)', \tag{4.3}$$

then the SRE (3.1) *admits a unique solution.*

Proof. We will show that (4.2) holds for $K = \varepsilon I$ with some $\varepsilon > 0$. To this end, for $\varepsilon > 0$ set $P_\varepsilon = \Psi(\varepsilon I) - \varepsilon I$. Then P_ε satisfies

$$\begin{cases} \dot{P}_\varepsilon + P_\varepsilon \left(A - \sum_{j=1}^m B_j B_j' \right) + \left(A - \sum_{j=1}^m B_j B_j' \right)' P_\varepsilon - \varepsilon^{-1} \sum_{j=1}^m P_\varepsilon B_j B_j' P_\varepsilon \\ \quad + \varepsilon \left(A + A' - \sum_{j=1}^m B_j B_j' \right) = 0, \\ P_\varepsilon(T) = I - \varepsilon I. \end{cases} \tag{4.4}$$

Therefore, under the assumption (4.3) and when $0 < \varepsilon < 1$, the above is a standard conventional Riccati equation which admits a unique solution $P_\varepsilon \geq 0$. This implies that (4.2) holds with $K = \varepsilon I$. The result follows then from Theorem 4.1. □

Remark 4.1. In [4], an algorithm of computing the solution to the SREs is given. Boiling down to the special SRE (3.1), the algorithm stipulates that one starts with $K = \varepsilon I$ (with $0 < \varepsilon < 1$) and solves the conventional Riccati equation (4.1) recursively. The resulting sequence of solutions will monotonically converge to the solution of SRE (3.1).

It should be noted that (4.3) only gives an (easily verifiable) *sufficient* condition for the solvability of SRE (3.1). In other special cases (see section 5 below), solvability of the SRE can also be shown without (4.3).

5. Black–Scholes model. We now apply the general results obtained in the previous sections to a contingent claim problem with the Black–Scholes setup. Suppose there is a market in which $m + 1$ assets (or securities) are traded continuously. One of the assets is the *bond* whose price process $P_0(t)$ is subject to the following (deterministic) ODE:

$$\begin{cases} dP_0(t) = r(t)P_0(t)dt, \quad t \in [0, T], \\ P_0(0) = p_0 > 0. \end{cases} \tag{5.1}$$

The other m assets are *stocks* whose price processes $P_1(t), \ldots, P_m(t)$ satisfy the following SDE:

$$\begin{cases} dP_i(t) = P_i(t) \left[b_i(t)dt + \sum_{j=1}^m \sigma_{ij}(t) dW^j(t) \right], \quad t \in [0, T], \\ P_i(0) = p_i > 0. \end{cases} \tag{5.2}$$

Define the *covariance matrix*

$$\sigma(t) = \begin{pmatrix} \sigma_1(t) \\ \vdots \\ \sigma_m(t) \end{pmatrix} \equiv (\sigma_{ij}(t))_{m \times m}. \tag{5.3}$$

The basic assumption throughout this section is

$$\sum(t) \equiv \sigma(t)\sigma(t)' \geq \delta I \quad \forall\, t \in [0, T] \tag{5.4}$$

for some $\delta > 0$. We also assume that all the given functions are measurable and uniformly bounded in t.

Consider an agent whose total wealth at time $t \geq 0$ is denoted by $x(t)$. Assume that the trading of shares takes place continuously and transaction cost and consump-

tions are not considered. Then $x(\cdot)$ satisfies (see, e.g., Karatzas and Shreve [13] and Elliott and Kopp [9])

$$\begin{cases} dx(t) &= \Big\{r(t)x(t) + \sum_{i=1}^m \big[b_i(t) - r(t)\big]\pi_i(t)\Big\}dt \\ &\quad + \sum_{j=1}^m \sum_{i=1}^m \sigma_{ij}(t)\pi_i(t)dW^j(t), \\ x(0) &= x > 0, \end{cases} \tag{5.5}$$

where $\pi_i(t)$, $i = 0, 1, 2 \ldots, m$, denotes the total market value of the agent's wealth in the ith bond/stock. We call $\pi(t) \equiv (\pi_1(t), \ldots, \pi_m(t))'$ a *portfolio* of the agent. Set

$$u(t) \equiv (u_1(t), \ldots, u_m(t))' = \sigma(t)'\pi(t), \text{ or } \pi(t) = \textstyle\sum(t)^{-1}\sigma(t)u(t). \tag{5.6}$$

On the other hand, due to (5.4), the model is *arbitrage free*, namely, there exists a *risk premium process* $\theta(\cdot)$ satisfying

$$\theta(t)\sigma(t)' = (b_1(t) - r(t), \ldots, b_m(t) - r(t)). \tag{5.7}$$

With the above notation (5.5) becomes

$$\begin{cases} dx(t) = \Big[r(t)x(t) + \sum_{j=1}^m \theta_j(t)u_j(t)\Big]\, dt + \sum_{j=1}^m u_j(t)dW^j(t), \\ x(0) = x. \end{cases} \tag{5.8}$$

The objective is, for each given initial wealth x and a contingent claim ξ (which is an $\mathcal{F}_T$-measurable square integrable random variable), to choose a (*mean-variance*) *hedging portfolio* $\pi(\cdot)$ (or, equivalently, a control $u(\cdot)$) so as to minimize

$$J(x, u(\cdot)) = \frac{1}{2}E|x(T) - \xi|^2. \tag{5.9}$$

Remark 5.1. Although the classical hedging problem is to make the difference on the right-hand side of (5.9) to be zero, it has become familiar to also call problem (5.9) a (mean-variance) hedging problem. The reason is at least two-fold. On the one hand, one may decompose the classical hedging problem into several problems of the type (5.9) with different initial conditions in order to characterize the price as well as the hedging portfolio for the original problem. This decomposition can be carried out based on an "extended asset method" recently introduced by Gourerioux, Laurent, and Pham [11] and extended by Laurent and Pham [15]. On the other hand, there is a growing interest in measuring risk when it is a priori known that with the initial capital x the contingent claim cannot be reached. In this case, the criterion (5.9) seems to be a viable alternative.

The problem (5.8)–(5.9) is a special case of the general model studied in section 4, so we can apply the results there. Interestingly, in this case the corresponding SRE is explicitly solvable due to the specific structure that the state variable is a *scalar* (and hence so is the solution to the SRE).

THEOREM 5.1. *The optimal portfolio of the hedging problem consisting of* (5.8) *and* (5.9) *is*

$$\pi^*(t) = -\textstyle\sum(t)^{-1}(b_1(t) - r(t), \ldots, b_m(t) - r(t))'[x^*(t) - p(t)] + \sum(t)^{-1}\sigma(t)q(t), \tag{5.10}$$

where $(p(\cdot), q(\cdot)) \equiv (p(\cdot), q_j(\cdot), j = 1, \ldots, m) \in L^2_{\mathcal{F}}(0, T; R^n) \times L^2_{\mathcal{F}}(0, T; R^{nm})$ *is the unique* $\mathcal{F}_t$*-adapted solution of the BSDE*

$$\begin{cases} dp(t) = \Big[r(t)p(t) + \sum_{j=1}^m \theta_j(t)q_j(t)\Big]\, dt + \sum_{j=1}^m q_j(t)dW^j(t), \\ p(T) = \xi. \end{cases} \tag{5.11}$$

Proof. The SRE (3.1) in the present case reduces to (noting that the unknown $P(t)$ of the equation is a scalar)

$$\begin{cases} \dot{P}(t) + 2r(t)P(t) - \sum_{j=1}^m \theta_j(t)^2 P(t) = 0, \\ P(T) = 1, \\ P(t) > 0, \ t \in [0,T]. \end{cases} \tag{5.12}$$

Denote $\rho(t) = \sum_{j=1}^m \theta_j(t)^2$. Then the above equation has a unique solution $P(t) = e^{-\int_t^T (\rho(s)-2r(s))ds}$. Note that the third inequality constraint in (5.12) is automatically satisfied by this solution. On the other hand, the associated equation (3.2) reads

$$\begin{cases} d\phi(t) = & -\Big\{[r(t)-\rho(t)]\phi(t) - \sum_{j=1}^m \theta_j(t)\beta_j(t) \\ & \quad + P(t)\left[g(t) + \sum_{j=1}^m \theta_j(t) z_j(t)\right]\Big\}dt \\ & \quad + \sum_{j=1}^m \beta_j(t) dW^j(t), \\ \phi(T) = & 0. \end{cases} \tag{5.13}$$

Applying Theorem 3.2 and noticing that $P(t)$ is now a scalar, we obtain

$$u_j^*(t) = -\theta_j(t)[x^*(t) - p(t)] + q_j(t). \tag{5.14}$$

Appealing to (5.6) and writing in a vector form, we obtain the desired result (5.10). □

Remark 5.2. The formula (5.10) has a straightforward interpretation in financial terms. Indeed, it is well known that the second term on the right-hand side of (5.10) is the *replicating portfolio* for the claim ξ when the initial wealth is the initial option price $p(0)$. The other term is exactly the *Merton portfolio* for a terminal utility function $c(x) = x^2$ (Merton [20]). Therefore, our optimal hedging policy (5.10) for our problem is the sum of the replicator for the claim and the Merton portfolio. Consequently, if the initial endowment x is different from the fair initial price $p(0)$ necessary to replicate the contingent claim ξ, then the difference $x - p(0)$ should be invested according to the Merton strategy.

Remark 5.3. In the Black–Scholes model, (3.17) reduces to

$$\begin{cases} \dot{\Gamma}(t) - 2r(t)\Gamma(t) + \sum_{j=1}^m \theta_j(t)^2 \Gamma(t) = 0, \\ \Gamma(0) = (x - p(0))^2. \end{cases} \tag{5.15}$$

This gives $\Gamma(t) = (x-p(0))^2 e^{-\int_0^t (\rho(s)-2r(s))ds}$. Hence the optimal value of (5.12) is $\frac{1}{2}\Gamma(T) = \frac{1}{2}P(0)(x-p(0))^2$.

Remark 5.4. By the explicit form of $P(\cdot)$, the solution to the SRE, as obtained in the proof of Theorem 5.1, we may understand $P(t)$ to be a sort of *normalizing* factor with respect to the "least-square" criterion in (5.9) as well as the discounting process in time. Moreover, by the first equality of (3.13), $\phi(t)$ is nothing but the *normalized* difference between the present expected value of the claim ξ and its present fair price.

6. A modified model. In the previous sections we investigated a model where only the terminal variance is to be minimized. It is more in line with the *European option* in the context of option theory where only the terminal situation is of interest.

To motivate the modified model we are going to formulate in this section let us consider the pricing problem of an American contingent claim ξ where the holder has the right to exercise the option at any time $\tau \in [0,T]$. This claim cannot be hedged by a self-financing portfolio, and it is necessary to introduce a (nondecreasing) consumption process $C(\cdot)$ in the price process

$$\begin{cases} dp(t) = \left[r(t)p(t) + \sum_{j=1}^m \theta_j(t)q_j(t)\right]dt + \sum_{j=1}^m q_j(t)dW^j(t) - dC(t), \\ p(T) = \xi \end{cases} \tag{6.1}$$

in order to push upwards the process $p(t)$ to keep it above the expected value of the claim $E(\xi|\mathcal{F}_t)$ at *any* time, while the amount of pushing should be minimal in the sense that

$$\int_0^T [p(t) - E(\xi|\mathcal{F}_t)]dC(t) = 0. \tag{6.2}$$

Pricing an American claim as a stopping time problem is extensively described in the literature (see, e.g., Karatzas and Shreve [13] and Elliott and Kopp [9]). Its equivalence to the existence problem of a solution $C(\cdot)$ to (6.1) can be found in El Karoui and Quenez [8].

In the modified mean-variance hedging problem below, the wealth $x(t)$ is to be kept near the expected value of the claim $E(\xi|\mathcal{F}_t)$ at *any* time $t \geq 0$ (rather than just at the terminal time). Whereas the solution to this modified problem does not really provide a solution to the original American contingent claim problem, in view of the formal analogies of the results below to those in pricing problems, we shall discuss in Remark 6.2 a possible interpretation of the modified model.

Motivated by the above, let us consider a modification of the model (2.1)–(2.2). Instead of cost functional (2.2), we consider

$$\tilde{J}(x,u(\cdot)) = \frac{1}{2}E\left[\int_0^T |x(t) - E(\xi|\mathcal{F}_t)|^2 dt + |x(T) - \xi|^2\right] \tag{6.3}$$

while keeping the same dynamics as in (2.1). (One can also put different weights on the running cost and the terminal cost, but we will not bother to do it here.)

Employing the same change of variable (2.3), we get the state equation (2.5) with the new cost functional

$$\tilde{J}(y,u(\cdot)) = \frac{1}{2}E\left[\int_0^T |y(t)|^2 dt + |y(T)|^2\right]. \tag{6.4}$$

To solve this problem, we only need to slightly modify the argument in section 3. Specifically, the SRE (3.1), for the present case, is changed to

$$\begin{cases} \dot{P} + PA + A'P - \sum_{j=1}^m PB_jP^{-1}B_j'P + I = 0, \\ P(T) = I, \\ P(t) > 0 \quad \forall\, t \in [0,T]. \end{cases} \tag{6.5}$$

The form of the associated equation (3.2) remains unchanged (but with the *new* $P(\cdot)$ in it as determined by (6.5)).

THEOREM 6.1. *If* (6.5) *and* (3.2) *admit solutions* $P \in C([0,T];\hat{S}^n_+)$ *and* $(\phi(\cdot),\beta_j(\cdot), j=1,\dots,m) \in L^2_{\mathcal{F}}(0,T;R^n)\times L^2_{\mathcal{F}}(0,T;R^{nm})$, *respectively, then the optimal control problem consisting of* (2.1) *and* (6.3) *has an optimal feedback control* $u^*(\cdot) \equiv (u_1^*(\cdot),\dots,u_m^*(\cdot))$, *where*

$$u_j^*(t) = -P(t)^{-1}B_j(t)'P(t)\big[x^*(t)-p(t)\big] + q_j(t),\ j=1,\dots,m, \tag{6.6}$$

where $(p(\cdot),q_j(\cdot),j=1,\dots,m) \in L^2_{\mathcal{F}}(0,T;R^n)\times L^2_{\mathcal{F}}(0,T;R^{nm})$ *is the unique* $\mathcal{F}_t$*-adapted solution of the BSDE*

$$\begin{cases} dp(t) = \Big[A(t)p(t) + \sum_{j=1}^m B_j(t)q_j(t) + f(t) - P(t)^{-2}\phi(t)\Big]\,dt \\ \qquad\quad + \sum_{j=1}^m q_j(t)dW^j(t),\ t\in[0,T], \\ p(T) = \xi, \end{cases} \tag{6.7}$$

or, equivalently,

$$\begin{cases} dp(t) = & \Big\{[A(t)+P(t)^{-1}]p(t) + \sum_{j=1}^m B_j(t)q_j(t) + f(t) \\ & -P(t)^{-1}E(\xi|\mathcal{F}_t)\Big\}dt \\ & + \sum_{j=1}^m q_j(t)dW^j(t),\ t\in[0,T], \\ p(T) = & \xi. \end{cases} \tag{6.8}$$

Proof. Consider the matrix-valued ODE

$$\begin{cases} \dot{Q} - AQ - QA' + \sum_{j=1}^m B_jQB_j' - Q^2 = 0, \\ Q(T) = I, \end{cases} \tag{6.9}$$

which is a conventional Riccati equation. Hence it admits a unique solution $Q(\cdot)$. Denote $S = PQ$, and then S satisfies

$$\begin{cases} \dot{S} = SA' - A'S + SQ - Q + \sum_{j=1}^m PB_jP^{-1}B_j'S - \sum_{j=1}^m PB_jP^{-1}SB_j', \\ S(T) = I. \end{cases} \tag{6.10}$$

This equation has the only solution $S\equiv I$, implying $Q(t) = P(t)^{-1}$. Now, performing the same change of variables in (3.13), we get that $(p(\cdot),q_j(\cdot),j=1,\dots,m)$ satisfies (6.7). Equation (6.8) is equivalent to (6.7) due to the fact that $P(t)^{-2}\phi(t) = P(t)^{-1}[E(\xi|\mathcal{F}_t) - p(t)]$ (see (3.13)). □

Remark 6.1. We note that in this case $(p(\cdot),q_j(\cdot),j=1,\dots,m)$ no longer satisfies the *original* BSDE (5.11) which is the starting point of the control problem under consideration in this paper. The reason is that the BSDE (5.11) only concerns the terminal situation but not any time between the present and the terminal times. Therefore a large deviation of $p(t)$ from the expected terminal value $E(\xi|\mathcal{F}_t)$ is allowed in the setup of (5.11). However, in our modified model, it is required that this deviation cannot be too large (which will be realized by the optimal control); therefore, in the optimal feedback control, one no longer compares against the original BSDE (5.11).

It is interesting that in this case the SRE (6.5) *automatically* admits a solution.

THEOREM 6.2. *The SRE* (6.5) *admits a unique solution.*

Proof. Employing the same argument as that in the proof of Theorem 4.2, set $P_\varepsilon = \Psi(\varepsilon I) - \varepsilon I$. Then P_ε satisfies

$$\begin{cases} \dot{P}_\varepsilon + P_\varepsilon \left(A - \sum_{j=1}^m B_j B_j'\right) + \left(A - \sum_{j=1}^m B_j B_j'\right)' P_\varepsilon \\ \quad -\varepsilon^{-1} \sum_{j=1}^m P_\varepsilon B_j B_j' P_\varepsilon \\ \quad +\varepsilon \left(A + A' - \sum_{j=1}^m B_j B_j'\right) + I = 0, \\ P_\varepsilon(T) = I - \varepsilon I. \end{cases} \tag{6.11}$$

When $\varepsilon > 0$ is small enough, $\varepsilon(A + A' - \sum_{j=1}^m B_j B_j') + I > 0$, hence the solution of the above equation (which is a conventional Riccati equation), $P_\varepsilon \geq 0$. This implies that (4.2) holds with $K = \varepsilon I$ for a sufficiently small $\varepsilon > 0$. The result follows then from Theorem 4.1. □

Now let us consider the corresponding Black–Scholes model. The SRE (5.12) is modified to

$$\begin{cases} \dot{P}(t) + 2r(t)P(t) - \sum_{j=1}^m \theta_j(t)^2 P(t) + 1 = 0, \\ P(T) = 1, \\ P(t) > 0,\ t \in [0, T]. \end{cases} \tag{6.12}$$

This equation has an explicit solution

$$P(t) = e^{-\int_t^T (\rho(s) - 2r(s))ds} + \int_t^T e^{-\int_t^s (\rho(\tau) - 2r(\tau))d\tau} ds > 0.$$

(The existence of solutions can also be concluded from Theorem 6.2.)

THEOREM 6.3. *The optimal (feedback) hedging portfolio for the modified Black–Scholes model is*

$$\pi^*(t) = -\sum(t)^{-1}(b_1(t) - r(t), \ldots, b_m(t) - r(t))'[x^*(t) - p(t)] + \sum(t)^{-1}\sigma(t)q(t), \tag{6.13}$$

where $(p(\cdot), q(\cdot)) \equiv (p(\cdot), q_j(\cdot), j = 1, \ldots, m) \in L^2_{\mathcal{F}}(0, T; R^n) \times L^2_{\mathcal{F}}(0, T; R^{nm})$ *is the unique* $\mathcal{F}_t$*-adapted solution of the following BSDE:*

$$\begin{cases} dp(t) = \left[r(t)p(t) + \sum_{j=1}^m \theta_j(t) q_j(t) + \frac{p(t) - E(\xi|\mathcal{F}_t)}{P(t)}\right] dt + \sum_{j=1}^m q_j(t) dW^j(t), \\ p(T) = \xi. \end{cases} \tag{6.14}$$

Proof. This follows immediately from Theorem 6.1 along with (5.6). □

Remark 6.2. Equation (6.14) is exactly in the same form as (6.1) with

$$C(t) = -\int_0^t \frac{p(s) - E(\xi|\mathcal{F}_s)}{P(s)} ds. \tag{6.15}$$

This process, $C(\cdot)$, plays a similar role to the cumulative consumption process in the original American contingent claim problem. It is no longer a nondecreasing process due to the underlying mean-variance criterion; however, the function of it is to correct any large deviation of the price $p(t)$ from the expected value of the claim $E(\xi|\mathcal{F}_t)$ at any time t. On the other hand, since in our model the price $p(t)$ is allowed to go

either above or under $E(\xi|\mathcal{F}_t)$, we do not have (6.2). Nevertheless, we do have the analogous relations

$$(6.16)\quad \int_0^T [p(t) - E(\xi|\mathcal{F}_t)]^+ [dC(t)]^- = 0, \quad \text{and} \quad \int_0^T [p(t) - E(\xi|\mathcal{F}_t)]^- [dC(t)]^+ = 0,$$

where $a^+ = \max\{a, 0\}$ and $a^- = \max\{-a, 0\}$. Therefore, the interpretation of the portfolio (6.13) is that one should hedge (in the mean-variance sense) the claim using the auxiliary process $C(\cdot)$ and invest the rest of the wealth according to the Merton portfolio.

7. Concluding remarks. In this paper we studied the relationship between BSDEs and stochastic control problems. We introduced a stochastic LQ model associated with a BSDE and solved the problem in a closed form by virtue of the stochastic LQ theory developed recently. The results were then applied to solve an optimal mean-variance hedging problem associated with a Black–Scholes contingent claim model. Our study suggested that the solution pair of a BSDE can be interpreted as the state-control pair of a stochastic control problem. This finding is expected to lead insights into the nature of the BSDEs as well as their applications in finance problems.

Acknowledgment. The authors would like to thank the two referees for their constructive comments that led to an improved version of the paper.

REFERENCES

[1] B. D. O. Anderson and J. B. Moore, *Optimal Control—Linear Quadratic Methods*, Prentice-Hall, Englewood Cliffs, NJ, 1989.
[2] J. M. Bismut, *Analyse Convexe et Probabilites*, These, Faculte des Sciences de Paris, Paris, 1973.
[3] J. M. Bismut, *An introductory approach to duality in optimal stochastic control*, SIAM Rev., 20 (1978), pp. 62–78.
[4] S. Chen, X. Li, and X. Y. Zhou, *Stochastic linear quadratic regulators with indefinite control weight costs*, SIAM J. Control Optim., 36 (1998), pp. 1685–1702.
[5] S. Chen and X. Y. Zhou, *Stochastic linear quadratic regulators with indefinite control weight costs* II, SIAM J. Control Optim., to appear.
[6] D. Duffie and L. Epstein, *Stochastic differential utility*, Econometrica, 60 (1992), pp. 353–394.
[7] N. El Karoui, S. Peng, and M. C. Quenez, *Backward stochastic differential equations in finance*, Math. Finance, 7 (1997), pp. 1–71.
[8] N. El Karoui and M. C. Quenez, *Nonlinear pricing theory and backward stochastic differential equations*, in Financial Mathematics (Bressanone, 1996), Lecture Notes in Math. 1656, Springer-Verlag, Berlin, 1997, pp. 191–246.
[9] R. J. Elliott and P. E. Kopp, *Mathematics of Financial Markets*, Springer-Verlag, New York, 1999.
[10] W. H. Fleming and H. M. Soner, *Controlled Markov Processes and Viscosity Solutions*, Springer-Verlag, New York, 1993.
[11] C. Gourerioux, J. P. Laurent, and H. Pham, *Mean variance hedging and numeraire*, Math. Finance, 8 (1998), pp. 179–200.
[12] R. E. Kalman, *Contributions to the theory of optimal control*, Bol. Soc. Mat. Mexicana, 5 (1960), pp. 102–119.
[13] I. Karatzas and S. E. Shreve, *Methods of Mathematical Finance*, Springer-Verlag, New York, 1998.
[14] H. Kunita and S. Watanabe, *On square integrable martingales*, Nagoya Math. J., 30 (1967), pp. 209–245.
[15] J. P. Laurent and H. Pham, *Dynamic programming and mean variance hedging*, Fin. Stoch., 3 (1999), pp. 83–110.

[16] J. Ma, P. Protter, and J. Yong, *Solving forward-backward stochastic differential equations explicitly—A four step scheme*, Probab. Theory Related Fields, 98 (1994), pp. 339–359.

[17] J. Ma and J. Yong, *Solvability of forward-backward SDEs and the nodal set of Hamilton–Jacobi–Bellman equations*, Chinese Ann. Math. Ser. B., 16 (1995), pp. 279–298.

[18] J. Ma and J. Yong, *Approximate Solvability of Forward-Backward Stochastic Differential Equations*, Tech. report, Department of Mathematics, Fudan University, Shanghai, China, 1999.

[19] J. Ma and J. Yong, *Forward-Backward Stochastic Differential Equations and Their Applications*, Lecture Notes in Math. 1702, Springer-Verlag, Berlin, 1999.

[20] R. Merton, *Optimum consumption and portfolio rules in a continuous time model*, J. Econom. Theory, 3 (1971), pp. 373–413; Erratum, 6 (1973), pp. 213–214.

[21] E. Pardoux and S. Peng, *Adapted solution of backward stochastic equation*, Systems Control Lett., 14 (1990), pp. 55–61.

[22] W. M. Wonham, *On a matrix Riccati equation of stochastic control*, SIAM J. Control, 6 (1968), pp. 681–697.

[23] J. Yong, *Linear forward-backward stochastic differential equations*, Appl. Math. Optim., 39 (1999), pp. 93–119.

[24] J. Yong and X. Y. Zhou, *Stochastic Controls*: *Hamiltonian Systems and HJB Equations*, Springer, New York, 1999.

SIAM J. CONTROL OPTIM.
Vol. 38, No. 5, pp. 1408–1427

SIMULTANEOUS EXACT CONTROLLABILITY AND SOME APPLICATIONS*

MARIUS TUCSNAK† AND GEORGE WEISS‡

Abstract. We study the exact controllability of two systems by means of a common finite-dimensional input function, a property called simultaneous exact controllability. Most of the time we consider one system to be infinite-dimensional and the other finite-dimensional. In this case we show that if both systems are exactly controllable in time T_0 and the generators have no common eigenvalues, then they are simultaneously exactly controllable in any time $T > T_0$. Moreover, we show that similar results hold for approximate controllability. For exactly controllable systems we characterize the reachable subspaces corresponding to input functions of class H^1 and H^2. We apply our results to prove the exact controllability of a coupled system composed of a string with a mass at one end. Finally, we consider an example of two infinite-dimensional systems: we characterize the simultaneously reachable subspace for two strings controlled from a common end. The result is obtained using a recent generalization of a classical inequality of Ingham.

Key words. linear system, operator semigroup, admissible control operator, Gramian, exact controllability, exact observability, simultaneous controllability, wave equation, boundary control, coupled system

AMS subject classifications. 93B28, 93C25, 93B03, 93C20

PII. S0363012999352716

1. Introduction.

We consider two control systems (possibly infinite-dimensional), with the states denoted by z_1, z_2, described by the equations

$$\begin{cases} \dot{z}_1(t) = A_1 z_1(t) + B_1 u(t), & z_1(0) = 0, \\ \dot{z}_2(t) = A_2 z_2(t) + B_2 u(t), & z_2(0) = 0. \end{cases} \tag{1.1}$$

Here, a dot denotes differentiation with respect to the time t, A_1, A_2 are generators of strongly continuous operator semigroups on the corresponding state spaces, and B_1, B_2 are admissible control operators for these semigroups. Note that the two systems receive the same input function u. These systems are called *simultaneously exactly controllable in time* T (where $T > 0$), if for any states f_1 and f_2, an L^2-function u can be found such that $z_1(T) = f_1$ and $z_2(T) = f_2$.

Simultaneous exact controllability was first considered by Russell in [22] and it is the subject of Chapter 5 in Lions [20]. The simultaneous controllability of two Riesz spectral systems (one hyperbolic and one parabolic) was studied in section 4 of Hansen [10] (see also Hansen and Zhang [12]). We were led to investigate simultaneous exact controllability in our study of coupled systems (sometimes called hybrid systems), such as a string with a mass at one end, or the SCOLE model of a beam clamped at one end and with a rigid body at the other end.

Our main result (proved in section 3) concerns the situation where one system is finite-dimensional. We show that, in this case, if A_1 and A_2 have no common eigenvalues and if both are exactly controllable in time T_0, then they are simultaneously

*Received by the editors March 1, 1999; accepted for publication (in revised form) October 4, 1999; published electronically May 11, 2000.
http://www.siam.org/journals/sicon/38-5/35271.html
†Department of Mathematics, University of Nancy-I, POB 239, Vandoeuvre les Nancy 54506, France (marius.tucsnak@iecn.u-nancy.fr).
‡Department of Electrical and Electronic Engineering, Imperial College of Science, Technology and Medicine, Exhibition Road, London SW7 2BT, UK (g.weiss@ic.ac.uk).

exactly controllable in any time $T > T_0$. For $T = T_0$ this is not always true, as we show in an example (see section 4).

The concept of simultaneous approximate controllability of two systems in time T is similar to the controllability concept defined earlier, but now the reachable pairs of states (f_1, f_2) must be dense in the product of the respective state spaces. Considering again one system to be finite-dimensional, we have a result that resembles our main result, but now we have no information on the time T needed for simultaneous approximate controllability: we only know that some $T > 0$ will work. Other results in section 3 concern the characterization of the reachable subspace of an exactly controllable system, when the input function u is constrained to be in the Sobolev space H^1 (or H^2) with $u(0) = 0$ (or with $u(0) = \dot{u}(0) = 0$).

In section 4 we give two applications to systems governed by partial differential equations (PDEs), both based on the (nonhomogeneous) one-dimensional wave equation. These two interdependent examples illustrate how simultaneous controllability results can be applied in the analysis of coupled systems. In section 5 we characterize the simultaneously reachable subspace of two systems describing vibrating strings. The results here are based on recent generalizations of an inequality of Ingham.

2. Some background on infinite-dimensional systems. In this section we gather, for easy reference, some basic facts about admissible control and observation operators, controllability, and observability. Some results here are new, but most are well known. For the latter, we do not give proofs; we only refer to the relevant literature.

We assume that X is a Hilbert space and $A : \mathcal{D}(A) \to X$ is the generator of a strongly continuous semigroup $\mathbb{T}$ on X. We define the Hilbert space X_1 as $\mathcal{D}(A)$ with the norm $\|z\|_1 = \|(\beta I - A)z\|$, where $\beta \in \rho(A)$ is fixed (this norm is equivalent to the graph norm). The Hilbert space X_{-1} is the completion of X with respect to the norm $\|z\|_{-1} = \|(\beta I - A)^{-1}z\|$. This space is isomorphic to $\mathcal{D}(A^*)^*$, and we have

$$X_1 \subset X \subset X_{-1}, \tag{2.1}$$

densely and with continuous embeddings. $\mathbb{T}$ extends to a semigroup on X_{-1}, denoted by the same symbol. The generator of this extended semigroup is an extension of A, whose domain is X, so that $A : X \to X_{-1}$.

We assume that U is a Hilbert space and $B \in \mathcal{L}(U, X_{-1})$ is an *admissible control operator* for $\mathbb{T}$, defined as in Weiss [24]. This means that if z is the solution of

$$\dot{z}(t) = Az(t) + Bu(t) \tag{2.2}$$

(an equation in X_{-1}), with $z(0) = z_0 \in X$ and $u \in L^2([0, \infty), U)$, then $z(t) \in X$ $\forall t \geq 0$. In this case, z is a continuous X-valued function of t. We have

$$z(t) = \mathbb{T}_t z_0 + \Phi_t u, \tag{2.3}$$

where $\Phi_t \in \mathcal{L}(L^2([0, \infty), U), X)$ is defined by

$$\Phi_t u = \int_0^t \mathbb{T}_{t-\sigma} B u(\sigma) d\sigma. \tag{2.4}$$

The above integration is done in X_{-1}, but the result is in X. The Laplace transform of z is

$$\hat{z}(s) = (sI - A)^{-1} \left[z_0 + B\hat{u}(s) \right].$$

B is called *bounded* if $B \in \mathcal{L}(U, X)$ (and unbounded otherwise).

We assume that Y is another Hilbert space and $C \in \mathcal{L}(X_1, Y)$ is an *admissible observation operator* for $\mathbb{T}$, defined as in Weiss [25]. This means that for every $T > 0$ there exists a $K_T \geq 0$ such that

$$\int_0^T \|C\mathbb{T}_t z_0\|^2 dt \leq K_T^2 \|z_0\|^2 \qquad \forall z_0 \in \mathcal{D}(A). \tag{2.5}$$

C is called *bounded* if it can be extended such that $C \in \mathcal{L}(X, Y)$.

We regard $L^2_{loc}([0,\infty), Y)$ as a Fréchet space with the seminorms being the L^2 norms on the intervals $[0, n]$, $n \in \mathbb{N}$. Then the admissibility of C means that there is a continuous operator $\Psi : X \to L^2_{loc}([0,\infty), Y)$ such that

$$(\Psi z_0)(t) = C\mathbb{T}_t z_0 \qquad \forall z_0 \in \mathcal{D}(A). \tag{2.6}$$

The operator Ψ is completely determined by (2.6), because $\mathcal{D}(A)$ is dense in X. We introduce the Λ-*extension* of C, denoted C_Λ, by

$$C_\Lambda z_0 = \lim_{\lambda \to +\infty} C\lambda(\lambda I - A)^{-1} z_0, \tag{2.7}$$

whose domain $\mathcal{D}(C_\Lambda)$ consists of all $z_0 \in X$ for which the limit exists. If we replace C by C_Λ, formula (2.6) becomes true $\forall z_0 \in X$ and for almost every $t \geq 0$. For $z_0 \in \mathcal{D}(A)$, Ψz_0 is almost everywhere (a.e.) differentiable and

$$\frac{d}{dt}\left(C\mathbb{T}_t z_0\right) = C_\Lambda \mathbb{T}_t A z_0 \qquad \text{for almost every } t \geq 0. \tag{2.8}$$

If $y = \Psi z_0$, then its Laplace transform is

$$\hat{y}(s) = C(sI - A)^{-1} z_0. \tag{2.9}$$

If $\mathbb{T}$ is exponentially stable, then $\Psi \in \mathcal{L}(X, L^2([0,\infty), Y))$.

The following duality result holds: if $\mathbb{T}$ is a semigroup on X with generator A, then $B \in \mathcal{L}(U, X_{-1})$ is an admissible control operator for $\mathbb{T}$ if and only if $B^* : \mathcal{D}(A^*) \to U$ is an admissible observation operator for the dual semigroup $\mathbb{T}^*$. Moreover, the adjoint of Φ_T from (2.4) is given by

$$(\Phi_T^* z_0)(t) = B_\Lambda^* \mathbb{T}_{T-t}^* z_0 \tag{2.10}$$

for almost every $t \in [0, T]$, where $B_\Lambda^* z = \lim_{\lambda \to +\infty} \lambda B^*(\lambda I - A^*)^{-1} z$, as in (2.7). For all the facts listed so far in this section, we refer to [24], [25], and [26].

For $C, \mathbb{T}$ as in (2.5) and for every $T > 0$, we introduce the bounded operator $\Psi_T : X \to L^2([0, T], Y)$ by truncating Ψ to $[0, T]$, i.e., $\forall t \in [0, T]$,

$$\left(\Psi_T z_0\right)(t) = C\mathbb{T}_t z_0 \qquad \forall z_0 \in \mathcal{D}(A). \tag{2.11}$$

The *observability Gramians* of (A, C) are the operators

$$P_T = \Psi_T^* \Psi_T \qquad \forall T \geq 0.$$

Thus, for $z_0 \in \mathcal{D}(A)$,

$$P_T z_0 = \int_0^T \mathbb{T}_t^* C^* C \mathbb{T}_t z_0 dt,$$

and, to get an expression valid $\forall z_o \in X$, we may replace C by C_Λ in the above formula. If $\mathbb{T}$ is exponentially stable, then we may also take $T = \infty$, defining the Gramian $P = \Psi^*\Psi$, which satisfies $A^*P + PA = -C^*C$. For more on Gramians we refer to Hansen and Weiss [11] or Russell and Weiss [23].

DEFINITION 2.1. *With the notation as in* (2.11) *the pair* (A, C) *is exactly observable in time* T *if* Ψ_T *is bounded from below, i.e., there exists* $k_T > 0$ *such that*

$$\int_0^T \|C\mathbb{T}_t z_0\|_Y^2 dt \geq k_T^2 \|z_0\|_X^2 \qquad \forall z_0 \in \mathcal{D}(A). \tag{2.12}$$

The pair (A, C) *is approximately observable in time* T *if* $\operatorname{Ker} \Psi_T = \{0\}$.

As is well known, for finite-dimensional systems the properties in Definition 2.1 are equivalent and independent of T, and if they hold, then we say that (A, C) is *observable*. We remark that $\int_0^T \|C\mathbb{T}_t z_0\|_Y^2 > 0$ $\forall z_0 \in \mathcal{D}(A)$ is not sufficient for approximate observability in time T.

Clearly, the following assertions hold true.

PROPOSITION 2.2. *The pair* (A, C) *is exactly observable in time* T *if and only if* P_T *is invertible. Similarly,* (A, C) *is approximately observable in time* T *if and only if* P_T *is one-to-one. If* $T > \tau$, *then* $P_T \geq P_\tau$.

With the notation from (2.11) it is easy to see that if $z_0 \in \mathcal{D}(A)$, then $\Psi_T z_0 \in H^1(0, T; Y)$. The following partial converse will be needed in section 3.

PROPOSITION 2.3. *With the notation as in* (2.11), *suppose that* (A, C) *is exactly observable in time* T_0. *If* $z_0 \in X$ *and* $T > T_0$ *are such that* $\Psi_T z_0 \in H^1(0, T; Y)$, *then* $z_0 \in \mathcal{D}(A)$. *For* $T = T_0$, *the implication is not true in general.*

Proof. Denote $y = \Psi_T z_0$, so that $y \in H^1(0, T; Y)$. Using, for example, Proposition VIII.3 (p. 124) in Brezis [6], we obtain

$$\sup_{\varepsilon \in (0, T-T_0)} \int_0^{T_0} \left\| \frac{y(t+\varepsilon) - y(t)}{\varepsilon} \right\|_Y^2 dt < \infty.$$

Since, for almost every $t \in [0, T_0]$, $y(t+\varepsilon) - y(t) = C_\Lambda \mathbb{T}_t (\mathbb{T}_\varepsilon - I) z_0$, it follows that

$$\sup_{\varepsilon \in (0, T-T_0)} \left\| \Psi_{T_0} \frac{\mathbb{T}_\varepsilon - I}{\varepsilon} z_0 \right\|_{L^2([0,T_0],Y)} < \infty.$$

Because of the exact observability estimate (2.12), this implies

$$\sup_{\varepsilon \in (0, T-T_0)} \left\| \frac{\mathbb{T}_\varepsilon - I}{\varepsilon} z_0 \right\|_X < \infty.$$

By a simple result on operator semigroups, see for instance Theorem 2.12 (p. 88) in Butzer and Berens [7], it follows that $z_0 \in \mathcal{D}(A)$. To see that for $T = T_0$ the implication is false, consider the left-shift semigroup $\mathbb{T}$ on $X = L^2[0, 1]$ with point observation at the left end. Thus $A = \frac{d}{d\xi}$, $\mathcal{D}(A) = \{x \in H^1(0,1) | x(1) = 0\}$, and $Cx = x(0)$. This system is exactly observable in time $T_0 = 1$. However, if $z_0(\xi) = 1$ $\forall \xi \in (0, 1)$, then $\Psi_1 z_0 \in H^1(0, 1)$, but $z_0 \notin \mathcal{D}(A)$. □

DEFINITION 2.4. *Let* A *be the generator of a strongly continuous semigroup* $\mathbb{T}$ *on* X *and let* $B \in \mathcal{L}(U, X_{-1})$ *be an admissible control operator for* $\mathbb{T}$. *The pair* (A, B) *is exactly controllable in time* $T > 0$, *if for every* $f_0 \in X$ *there exists a* $u \in L^2([0, T], U)$ *such that*

$$\int_0^T \mathbb{T}_{T-\sigma} B u(\sigma) d\sigma = f_0.$$

(A, B) is approximately controllable in time T if the set of those f_0 for which the above property holds is dense.

In other words, we say that (A, B) is exactly controllable in time T if Φ_T is onto, i.e., Ran $\Phi_T = X$, and (A, B) is approximately controllable in time T if Ran Φ_T is dense in X. For finite-dimensional systems the above properties are equivalent and independent of T, and if they hold we say that (A, B) is *controllable*.

PROPOSITION 2.5. *We assume that A is the generator of a semigroup $\mathbb{T}$ on X and $B \in \mathcal{L}(U, X_{-1})$ is an admissible control operator for $\mathbb{T}$. Then (A, B) is exactly controllable in time T if and only if (A^*, B^*) is exactly observable in time T. Similarly, (A, B) is approximately controllable in time T if and only if (A^*, B^*) is approximately observable in time T.*

This proposition is an easy consequence of (2.10). It is used frequently in the literature on control of systems governed by PDEs (see, e.g., the HUM method of Lions [20]). For more details on exact controllability (observability) in a functional-analytic setting we refer to Avdonin and Ivanov [2] or [23] and the references therein. In the PDE's-setting, the relevant literature is overwhelming, and we mention the books of Lions [20], Lagnese and Lions [16], and Komornik [21] and the paper of Bardos, Lebeau, and Rauch [5].

3. Main results. First we give the definition of the simultaneous controllability concepts used.

DEFINITION 3.1. *For $j \in \{1, 2\}$, let A_j be the generators of the strongly continuous semigroups $\mathbb{T}^j$ acting on the Hilbert spaces X^j. Let U be a Hilbert space and let $B_j \in \mathcal{L}(U, X^j_{-1})$ be admissible control operators for $\mathbb{T}^j$.*

The pairs (A_j, B_j) are called simultaneously exactly controllable in time $T > 0$ if for every state $f_j \in X^j$ there exists a function $u \in L^2([0, T], U)$ such that

$$\int_0^T \mathbb{T}^j_{T-\sigma} B_j u(\sigma) d\sigma = f_j.$$

The same pairs are called simultaneously approximately controllable in time $T > 0$ if the property described above holds for (f_1, f_2) in a dense subspace of $X^1 \times X^2$.

It is clear that the concepts introduced in the last definition are equivalent to the exact (approximate) controllability in time T of the pair

$$A = \begin{bmatrix} A_1 & 0 \\ 0 & A_2 \end{bmatrix}, \qquad B = \begin{bmatrix} B_1 \\ B_2 \end{bmatrix}.$$

Using Proposition 2.5, the above concepts can be characterized by duality.

PROPOSITION 3.2. *With the notation of Definition 3.1, we have:*

1. *The pairs (A_1, B_1) and (A_2, B_2) are simulteanously exactly controllable in time T if and only if there exists $k_T > 0$ such that $\forall (z_0^1, z_0^2) \in \mathcal{D}(A_1^*) \times \mathcal{D}(A_2^*)$ we have*

$$\int_0^T \|B_1^* \mathbb{T}_t^{1*} z_0^1 + B_2^* \mathbb{T}_t^{2*} z_0^2\|_U^2 \geq k_T^2 \left(\|z_0^1\|_{X^1}^2 + \|z_0^2\|_{X^2}^2 \right). \tag{3.1}$$

2. *The pairs (A_1, B_1) and (A_2, B_2) are simultaneously approximately controllable in time T if and only if the following statement holds.*
   If $(z_0^1, z_0^2) \in X^1 \times X^2$ are such that

$$B_{1\Lambda}^* \mathbb{T}_t^{1*} z_0^1 + B_{2\Lambda}^* \mathbb{T}_t^{2*} z_0^2 = 0 \qquad \text{for almost every } t \in [0, T], \tag{3.2}$$

then $(z_0^1, z_0^2) = (0, 0)$.

We mention that in (3.2) we must use the Λ-extensions as in (2.10). The reason is that it is not possible to use only $(z_0^1, z_0^2) \in \mathcal{D}(A_1^*) \times \mathcal{D}(A_2^*)$ (this follows from the comments after Definition 2.1).

The main result of this section is the following theorem.

THEOREM 3.3. *Let A be the generator of the strongly continuous semigroup $\mathbb{T}$ acting on the Hilbert space X. Let $B \in \mathcal{L}(\mathbb{C}^m, X)$ be an admissible control operator for $\mathbb{T}$ and assume that (A, B) is exactly controllable in time T_0. Let $a \in \mathbb{C}^{n\times n}$ and $b \in \mathbb{C}^{n\times m}$ be matrices such that (a, b) is controllable. Assume that A and a have no common eigenvalues. Then the pairs (A, B) and (a, b) are simultaneously exactly controllable in any time $T > T_0$.*

First we prove the following approximate controllability result.

LEMMA 3.4. *Suppose that $T > T_0$ and that (A, B), (a, b) satisfy the assumptions of Theorem* 3.3. *Then these two pairs are simultaneously approximately controllable in time T for every $T > T_0$.*

Proof. Let $T > T_0$ be fixed. Denote by V the set of all $v_0 \in \mathbb{C}^n$ such that there exists a $z_0 \in X$ with

$$B_\Lambda^* \mathbb{T}_t^* z_0 + b^* e^{a^* t} v_0 = 0 \qquad \text{for almost every } t \in [0, T]. \tag{3.3}$$

Using the approximate controllability of (A, B) in time T_0 and Proposition 2.5, we see that the function $t \to B_\Lambda^* \mathbb{T}_t^* z_0$, $t \in [0, T]$, determines z_0. By (3.3), this function is determined by v_0. Thus, if $v_0 \in V$, then z_0 satisfying (3.3) is unique and depends linearly on v_0: $z_0 = Qv_0$. Since the function $t \to b^* e^{a^* t} v_0$ is smooth, by Proposition 2.3 we have that

$$Qv_0 \in \mathcal{D}(A^*) \qquad \forall v_0 \in V.$$

Now we show that $\forall v_0 \in V$, we have

$$Qa^* v_0 = A^* Q v_0. \tag{3.4}$$

Indeed, by differentiating (3.3) with respect to time and using (2.8), we obtain that

$$B_\Lambda^* \mathbb{T}_t^* A^* Q v_0 + b^* e^{a^* t} a^* v_0 = 0 \tag{3.5}$$

for almost every $t \in [0, T]$, which shows that $a^* v_0 \in V$ and (3.4) holds.

Let $\widetilde{a}$ denote the restriction of a^* to its invariant subspace V. If $V \neq \{0\}$, then $\widetilde{a}$ must have an eigenvalue $\lambda \in \sigma(a^*)$ and a corresponding eigenvector $\widetilde{v}$. Formula (3.4) implies that $A^* Q\widetilde{v} = \lambda Q\widetilde{v}$. Since Q is one-to-one, we have that $Q\widetilde{v} \neq 0$, so that λ is an eigenvalue of A^*. This is in contradiction to the assumption in Theorem 3.3, and hence we must have $V = \{0\}$. Thus, (3.3) implies that $(z_0, v_0) = (0, 0)$ and we can apply the second part of Proposition 3.2. ∎

Proof of Theorem 3.3. Let $T > T_0$ be fixed. According to Proposition 2.5 it suffices to show that the pair

$$\mathcal{A}^* = \begin{bmatrix} A^* & 0 \\ 0 & a^* \end{bmatrix}, \qquad \mathcal{B}^* = \begin{bmatrix} B^* & b^* \end{bmatrix} \tag{3.6}$$

is exactly observable in time T. We already know from Lemma 3.4 and Proposition 2.5 that $(\mathcal{A}^*, \mathcal{B}^*)$ is approximately observable in time T. Let $\mathcal{P}_T$ denote the observability

Gramian of $(\mathcal{A}^*, \mathcal{B}^*)$, so that $\mathcal{P}_T > 0$. We partition $\mathcal{P}_T$ in a natural way, according to the product space $X \times \mathbb{C}^n$:

$$\mathcal{P}_T = \begin{bmatrix} P_T & L \\ L^* & p_T \end{bmatrix}.$$

We want to show that $\mathcal{P}_T$ is invertible (i.e., bounded from below). It is not difficult to see that P_T is the observability Gramian of (A^*, B^*) and p_T is the observability Gramian of (a^*, b^*). As (A^*, B^*) and (a^*, b^*) are exactly observable in time T, by Proposition 2.2, both P_T and p_T are positive and boundedly invertible. We bring in the Schur-type factorization

$$\begin{bmatrix} P_T & L \\ L^* & p_T \end{bmatrix} = \begin{bmatrix} P_T & 0 \\ L^* & I \end{bmatrix} \begin{bmatrix} P_T^{-1} & 0 \\ 0 & \Delta \end{bmatrix} \begin{bmatrix} P_T & L \\ 0 & I \end{bmatrix},$$

where $\Delta = p_T - L^* P_T^{-1} L$ (this is checked by multiplying out). Notice that the first factor is the adjoint of the last, and they are invertible. Therefore, $\mathcal{P}_T$ is invertible if and only if the middle factor is invertible. Since P_T^{-1} is obviously bounded from below, we see that $\mathcal{P}_T$ is bounded from below if and only if Δ is bounded from below. Since $\mathcal{P}_T > 0$, from the factorization we see that $\Delta > 0$. But Δ is a matrix, so that $\Delta > 0$ implies that Δ is invertible. Thus we have proved that $\mathcal{P}_T$ is invertible. By Proposition 2.2, $(\mathcal{A}^*, \mathcal{B}^*)$ is exactly observable in time T. □

Remark 3.5. Under the assumptions of Theorem 3.3, in general, the two systems will not be simultaneously exactly controllable in time T_0. An example to illustrate this will be given in section 4.

In the rest of this section we shall investigate simultaneous approximate controllability. With the assumptions of Theorem 3.3 we obviously obtain simultaneous approximate controllability, but the result is not sharp as it asks for exact controllability of each component. We give below a simultaneous approximate controllability result by supposing only approximate controllability of each component.

At this point we introduce some notation. Let A be the generator of a strongly continuous semigroup. Then the resolvent set $\rho(A)$ contains a right half-plane. The resolvent set is not necessarily connected, and we denote by $\rho_\infty(A)$ the connected component of $\rho(A)$ which contains some right half-plane. (Obviously, there is only one such component.) In particular, if $\sigma(A)$ is countable, as is often the case in applications, then $\rho_\infty(A) = \rho(A)$.

PROPOSITION 3.6. *Let A be the generator of the strongly continuous semigroup $\mathbb{T}$ acting on the Hilbert space X. Let $B \in \mathcal{L}(\mathbb{C}^m, X_{-1})$ be an admissible control operator for $\mathbb{T}$ and assume that (A, B) is approximately controllable in time T_0. Let $a \in \mathbb{C}^{n\times n}$ and $b \in \mathbb{C}^{n\times m}$ be matrices such that (a, b) is controllable. Further, assume that*

$$\sigma(a) \subset \rho_\infty(A). \tag{3.7}$$

Then there exists $T > 0$ such that the pairs (A, B) and (a, b) are simultaneously approximately controllable in time T.

Proof. To arrive at a contradiction, we assume that the opposite holds: $(\mathcal{A}, \mathcal{B})$ from (3.6) is not approximately controllable in any time. Then it follows from Proposition 3.2 that for every $k \in \mathbb{N}$ there exists a $z_k \in X$ and a $v_k \in \mathbb{C}^n$ such that $(z_k, v_k) \neq (0, 0)$ and

$$B_\Lambda^* \mathbb{T}_t^* z_k + b^* e^{a^* t} v_k = 0 \qquad \forall\, t \in [0, k]. \tag{3.8}$$

It follows from the approximate observability in time T_0 of (A^*, B^*) that $\forall k > T_0$ we must have $v_k \neq 0$. Hence we may assume without loss of generality that $\|v_k\|_{\mathbb{C}^n} = 1$. By the compactness of the unit ball in $\mathbb{C}^n$, we may assume further that the sequence (v_k) is convergent: $\lim v_k = v_0$. Then it follows that if we define the functions $y_k \in L^2_{loc}([0,\infty), \mathbb{C}^m)$ by

$$y_k(t) = b^* e^{a^* t} v_k \qquad \text{for } k \in \{0, 1, 2, \dots\},$$

then $\lim y_k = y_0$ (in L^2_{loc}). Let Ψ_{T_0} be the operator defined by

$$\Psi_{T_0} z_0 = B^*_\Lambda \mathbb{T}^*_t z_0 \qquad \forall t \in [0, T_0],$$

and let Π_{T_0} denote the truncation of a function defined on $[0,\infty)$ to $[0, T_0]$. Then (3.8) implies that

$$\Psi_{T_0} z_k + \Pi_{T_0} y_k = 0 \qquad \forall k \geq T_0.$$

Since Ker $\Psi_{T_0} = \{0\}$, the above equation shows that z_k is uniquely determined by y_k, which in turn is obtained from v_k. All these dependencies are linear, so that there is an operator $R : \mathbb{C}^n \to X$ (possibly nonunique, depending on the span of all v_k) such that $z_k = Rv_k \ \forall k \in \mathbb{N}$. Hence, the sequence (z_k) is convergent, and we put $z_0 = \lim z_k = Rv_0$. Now it is easy to conclude from (3.8) that

$$(\Psi z_0)(t) + b^* e^{a^* t} v_0 = 0 \qquad \text{for almost every } \ t \geq 0.$$

Taking Laplace transforms, we obtain from the last formula that for some $\alpha \in \mathbb{R}$ and every $s \in \mathbb{C}$ with Re $s > \alpha$,

$$B^*(sI - A^*)^{-1} z_0 + b^*(sI - a^*)^{-1} v_0 = 0. \tag{3.9}$$

By analytic continuation, this formula remains valid on $\rho_\infty(A^*) \setminus \sigma(a^*)$. (On the other connected components of $\rho(A^*)$ we have no such information.) Since $v_0 \neq 0$ (actually, its norm is 1) and (a^*, b^*) is observable, the rational function $b^*(sI - a^*)^{-1} v_0$ is not zero. Therefore it has poles at a nonempty subset of $\sigma(a^*)$, which by (3.7) is contained in $\rho_\infty(A^*)$. The first term in (3.9) being analytic around $\sigma(a^*)$, it follows that the left-hand side of (3.9) has poles, which is absurd. Thus we have proved that $(\mathcal{A}, \mathcal{B})$ must be approximately controllable in some time T. □

Note that the lemma says nothing about the time T in which $(\mathcal{A}, \mathcal{B})$ is approximately controllable. If T_0 is minimal for (A, B), then of course $T \geq T_0$.

In the last part of this section we characterize the reachable subspaces of an exactly controllable system, when the input function is restricted to Sobolev type spaces strictly included in L^2.

Let A be the generator of a strongly continuous semigroup $\mathbb{T}$ on X and let $B \in \mathcal{L}(U, X_{-1})$ be an admissible control operator for $\mathbb{T}$. Suppose that the pair (A, B) is *exactly controllable* in time T, in the sense of Definition 2.4. This means that the range of the operator Φ_T defined by (2.4) is equal to X. A natural question is the characterization of the states which can be reached by more regular inputs. Define

$$H^1_L(0, T; U) = \{\psi \in H^1(0, T; U) \mid \psi(0) = 0\}.$$

The existence and uniqueness result below shows that the space reachable by means of controls in $H^1_L(0, T; U)$ cannot be larger than the space Z defined by

$$Z = X_1 + (\beta I - A)^{-1} BU = (\beta I - A)^{-1}(X + BU), \tag{3.10}$$

where $\beta \in \rho(A)$ (Z does not depend on the choice of β). The norm on Z is defined by

$$\|z\|_Z^2 = \inf\left\{\|x\|^2 + \|u\|^2 \mid x \in X,\, u \in U,\, z = (\beta I - A)^{-1}(x + Bu)\right\}.$$

LEMMA 3.7. *For any $u \in H^1_L(0,T;U)$, the solution z of (2.2) with $z(0) = 0$ is such that*

$$z \in C(0,T;Z) \cap C^1(0,T,X).$$

Proof. Let $u \in H^1_L(0,T;U)$ and denote by w the solution of

$$\dot{w} = Aw + B\dot{u}, \qquad w(0) = 0.$$

As B is an admissible control operator we have that $w \in C([0,T];X)$. Moreover it is easily checked that the function $t \to \int_0^T w(s)ds$ satisfies (2.2). Since the solution of (2.2) with $z(0) = 0$ is unique, we obtain

$$z(t) = \int_0^T w(s)ds,$$

which obviously yields that

$$z \in C^1([0,T],X). \tag{3.11}$$

On the other hand (2.2) gives

$$(\beta I - A)z(t) = \beta z(t) - \dot{z}(t) + Bu(t) \qquad \forall t \in [0,T]. \tag{3.12}$$

Since $\beta z - \dot{z} + Bu \in C([0,T], X + BU)$, relation (3.12) with $\beta \in \rho(A)$ implies

$$z \in C([0,T],Z). \tag{3.13}$$

From (3.11) and (3.13) we clearly obtain the conclusion of the lemma. □

We can now characterize the states which are reachable by means of input functions in $H^1_L(0,T;U)$ as follows.

PROPOSITION 3.8. *Suppose that the pair (A,B) is exactly controllable in time T_0. Then $\forall T > T_0$, the reachable space by means of input functions $u \in H^1_L(0,T;U)$ is the space Z from* (3.10).

Proof. We know from Lemma 3.7 that the reachable space is included in Z. To show that Z is contained in the reachable space, take $\beta \in \rho(A)$ and consider two systems with states w and v and input u_1, described by

$$\dot{w} = (A - \beta I)w + Bu_1, \tag{3.14}$$

$$\dot{v} = u_1. \tag{3.15}$$

For an arbitrary $z^0 \in Z$ choose $w^0 \in X,\ v^0 \in U$ such that

$$z^0 = (\beta I - A)^{-1}[w^0 - Bv^0]. \tag{3.16}$$

Since 0 is not an eigenvalue of $A - \beta I$, by Theorem 3.3 the systems (3.14) and (3.15) are simultaneously exactly controllable in any time $T > T_0$. Hence we can find $u_1 \in L^2([0,T];U)$ such that the solutions $w,\ v$ of (3.14) and (3.15) satisfy

$$w(0) = 0, \qquad w(T) = e^{-\beta T}w^0, \qquad v(0) = 0, \qquad v(T) = e^{-\beta T}v^0. \tag{3.17}$$

We define the function z_1 by

$$z_1(t) = (\beta I - A)^{-1}(w(t) - Bv(t)) \qquad \forall t \in [0,T].$$

Then it is easy to see that

$$z_1(0) = 0, \qquad z_1(T) = e^{-\beta T} z^0. \tag{3.18}$$

Moreover, after a simple calculation, (3.14) and (3.15) imply that

$$\dot{z}_1(T) = -w(t) = (A - \beta I) z_1(t) - Bv(t) \qquad \forall t \in (0,T). \tag{3.19}$$

If we define now

$$z(t) = e^{\beta t} z_1(t), \qquad u(t) = e^{\beta t} v(t),$$

relations (3.18) and (3.19) imply that z and u satisfy (2.2) together with $z(0) = 0$ and $z(T) = z^0$. This means that Z is included in the space reachable by means of input functions $u \in H^1_L(0,T;U)$, as claimed. □

4. Applications.

4.1. Applications to the equation of a vibrating string. In this subsection we apply the results obtained in previous sections to the equation of a nonhomogeneous vibrating string. First we show that, with suitably chosen spaces, the system corresponding to the string equation and an integrator are simultaneously exactly controllable. In the case of a homogeneous string we show that the simultaneous exact controllability time is strictly larger than the exact controllability time for the string alone, i.e., we give the counterexample announced in Remark 3.5. In the second part of this subsection we characterize the space of the states which are reachable by means of an H^1 or H^2 input function u with $u(0) = 0$ and, in the case $u \in H^2$, also $\dot{u}(0) = 0$.

Let us consider the initial and boundary value problem

$$\begin{cases} \ddot{w}(x,t) = [m(x) w_x(x,t)]_x, & 0 < x < 1, \\ w(0,t) = 0, & w(1,t) = u(t), \\ w(x,0) = 0, & \dot{w}(x,0) = 0 \end{cases} \tag{4.1}$$

with

$$m \in W^{1,\infty}(0,1), \qquad m(x) \geq m_0 > 0 \qquad \forall x \in (0,1). \tag{4.2}$$

The equations above represent the simplest model of a nonhomogeneous elastic string. Following well-known ideas (see for instance Lasiecka and Triggiani [18], [19]) the system (4.1) can be written in the abstract form (2.2), provided we use the notation

$$z = \begin{bmatrix} w \\ \dot{w} \end{bmatrix}, \qquad X = L^2[0,1] \times H^{-1}(0,1), \qquad U = \mathbb{C},$$

$$A = \begin{bmatrix} 0 & I \\ A_0 & 0 \end{bmatrix}, \qquad B = \begin{bmatrix} 0 \\ -A_0 D \end{bmatrix}, \tag{4.3}$$

where

$$\mathcal{D}(A_0) = H_0^1(0,1), \qquad A_0 : \mathcal{D}(A_0) \to H^{-1}(0,1), \qquad A_0 h = (m(x)h_x)_x,$$

so that $A_0 < 0$, and the Dirichlet map $D : \mathbb{C} \to L^2[0,1]$ is defined by

$$D\alpha = y \Longleftrightarrow \{(m(x)y_x)_x = 0 \text{ in } (0,1),\ y(0) = 0,\ y(1) = \alpha\}$$

(see also [1]). From the above it clearly follows that $A : \mathcal{D}(A) \to X$, with

$$\mathcal{D}(A) = H_0^1(0,1) \times L^2[0,1],$$

and that A is skew-adjoint: $A^* = -A$. Note that $B^* = \begin{bmatrix} 0 & D^* \end{bmatrix}$ and, for every $h \in H^2(0,1) \cap H_0^1(0,1)$, $D^* A_0 h = m(1)h_x(1)$. We denote by $\mathbb{T}$ the semigroup generated by A. Well-known computations, using the above expressions for A^* and B^* (see again [18], [19]) give that

$$B^* \mathbb{T}_t^* \begin{bmatrix} z^0 \\ z^1 \end{bmatrix} = m(1)\phi_x(1,t) \qquad \forall \begin{bmatrix} z^0 \\ z^1 \end{bmatrix} \in \mathcal{D}(A), \tag{4.4}$$

where ϕ solves the corresponding homogeneous problem

$$\ddot{\phi}(x,t) = (m(x)\phi_x(x,t))_x, \qquad 0 < x < 1,\ t \in (0,T), \tag{4.5}$$

$$\phi(0,t) = \phi(1,t) = 0, \qquad t \in [0,T], \tag{4.6}$$

$$\phi(\cdot,0) = \phi^0 = A_0^{-1} z^1 \in H^2(0,1) \cap H_0^1(0,1), \tag{4.7}$$

$$\dot{\phi}(\cdot,0) = \phi^1 = z^0 \in H_0^1(0,1). \tag{4.8}$$

It is by now well known that B is an admissible control operator and the couple (A,B) is exactly controllable in any time $T > T_0$, where $T_0 = \frac{2}{\sqrt{m_0}}$ (see for instance Zuazua [27]). Moreover, if $m = 1$, then the system (A,B) is exactly controllable in time 2 (see for instance Haraux [15]).

Consider now the following system of two scalar differential equations with the same input u:

$$\begin{cases} \dot{v} = u, \\ \dot{w} = w + u. \end{cases} \tag{4.9}$$

The result below, concerning the simultaneous exact controllability of (4.1) and (4.9), gives, in particular, the counterexample announced in Remark 3.5.

PROPOSITION 4.1. *The systems* (4.1) *and* (4.9) *are simultaneously exactly controllable in any time $T > T_0$, where $T_0 = \frac{2}{\sqrt{m_0}}$. However, if $m = 1$, then the systems* (4.1) *and* (4.9) *are not simultaneously approximately controllable in time $T_0 = 2$.*

Proof. We can write the system (4.9) in the form $\dot{q} = aq + bu$, where

$$q = \begin{bmatrix} v \\ w \end{bmatrix}, \qquad a = \begin{bmatrix} 0 & 0 \\ 0 & 1 \end{bmatrix}, \qquad b = \begin{bmatrix} 1 \\ 1 \end{bmatrix}, \tag{4.10}$$

and it is clear that (a, b) is controllable. The eigenvalues of A from (4.3) are on the imaginary axis and nonzero. The simultaneous exact controllability in any time $T > T_0$ follows from the exact controllability of the system in (4.1) in any time $T > T_0$, by applying Theorem 3.3.

We still have to prove the lack of simultaneous approximate controllability in time 2, in the case of a homogeneous string with $m = 1$. Choose $w_0 \in \mathbb{R}$, $w_0 \neq 0$. As the family formed by $(\sin(n\pi t)_{n\geq 1}, \cos(n\pi t)_{n\geq 1})$ together with the constant function $1/\sqrt{2}$ is an orthonormal basis in $L^2(0,2)$, we can find sequences $(a_n)_{n\geq 1}$ and $(b_n)_{n\geq 1}$ in l^2 and $v_0 \in \mathbb{R}$ such that

$$\sum_{n=1}^{\infty}(-1)^n[a_n \cos(n\pi t) + b_n \sin(n\pi t)] + v_0 + e^t w_0 = 0 \qquad \text{for a.e. } t \in [0,2]. \tag{4.11}$$

Note that the functions $\sin(n\pi x)$ $(x \in (0,1))$ are eigenvectors of A_0. If we denote

$$z^0(x) = \pi \sum_{n=1}^{\infty} b_n \sin(n\pi x), \qquad z^1(x) = \pi^2 \sum_{n=1}^{\infty} n a_n \sin(n\pi x),$$

then $z^0 \in L^2[0,1]$ and $z^1 \in H^{-1}(0,1)$. Now using (4.4) and (4.10), relation (4.11) can be written as

$$B_\Lambda^* \mathbb{T}_t^* \begin{bmatrix} z^0 \\ z^1 \end{bmatrix} + b^* e^{a^* t} \begin{bmatrix} v_0 \\ w_0 \end{bmatrix} = 0 \qquad \text{for almost every } t \in [0,2].$$

Since $w_0 \neq 0$, this relation with Proposition 3.2 implies that the systems (4.1) and (4.9) are not simultaneously approximately controllable in time $T_0 = 2$. □

For $l > 0$ we define the space

$$H_L^2(0,l) = \{u \in H^2(0,l) \mid u(0) = \dot{u}(0) = 0\}.$$

The states of the system (4.1) which can be reached by means of H_L^1 and H_L^2 input functions can be characterized as follows.

PROPOSITION 4.2. *Suppose that $m(x)$ satisfies (4.2) and $T > T_0 = \frac{2}{\sqrt{m_0}}$. Then the space of all states $(w(T), \dot{w}(T))$ which can be reached in time T by means of input functions $u \in H_L^1(0,T)$ is $Z = H_L^1(0,1) \times L^2[0,1]$.*

Moreover, the space of all states $(w(T), \dot{w}(T))$ which can be reached in time T by means of input functions $u \in H_L^2(0,T)$ is $Z_1 = [H_L^1(0,1) \cap H^2(0,1)] \times H_L^1(0,1)$.

Proof. For $u \in H_L^1(0,T)$ it suffices to apply Proposition 3.8 and to notice that, with the notation (4.3), the space Z defined by (3.10) is $H_L^1(0,1) \times L^2[0,1]$. For $u \in H_L^2(0,T)$ we consider the new input $\widetilde{u} = \dot{u}$, a new state space equal to $H_L^1(0,1) \times L^2[0,1]$, and we apply again Proposition 3.8. □

4.2. Controllability of a coupled system. Consider a vertical string whose horizontal displacement in a given plane is described by the wave equation on the spatial domain $(0,1)$. The upper end (corresponding to $x = 0$) is kept fixed and an object of mass M is attached at the lower end (corresponding to $x = 1$). The external input is a horizontal force v acting on the object, and it is contained in the plane mentioned earlier. We neglect the moment of inertia of the object (i.e., we imagine the object to be very small). From simple physical considerations, and taking a certain constant to be one, we obtain that this system is described by the following equations,

valid $\forall x \in (0,1)$ and $\forall t \in (0,\infty)$:

$$\begin{cases} \ddot{w}(x,t) = [m(x)w_x]_x(x,t), w(0,t) = 0, \\ M\ddot{w}(1,t) - w_x(1,t) = v(t), \\ w(x,0) = \dot{w}(x,0) = 0, \ x \in (0,1). \end{cases} \tag{4.12}$$

Here, w is the controlled wave (horizontal displacement) and $\dot{w}$ is the horizontal velocity. The appropriate spaces for all these functions will be specified later. The point $x = 0$ is just reflecting waves, while the active end $x = 1$ is where both the observation and the control take place. We shall often write $w(t)$ to denote a function of x, meaning that $w(t)(x) = w(x,t)$, and similarly for other functions.

A direct analysis of the well-posedness, controllability, and observability of this system is not trivial, in spite of the simplicity of the system. We will show below that we can obtain a sharp result by simply applying Proposition 4.2. We begin by identifying the natural state space of (4.12).

PROPOSITION 4.3. *Suppose that $m(\cdot)$ satisfies* (4.2) *and that $v \in L^2[0,T]$. Then the initial and boundary value problem* (4.12) *admits a unique solution*

$$w \in C(0,T; H^1_L(0,1) \cap H^2(0,1)) \cap C^1(0,T; H^1_L(0,1)). \tag{4.13}$$

Proof. Using semigroups or a standard Galerkin method, it is easy to prove that $\forall v \in L^2[0,T]$, the problem (4.12) admits a unique solution

$$w \in C(0,T; H^1_L(0,1)) \cap C^1(0,T; L^2[0,1]), \tag{4.14}$$

which satisfies the first equation from (4.12) in $\mathcal{D}'((0,1)\times(0,T))$ and the second in $\mathcal{D}'(0,T)$ (notice that $w_x(1,\cdot)$ makes sense in $H^{-2}(0,T)$). Consider a sequence (v_n) in $\mathcal{D}(0,T)$ such that $v_n \to v$ in $L^2[0,T]$. If we denote by (w_n) the corresponding sequence of smooth solutions of (4.12), it is clear that

$$w_n \to w \ \text{ in } L^\infty(0,T; H^1_L(0,1)) \cap W^{1,\infty}(0,T; L^2[0,1]), \tag{4.15}$$

$$w_n(1,t) = \dot{w}_n(1,t) = 0 \qquad \forall n \geq 1. \tag{4.16}$$

Moreover, by multiplying the equation

$$(\ddot{w}_m - \ddot{w}_n)(x,t) = [m(x)(w_m - w_n)_x]_x(x,t)$$

by $x\frac{\partial}{\partial x}(w_m - w_n)(x,t)$ and by integrating over $[0,1]\times[0,T]$, we obtain, after well-known calculations, the existence of a constant $C > 0$ such that

$$\int_0^T |(w_m - w_n)_x(1,t)|^2 \, dt$$

$$\leq C\left(\|w_n - w_m\|_{L^\infty(0,T;H^1(0,1))} + \|\dot{w}_n - \dot{w}_m\|_{L^\infty(0,T;L^2[0,1])}\right). \tag{4.17}$$

Since

$$M\ddot{w}_n(1,t) - (w_n)_x(1,t) = v_n(t),$$

relation (4.17) implies that $\ddot{w}_n(1,\cdot)$ is a Cauchy sequence in $L^2[0,T]$. By using (4.15) and (4.16), we obtain that $w(1,\cdot) \in H_L^2(0,T)$. The regularity (4.13) follows now from Proposition 4.2. □

PROPOSITION 4.4. *Suppose that m satisfies* (4.2) *and $T > T_0 = \frac{2}{\sqrt{m_0}}$. Then the system* (4.12) *is well posed and exactly controllable in time T in the space $X = [H_L^1(0,1) \cap H^2(0,1)] \times H_L^1(0,1)$. In other words, $(w^0, w^1) \in [H_L^1(0,1) \cap H^2(0,1)] \times H_L^1(0,1)$ if and only if there exists $v \in L^2[0,T]$ such that the solution of* (4.12) *satisfies*

$$w(T) = w^0, \qquad \dot{w}(T) = w^1. \tag{4.18}$$

Proof. By Proposition 4.2, for any $(w^0, w^1) \in [H_L^1(0,1) \cap H^2(0,1)] \times H_L^1(0,1)$ there exist

$$w \in C(0,T; H^2(0,1)), \qquad u \in H_L^2(0,T) \tag{4.19}$$

satisfying (4.1) and (4.18). From (4.19) it obviously follows that if we define

$$v(t) = m\ddot{u}(t) - w_x(1,t),$$

then $v \in L^2[0,T]$ and w, v satisfy (4.12) and (4.18). □

5. The simultaneously reachable subspace of two infinite-dimensional systems. In this section we study an example showing that for certain pairs of infinite-dimensional systems it is still possible to derive results similar to those obtained in the previous section. However, the reachable space and the reachability time are more difficult to characterize. The problem we tackle is the one-dimensional version of an open question raised in Lions [20]. We give here only the results which are simple consequences of recent work on nonharmonic Fourier series. A detailed study of this problem requires new techniques and is the subject of the forthcoming paper by Avdonin and Tucsnak [3].

For $\xi \in (0,1)$ we consider the problems

$$\begin{cases} \ddot{w}_1(x,t) - (w_1(x,t))_{xx} = 0 \quad \forall x \in (0,\xi), \quad \forall t \in (0,\infty), \\ w_1(0,t) = 0, \ w_1(\xi,t) = u(t) \quad \forall t \in (0,\infty), \\ w_1(x,0) = 0, \ \dot{w}_1(x,0) = 0 \quad \forall x \in (0,\xi) \end{cases} \tag{5.1}$$

and

$$\begin{cases} \ddot{w}_2(x,t) - (w_2(x,t))_{xx} = 0 \quad \forall x \in (\xi,1), \quad \forall t \in (0,\infty), \\ w_2(1,t) = 0, \ w_2(\xi,t) = u(t) \quad \forall t \in (0,\infty), \\ w_2(x,0) = 0, \ \dot{w}_2(x,0) = 0 \quad \forall x \in (\xi,1). \end{cases} \tag{5.2}$$

The systems above model the vibrations of two strings joined at a common end at $x = \xi$, the input being the displacement of this common point.

By using notation similar to the one used in (4.3), we can easily define the operators (A_i, B_i), $i = 1, 2$ such that the equations (5.1), (5.2) can be written as in (1.1), with state spaces $X^1 = L^2[0,\xi] \times H^{-1}(0,\xi)$ and $X^2 = L^2[\xi,1] \times H^{-1}(\xi,1)$. According to classical results, B_1 (resp., B_2) is an admissible control operator and the system (A_1, B_1) (resp., (A_2, B_2)) is exactly controllable in time 2ξ (resp., $2(1-\xi)$). The aim of this section is to describe, to some extent, the space of the states in $X^1 \times X^2$ which are reachable by means of an input function $u \in L^2[0,T]$, with sufficiently large T.

We cannot give a precise characterization of this reachable space but we give sharp embedding results in appropriate Sobolev spaces.

For $s > -\frac{1}{2}$, we introduce the space $\mathcal{W}_s \subset X^1 \times X^2$ of quadruples of functions $(w_1^0, w_1^1, w_2^0, w_2^1)$ satisfying

$$(w_1^0, w_1^1, w_2^0, w_2^1) \in H^{s+1}(0,\xi) \times H^s(0,\xi) \times H^{s+1}(\xi,1) \times H^s(\xi,1),$$

$$w_1^0(0) = 0, \qquad w_2^0(1) = 0, \qquad w_1^0(\xi) = w_2^0(\xi).$$

Denote by $\mathbb{Q}$ the set of rational numbers. We denote by $\mathcal{S}$ the set of all numbers $\rho \in (0,1)$ such that $\rho \notin \mathbb{Q}$ and if $[0, a_1, \dots, a_n, \dots]$ is the expansion of ρ as a continuous fraction, then (a_n) is bounded. Note that $\mathcal{S}$ is uncountable and, by classical results on diophantine approximation (cf. [8, p. 120]), its Lebesgue measure is zero. Roughly speaking, the set $\mathcal{S}$ contains the irrationals which are "badly" approximable by rational numbers. In particular, by the Euler–Lagrange theorem (cf. [17, p. 57]) $\mathcal{S}$ contains all $\xi \in (0,1)$ such that ξ is an irrational quadratic number (i.e., satisfying a second degree equation with rational coefficients). According to a classical result (see, for instance, [17]), if $\xi \in \mathcal{S}$, then there exists a constant $C_\xi > 0$ such that

$$\left|\xi - \frac{p}{q}\right| \geq \frac{C_\xi}{q^2} \qquad \forall p, q \in \mathbb{N}. \tag{5.3}$$

We can now state our main result concerning the lack of simultaneous exact controllability of the two strings, which also gives some information on the simultaneously reachable space as a function of ξ.

THEOREM 5.1. *Suppose that $T > \max\{4\xi, 4(1-\xi)\}$. Then the following holds.*

(a) *For any $\xi \in \mathcal{S}$, all the elements of $\mathcal{W}_0$ can be reached in time T by means of an input $u \in L^2[0,T]$.*

(b) *For almost all $\xi \in [0,1]$ and $\forall s > 0$, all the states in $\mathcal{W}_s$ can be reached in time T by means of an input $u \in L^2[0,T]$.*

(c) *The results above are sharp in the sense that, for any $\xi \in (0,1)$ and $s \in (-\frac{1}{2}, 0)$, we can find a state in $\mathcal{W}_s$ which is not reachable by means of an input $u \in L^2[0,T]$. In particular, for any $T > 0$, the systems (5.1), (5.2) are not simultaneously exactly controllable in time T (in the natural energy space $X^1 \times X^2$).*

As a tool in our proof, $\forall s > -\frac{1}{2}$ we introduce the space

$$\mathcal{V}_s = H_0^{s+1}(0,\xi) \times H^s(0,\xi) \times H_0^{s+1}(\xi,1) \times H^s(\xi,1).$$

It is clear that $\mathcal{V}_s$ is a subspace of W_s (with finite codimension). In order to prove Theorem 5.1, we notice first that for $s < \frac{1}{2}$, the reachability of $\mathcal{W}_s$ is equivalent to the reachability of its subspace $\mathcal{V}_s$. More precisely, we have the following lemma.

LEMMA 5.2. *Let $s \in (-\frac{1}{2}, \frac{1}{2})$. Then all the elements of $\mathcal{W}_s$ can be reached in time T by means of an input $u \in L^2[0,T]$ if and only if the same property holds for $\mathcal{V}_s$.*

Proof. One of the implications is trivial. Take $(w_1^0, w_1^1, w_2^0, w_2^1) \in \mathcal{W}_s$ for some fixed $s \in (-\frac{1}{2}, \frac{1}{2})$ and denote $\alpha = w_1^0(\xi) = w_2^0(\xi)$. Let $\psi_1(x,t)$, $\psi_2(x,t)$ be the solutions of (5.1), (5.2) with $u = u_\psi$, where

$$u_\psi(t) = \frac{\alpha}{T^2} t^2.$$

It can be checked, arguing similarly as in the proof of Lemma 3.7, but differentiating twice, that

$$(\psi_1, \dot\psi_1, \psi_2, \dot\psi_2) \in C([0,T]; \mathcal{W}_1).$$

In particular, this implies that the above statement is true with $\mathcal{W}_s$ in place of $\mathcal{W}_1$. Moreover, we have

$$\psi_1(0,T) = \psi_2(1,T) = 0, \qquad \psi_1(\xi,T) = \psi_2(\xi,T) = \alpha.$$

The above equalities (together with $s < \frac{1}{2}$) imply that

$$(w_1^0 - \psi_1(\cdot,T), w_1^1 - \dot\psi_1(\cdot,T), w_2^0 - \psi_2(\cdot,T), w_2^1 - \dot\psi_2(\cdot,T)) \in \mathcal{V}_s.$$

Suppose now that all the elements of $\mathcal{V}_s$ can be reached in time T by means of an input in $L^2[0,T]$. It follows that there exists an input $u_\varphi \in L^2[0,T]$ such that the solutions φ_1, φ_2 of (5.1) and (5.2) with $u = u_\varphi$ satisfy the conditions

$$\varphi_1(x,T) = w_1^0(x) - \psi_1(x,T),\ \dot\varphi_1(x,T) = w_1^1(x) - \dot\psi_1(x,T), \quad \text{in } L^2[0,\xi], \tag{5.4}$$

$$\varphi_2(x,T) = w_2^0(x) - \psi_2(x,T),\ \dot\varphi_2(x,T) = w_2^1(x) - \dot\psi_2(x,T), \quad \text{in } L^2[\xi,1]. \tag{5.5}$$

If we define the input $u \in L^2[0,T]$ by $u = u_\psi + u_\varphi$, then the corresponding solutions w_1 and w_2 of (5.1), (5.2) satisfy

$$w_1(x,T) = w_1^0(x),\ \dot w_1(x,T) = w_1^1(x),\ w_2(x,T) = w_2^0(x),\ \dot w_2(x,T) = w_2^1(x).$$

Thus, the elements of $\mathcal{W}_s$ can be reached in time T by an input $u \in L^2[0,T]$. □

The main tool used in the proof of Theorem 5.1 is a recent generalization of a classical inequality of Ingham. This result was first proved in Jaffard, Tucsnak, and Zuazua [14] for $T > \frac{12\sqrt{6}}{\delta}$ and then improved in Baiochi, Komornik, and Loreti [4] for $T > \frac{4\pi}{\delta}$. Its statement (following [4]) is the following theorem.

THEOREM 5.3. *Let $M > 0$ and let (λ_n) be a strictly increasing real sequence over $\mathbb{Z}$ satisfying*

$$\lambda_{n+2} - \lambda_n \geq \delta > 0 \qquad \forall n \in \mathbb{Z} \ \text{ with } \ |n| \geq M. \tag{5.6}$$

Then $\forall T > \dfrac{4\pi}{\delta}$ there exist constants $C_1, C_2 > 0$ such that

$$C_1 \sum \left[\left(|a_n|^2 + |a_{n+1}|^2\right) |\lambda_{n+1} - \lambda_n|^2 + |a_n + a_{n+1}|^2 \right] \leq \int_0^T \left| \sum a_n e^{i\lambda_n t} \right|^2 dt$$

$$\leq C_2 \sum \left[\left(|a_n|^2 + |a_{n+1}|^2\right) |\lambda_{n+1} - \lambda_n|^2 + |a_n + a_{n+1}|^2 \right] \quad \forall (a_n) \in l^2.$$

Let us now consider the initial and boundary value problems

$$\ddot\phi_1(x,t) - \frac{\partial^2 \phi_1}{\partial x^2}(x,t) = 0 \qquad \forall x \in (0,\xi) \qquad \forall t \in (0,\infty), \tag{5.7}$$

$$\phi_1(0,t)=\phi_1(\xi,t)=0 \qquad \forall t\in(0,\infty), \tag{5.8}$$

$$\phi_1(x,0)=\phi_1^0(x), \qquad \dot\phi_1(x,0)=\phi_1^1(x) \qquad \forall x\in(0,\xi), \tag{5.9}$$

and

$$\ddot\phi_2(x,t)-\frac{\partial^2\phi_2}{\partial x^2}(x,t)=0 \qquad \forall x\in(\xi,1) \qquad \forall t\in(0,\infty), \tag{5.10}$$

$$\phi_2(1,t)=\phi_2(\xi,t)=0 \qquad \forall t\in(0,\infty), \tag{5.11}$$

$$\phi_2(x,0)=\phi_2^0(x), \quad \dot\phi_2(x,0)=\phi_2^1(x) \qquad \forall x\in(\xi,1). \tag{5.12}$$

We will use the following duality result, which is related to Proposition 3.2. This result follows from Theorem 2.1 in Dolecki and Russell [9] or from the HUM method of Lions (see [20]).

LEMMA 5.4. *The space of the states of* (5.1), (5.2) *which can be reached by means of the same input* $u\in L^2[0,T]$ *contains the space* $\mathcal{V}_s$, $s\in(-\frac12,\frac12)$ *if and only if there exist* $C,T>0$ *such that the solutions* ϕ_1,ϕ_2 *of* (5.7)–(5.12) *satisfy*

$$\int_0^T\left|\frac{\partial\phi_2}{\partial x}(\xi,t)-\frac{\partial\phi_1}{\partial x}(\xi,t)\right|^2dt$$

$$\geq\left(\|\phi_1^0\|^2_{H^{-s}(0,\xi)}+\|\phi_1^1\|^2_{H^{-1-s}(0,\xi)}+\|\phi_2^0\|^2_{H^{-s}(\xi,1)}+\|\phi_2^1\|^2_{H^{-1-s}(0,\xi)}\right)$$

$$\forall(\phi_1^0,\phi_1^1,\phi_2^0,\phi_2^1)\in\big(H^2(0,\xi)\cap H_0^1(0,\xi)\big)\times H_0^1(0,\xi)\times\big(H^2(\xi,1)\cap H_0^1(\xi,1)\big)\times H_0^1(\xi,1).$$

Proof of Theorem 5.1. If $\phi_1^0\in H^2(0,\xi)\cap H_0^1(0,\xi)$, $\phi_1^1\in H_0^1(0,\xi)$, $\phi_2^0\in H^2(\xi,1)\cap H_0^1(\xi,1)$, $\phi_2^1\in H_0^1(\xi,1)$, it is known that we have the expansions

$$\left.\begin{array}{ll}\phi_1^0(x)= & \sum_{n\geq1}c_n\sin\left(\frac{n\pi x}{\xi}\right)\\ \phi_1^1(x)= & \frac{\pi}{\xi}\sum_{n\geq1}nd_n\sin\left(\frac{n\pi x}{\xi}\right)\end{array}\right\}\ x\in(0,\xi),$$

$$\left.\begin{array}{ll}\phi_2^0(x)= & \sum_{n\geq1}e_n\sin\left(\frac{n\pi(1-x)}{1-\xi}\right)\\ \phi_2^1(x)= & \frac{\pi}{1-\xi}\sum_{n\geq1}nf_n\sin\left(\frac{n\pi(1-x)}{1-\xi}\right)\end{array}\right\}\ x\in(\xi,1),$$

where the sequences (n^2c_n), (n^2d_n), (n^2e_n), and (n^2f_n) are in l^2. A standard calculation shows that the solutions ϕ_1,ϕ_2 of (5.7)–(5.12) are given by

$$\phi_1(x,t)=\sum_{n\in\mathbb{Z}}a_ne^{i\frac{n\pi}{\xi}t}\sin\left(\frac{n\pi x}{\xi}\right), \qquad x\in(0,\xi), \tag{5.13}$$

$$\phi_2(x,t)=\sum_{n\in\mathbb{Z}}b_ne^{i\frac{n\pi}{1-\xi}t}\sin\left(\frac{n\pi(1-x)}{1-\xi}\right), \qquad x\in(\xi,1), \tag{5.14}$$

where

$$a_n = \begin{cases} \frac{c_n - id_n}{2} & \text{for } n \geq 1, \\ \frac{c_{-n} + id_{-n}}{2} & \text{for } n \leq -1, \\ 0 & \text{for } n = 0, \end{cases} \tag{5.15}$$

$$b_n = \begin{cases} \frac{e_n - if_n}{2} & \text{for } n \geq 1, \\ \frac{e_{-n} + if_{-n}}{2} & \text{for } n \leq -1, \\ 0 & \text{for } n = 0. \end{cases} \tag{5.16}$$

If we denote by $(\lambda_n)_{n\in\mathbb{Z}}$ the strictly increasing sequence formed by the elements of the set

$$\Lambda = \left[\cup_{n\in\mathbb{Z}} \left\{\frac{n\pi}{\xi}\right\}\right] \bigcup \left[\cup_{n\in\mathbb{Z}} \left\{\frac{n\pi}{1-\xi}\right\}\right],$$

we can easily check that

$$\lambda_{n+2} - \lambda_n \geq in\left\{\frac{\pi}{\xi}, \frac{\pi}{1-\xi}\right\} \qquad \forall n \in \mathbb{Z}. \tag{5.17}$$

On the other hand, from (5.3) it easily follows (see [13] for details) that, $\forall \xi \in \mathcal{S}$, there exists a constant $C_\xi > 0$ with

$$\lambda_{n+1} - \lambda_n \geq \frac{C_\xi}{|\lambda_n|} \qquad \forall n \in \mathbb{Z}^*, \tag{5.18}$$

where $\mathbb{Z}^* = \mathbb{Z} \setminus \{0\}$. Moreover (5.13), (5.14) imply

$$\frac{\partial \phi_2}{\partial x}(\xi, t) - \frac{\partial \phi_1}{\partial x}(\xi, t) = \sum_{n\in\mathbb{Z}^*} (-1)^{n+1} n\pi \left(\frac{a_n}{\xi} e^{i\frac{n\pi t}{\xi}} + \frac{b_n}{1-\xi} e^{i\frac{n\pi t}{1-\xi}}\right), \tag{5.19}$$

which yields

$$\frac{\partial \phi_2}{\partial x}(\xi, t) - \frac{\partial \phi_1}{\partial x}(\xi, t) = \sum_{n\in\mathbb{Z}^*} k_n \lambda_n e^{i\lambda_n t}, \tag{5.20}$$

with the sequence (k_n) satisfying

$$\sum_{n\in\mathbb{Z}^*} |k_n|^2 = \sum_{n\in\mathbb{Z}^*} (|a_n|^2 + |b_n|^2). \tag{5.21}$$

Relations (5.18), (5.20), (5.21), and Theorem 5.3 imply that there exists a constant $K_\xi > 0$ such that

$$\int_0^T \left|\frac{\partial \phi_2}{\partial x}(\xi, t) - \frac{\partial \phi_1}{\partial x}(\xi, t)\right|^2 dt \geq K_\xi \sum_{n\in\mathbb{Z}} \left(|a_n|^2 + |b_n|^2\right) \tag{5.22}$$

$\forall \xi \in \mathcal{S}$ and $\forall T > \max\{4\xi, 4(1-\xi)\}$. Inequality (5.22) combined with Lemma 5.4 implies that the elements in $\mathcal{V}_0$ are reachable by means of an input in $L^2[0, T]$. By using Lemma 5.2 we obtain assertion (a) of Theorem 5.1.

According to Lemma 7.3 in [13], $\forall \varepsilon > 0$ there exists a set $B_\varepsilon \subset (0,1)$, of Lebesgue measure 1, such that $\forall \xi \in B_\varepsilon$, there exists a constant $C_\xi > 0$ with

$$\lambda_{n+1} - \lambda_n \geq \frac{C_\xi}{|\lambda_n|^{1+\varepsilon}} \qquad \forall n \in \mathbb{Z}^*. \tag{5.23}$$

Relations (5.20), (5.21), (5.23), and Theorem 5.3 imply that there exists a constant $K_\xi > 0$ such that

$$\int_0^T \left| \frac{\partial \phi_2}{\partial x}(\xi, t) - \frac{\partial \phi_1}{\partial x}(\xi, t) \right|^2 dt \geq K_\xi \sum_{n \in \mathbb{Z}} \left(\frac{|a_n|^2 + |b_n|^2}{|\lambda_n|^{2\varepsilon}} \right) \tag{5.24}$$

$\forall \xi \in B_\varepsilon$ and $\forall T > \max\{4\xi, 4(1-\xi)\}$. Lemma 5.4 combined with (5.24) implies that $\forall s \in (0, \frac{1}{2})$, the elements in $\mathcal{V}_s$ are reachable by an input in $L^2[0,T]$. By applying again Lemma 5.2 we get assertion (b) of Theorem 5.1 for $s < \frac{1}{2}$. For $s \geq \frac{1}{2}$ the assertion remains true because $\mathcal{W}_s \subset \mathcal{W}_r$ for $s > r$.

In order to prove assertion (c) we notice that, $\forall \xi \in (0,1)$, we can use the continuous fractions expansion of $\frac{1-\xi}{\xi}$ to construct a sequence $(p(n))$ with values in $\mathbb{N}$, with $\lim_{n \to \infty} p(n) = \infty$, such that

$$\lambda_{p(n)+1} - \lambda_{p(n)} \leq \frac{C}{p(n)} \qquad \forall n \in \mathbb{N}. \tag{5.25}$$

If we denote by (ϕ_{1n}) (resp., by (ϕ_{2n})) the sequence of solutions of (5.7)–(5.9) (resp., of (5.10)–(5.12)) having initial data $(\sin(\frac{p(n)\pi}{\xi}), 0)$ (resp., $(\sin(\frac{(p(n)+1)\pi}{1-\xi}), 0)$, relations (5.13), (5.14), and (5.25) imply that

$$\lim_{n \to \infty} \frac{\displaystyle\int_0^T \left| \frac{\partial \phi_{2n}}{\partial x}(\xi, t) - \frac{\partial \phi_{1n}}{\partial x}(\xi, t) \right|^2 dt}{\|\phi_{1n}(0)\|^2_{H^s(0,\xi)} + \|\phi_{2n}(0)\|^2_{H^s(\xi,1)}} = 0$$

$\forall s < 0$. Using again Lemma 5.4 we conclude that (c) also holds. □

Remark 5.5. The fact that (5.24) holds for any $T > \max\{4\xi, 4(1-\xi)\}$ was proved in [4]. Earlier versions of this inequality (corresponding to larger values of T) were given in [13] and [14]. Notice that (5.24) and the standard duality argument imply only reachability of elements in $\mathcal{V}_s$. In order to get the reachability of elements in $\mathcal{W}_s$ we need a different argument, namely Lemma 5.2.

Remark 5.6. Intuitively it does not seem reasonable to have a minimal simultaneous reachability time depending on ξ. This question and other related issues (simultaneous approximate controllability, simultaneous spectral controllability) are tackled in [3]. In this work it is shown that the minimal time for these various types of controllability is $T = 2$.

Acknowledgments. The authors wish to thank Enrique Zuazua for suggesting a method to obtain the optimal approximate controllability time in Lemma 3.4. We are grateful to Sergei Avdonin, Francis Conrad, Scott Hansen, and Olof Staffans for helpful discussions and references.

REFERENCES

[1] G. Avalos and G. Weiss, *The Wave Equation as a Conservative Regular Linear System*, in preparation.

[2] S.-A. Avdonin and S.A. Ivanov, *Families of Exponentials—The Method of Moments in Controllability Problems for Distributed Parameter Systems*, Cambridge University Press, Cambridge, UK, 1995.

[3] S.-A. Avdonin and M. Tucsnak, *Simultaneous Controllability in Short Time for Two Elastic Strings*, preprint, 1999.

[4] C. Baiochi, V. Komornik, and P. Loreti, *Ingham type theorems and applications to control theory*, Boll. Un. Mat. Ital. B, II B (1999), pp. 33–63.

[5] C. Bardos, G. Lebeau, and J. Rauch, *Sharp sufficient conditions for the observation, control and stabilization of waves from the boundary*, SIAM J. Control Optim., 30 (1992), pp. 1024–1065.

[6] H. Brezis, *Functional Analysis*, Springer-Verlag, New York, 1988 (French original published by Masson, Paris).

[7] P.L. Butzer and H. Berens, *Semi-groups of Operators and Approximation*, Springer-Verlag, Berlin, 1967.

[8] J.W.S. Cassels, *An Introduction to Diophantine Approximation*, Cambridge University Press, Cambridge, UK, 1965.

[9] S. Dolecki and D.L. Russell, *A general theory of observation and control*, SIAM J. Control Optim., 15 (1977), pp. 185–220.

[10] S.W. Hansen, *Boundary control of a one-dimensional linear thermoelastic rod*, SIAM J. Control Optim., 32 (1994), pp. 1052–1074.

[11] S.W. Hansen and G. Weiss, *New results on the operator Carleson measure criterion*, IMA J. Math. Control Inform., 14 (1997), pp. 3–32.

[12] S.W. Hansen and B.Y. Zhang, *Boundary control of a linear thermoelastic beam*, J. Math. Anal. Appl., 210 (1997), pp. 182–205.

[13] S. Jaffard, M. Tucsnak, and E. Zuazua, *Singular internal stabilization of the wave equation*, J. Differential Equations, 145 (1998), pp. 184–215.

[14] S. Jaffard, M. Tucsnak, and E. Zuazua, *On a theorem of Ingham*, J. Fourier Anal. Appl., 3 (1997), pp. 577–582.

[15] A. Haraux, *Remarques sur la contrôlabilité ponctuelle et spectrale de systèmes distribuées*, Publications du laboratoire d'analyse numérique R89017, Preprint 2000/4, Institut Elie Cartan de Nancy, France, 1989.

[16] J.E. Lagnese and J.L. Lions, *Modelling, Analysis and Control of Thin Plates*, Masson, Paris, 1988.

[17] S. Lang, *Introduction to Diophantine Approximations*, Addison-Wesley, New York, 1966.

[18] I. Lasiecka and R. Triggiani, *A cosine operator approach to the modelling of $L_2(0,t;L^2(\Gamma))$-boundary input hyperbolic equations*, Appl. Math. Optim., 7 (1981), pp. 35–83.

[19] I. Lasiecka and R. Triggiani, *Regularity of hyperbolic equations under boundary terms*, Appl. Math. Optim., 7 (1981), pp. 35–83.

[20] J.-L. Lions, *Controlabilité exacte perturbations et stabilisation de systèmes distribués* 1, Rech. Math. Appl. 8, Masson, Paris, 1988.

[21] V. Komornik, *Exact Controllability and Stabilization—The Multiplier Method*, RAM Res. Appl. Math., John Wiley, Chichester, UK, Masson, Paris, 1994.

[22] D.L. Russell, *The Dirichlet–Neumann boundary control problem associated with Maxwell's equations in a cylindrical region*, SIAM J. Control Optim., 24 (1986), pp. 199–229.

[23] D.L. Russell and G. Weiss, *A general necessary condition for exact observability*, SIAM J. Control Optim., 32 (1994), pp. 1–23.

[24] G. Weiss, *Admissibility of unbounded control operators*, SIAM J. Control Optim., 27 (1989), pp. 527–545.

[25] G. Weiss, *Admissible observation operators for linear semigroups*, Israel J. Math., 65 (1989), pp. 17–43.

[26] G. Weiss, *Regular linear systems with feedback*, Math. Control Signals Systems, 7 (1994), pp. 23–57.

[27] E. Zuazua, *An Introduction to Exact Controllability for Distributed Systems*, Lecture notes, University of Lisbon, Lisbon, Portugal, 1990.

SIAM J. CONTROL OPTIM.
Vol. 38, No. 5, pp. 1428–1452

BIFURCATION CONTROL VIA STATE FEEDBACK FOR SYSTEMS WITH A SINGLE UNCONTROLLABLE MODE*

WEI KANG†

Abstract. The state feedback control of bifurcations with quadratic or cubic degeneracy is addressed for systems with a single uncontrollable mode. Based on normal forms and invariants, the classification of bifurcations for systems with a single uncontrollable mode is obtained (Table 1). Using invariants, stability characterizations are derived for a family of bifurcations, including saddle-node bifurcations, transcritical bifurcations, pitchfork bifurcations, and bifurcations with a cusp or hysteresis phenomenon. Bifurcations in systems under perturbed feedbacks are also addressed. In the case of a saddle-node bifurcation, continuous but not differentiable feedbacks are introduced to locally remove the bifurcation and to achieve the stability.

Key words. bifurcation, invariants, normal form, feedback control

AMS subject classifications. 93C10, 93C15

PII. S0363012997325927

1. Introduction.

Nonlinear dynamical systems exhibit complicated performance around bifurcation points. As the parameter of a system is varied, changes may occur in the qualitative structure of its solutions around a point of bifurcation. Using a feedback to stabilize a system with bifurcations has been studied by many authors (see, for instance, [1], [5], [8], [11], [20], and [25]). Many engineering applications of bifurcation control can be found in the literature (e.g., control of surge and rotating stall in engine compressors, flight control under high angle-of-attack). Quadratic and cubic feedbacks were introduced in [1] for the stabilization of bifurcated equilibria. It was proved in [1] that the periodic solution of a Hopf bifurcation can be stabilized by using state feedbacks. For the period doubling bifurcation, the method of harmonic balance was introduced in [8]. A feedback design method for delaying and stabilizing period doubling bifurcations was obtained. In [25], control laws were designed for the suppression of chaos in a thermal convection system model. A review of bifurcation and chaos in control systems can be found in [5]. More references on related topics can be found in [4], a bibliography of publications on bifurcation and chaos in control systems.

The main goal of this paper is to develop a framework for the analysis and control of bifurcations. Stability characterizations are obtained for control systems around bifurcations with quadratic or cubic degeneracy. Several well-known static bifurcations are addressed in this paper in a unified approach. This is made possible by using normal forms. The main results in sections 4 and 5 are summarized in Table 1, which is a complete classification of bifurcations with quadratic or cubic degeneracy for systems having a single uncontrollable mode. Because the system has only one uncontrollable mode, it does not have the Hopf bifurcation if the feedback stabilizes the controllable part. The Hopf bifurcation occurs for normal forms with two uncontrollable modes (see [7]).

What makes this paper unique is the approach based on the normal form and the

*Received by the editors August 11, 1997; accepted for publication (in revised form) June 29, 1999; published electronically May 11, 2000.
http://www.siam.org/journals/sicon/38-5/32592.html
†Mathematics Department, Naval Postgraduate School, Monterey, CA 93943 (wkang@math.nps.navy.mil).

TABLE 1

The bifurcations under state feedbacks. In the table, Q_1 is defined by (3.5), $\bar{Q}$ is defined by (4.1), D is defined by (4.23), and $\tilde{Q}$ is a function defined by $\tilde{Q}(a,b,c) = [a_1 \ -a_z \ 0]\ Q\ [a\ b\ c]^T$.

System	Condition		Bifurcation	Stability
System (2.2)	$Q_1(a_1,-a_z) \neq 0$	$\det(\bar{Q}) > 0$	The origin is an insolated equilibrium point	The system is unstable
		$\det(\bar{Q}) < 0$	Transcritical bifurcation	The system is stable at (z,x,μ) if $\hat{Q}(z,x_1,\mu) > 0$
				The system is unstable at (z,x,μ) if $\hat{Q}(z,x_1,\mu) < 0$
	$Q_1(a_1,-a_z) = 0$	$D \neq 0$ $\tilde{Q}(0,-a_\mu,a_1) \neq 0$	Pitchfork bifurcation	Supercritical if $D < 0$
				Subcritical if $D > 0$
System (2.3)	$Q_1(a_1,-a_z) \neq 0$		Saddle-node bifurcation	The system is stable at (z,x,μ) if $Q_1(z,x_1)z < 0$
				The system is unstable at (z,x,μ) if $Q_1(z,x_1)z > 0$
	$Q_1(a_1,-a_z) = 0$	$D \neq 0$	No bifurcation, there exists a unique equilibrium point for every fixed value of μ	All equilibrium points are stable if $D < 0$
				All equilibrium points are unstable if $D > 0$

invariants of nonlinear control systems. The two successful methods in the classical bifurcation theory are the normal form method and the projection method. They have been applied to control systems [1], [2], [21], [22]. The control system normal form adopted in the present paper is different from those used in the literature of nonlinear dynamical systems without control inputs. Why is it necessary to introduce the control system normal form instead of adopting the Poincaré normal form of vector fields? In fact, even for a linear control system $\dot{x} = Ax + Bu$, the controller normal form is more useful than the diagonal form of A in the feedback design. The normal form of nonlinear control systems generalizes the linear controller form. An affine control system $\dot{x} = f(x) + g(x)u$ has two vector fields $f(x)$ and $g(x)$. Therefore, the normal form of a control system requires the simplification of both f and g simultaneously. The simplification of f does not necessarily result in a simple form for g. Furthermore, the transformation group of control systems consists of changes of coordinates and feedbacks. This is different from the normal form theory of dynamical systems where feedbacks are not considered. The resonant terms defined for the control system normal form characterize the nature of a control system because they are invariant under both changes of coordinates and state feedbacks. The results obtained in this paper are intrinsic. They link the qualitative properties such as the bifurcation of control systems and its stability with their invariants.

An advantage of using the control system normal form is that the stability around bifurcations for a family of control systems is equivalent to the stability of their normal forms. This equivalence relation significantly simplifies the problem. It enables us to study a family of control systems with various bifurcations in a unified approach. The results proved in this paper provide a complete classification of bifurcations with quadratic or cubic degeneracy for systems having a single uncontrollable mode (Table 1).

Another advantage of the normal form approach is that the set of all equilibria of a control system (without feedback) in normal form can be found, and it is ap-

proximately a quadratic surface. Based on the geometric interpretation of equilibrium sets, we can describe how feedbacks change the distribution of the equilibrium points in the closed-loop system. This is important because the graph of the equilibrium points determines the type of the bifurcation. For certain systems, the position of an equilibrium point determines its stability as well. Furthermore, understanding the geometry of equilibrium sets enables us to characterize bifurcations under feedbacks which are not zero at the critical point.

The paper is organized as follows. Two bifurcation problems are formulated in section 2. Normal forms and invariants, the foundation of the framework, are introduced in section 3. The bifurcation control for systems with a single zero uncontrollable mode are studied in section 4 and section 5. In section 6, general feedbacks which are not zero at the critical point are studied. The classification of bifurcations are summarized in Table 1.

2. The problem formulation. Consider the following control system with a parameter

$$\dot{x} = f(x,\mu) + g(x,\mu)u, \qquad f(0,0) = 0, \tag{2.1}$$

where $x \in \mathbb{R}^n$ is the state variable, $u \in \mathbb{R}^m$ is the control input, and μ is the parameter. We assume that the rank of $g(x)$ is m at the point of interest. Unless it is otherwise specified, all vector fields and state feedbacks in this paper are C^k for some $k > 0$ sufficiently large. System (2.1) is said to be *linearly controllable* at $(x,\mu) = (0,0)$ if its linearization (A,B),

$$A = \frac{\partial f}{\partial x}(0,0), \qquad B = g(0,0)$$

is controllable. The origin $(x,\mu) = (0,0)$ is called an equilibrium or equilibrium point of (2.1) because $x(t) = 0$ is a constant solution if $\mu = 0$ and $u = 0$. Constant solutions may exist for other values of (x,μ,u). The *equilibrium set* is defined by

$$E = \{(x,\mu) | \text{ there exists } u_0 \in \mathbb{R} \text{ such that } f(x,\mu) + g(x,\mu)u_0 = 0\}.$$

A point in E is called an *equilibrium* or *equilibrium point*. Feedbacks are not involved in this definition. If the control input u is substituted by a feedback $u = u(x)$, a *closed-loop equilibrium*, (x_0,μ_0), is defined by $f(x_0,\mu_0) + g(x_0,\mu_0)u(x_0) = 0$. The set of all closed-loop equilibria is

$$E_c = \{(x,\mu) | f(x,\mu) + g(x,\mu)u(x) = 0\}.$$

The concept of an equilibrium set plays an important role in this paper. It is known that the closed-loop equilibrium set E_c, in general, is changed if the feedback is varied. However, the set E_c under any state feedback must be a subset of E. So, E consists of all possible closed-loop equilibria. The topology of E_c is induced from E.

The classical bifurcation theory studies the change of qualitative properties of dynamical systems as the parameters are varied. Qualitative properties include the topology of the equilibrium set, the stability, the existence of periodic solutions, etc. Control systems have two types of qualitative properties, which are those invariant under regular feedbacks (for example, the controllability, the stabilizability, and the topology of E) and those determined by the closed-loop system (for example, the closed-loop equilibria and the stability under a state feedback). Studying how these

properties are changed with parameters leads to the following two bifurcation problems for control systems.

PROBLEM 2.1. *Bifurcation of control systems. The problem focuses on the change of qualitative properties of control systems (such as controllability and stabilizability).*

PROBLEM 2.2. *Bifurcation control by using feedbacks. The problem focuses on the feedback design to achieve the stability around a critical point, or to achieve the desired performance by qualitatively changing a bifurcation.*

Problem 2.1 was addressed in [14] and [15] for systems with a single uncontrollable mode. In the present paper, we focus on Problem 2.2. It is proved that the same control system may exhibit several different kinds of bifurcations under different feedbacks. Instead of focusing on a single bifurcation, we ask the following questions. What kinds of bifurcations can occur in a control system, and what is the relationship between bifurcations and control laws? It is a different viewpoint from the existing bifurcation control approaches. It is known that local bifurcations at a linearly controllable point can be either removed or delayed by pole placement. In this paper, we study systems which are not linearly controllable. The work is motivated by engineering problems such as engine compressors and submersible vehicles [23], [20], [17]. In addition to the engineering applications, our research on uncontrollable systems is also motivated by the fact that qualitative properties such as controllability and stabilizability of control systems are generic (they are not changed by a small variation of parameters) at a linearly controllable point. If a system is not linearly controllable at a point, nonlinear phenomena such as bifurcations are expected around the critical point.

It is assumed throughout this paper that there exists a single uncontrollable mode (denoted by λ) in the linearization. The dimension of the state space is at least two ($n \geq 2$). If $\lambda \neq 0$, the sign of λ determines the stabilizability of the uncontrollable dynamics. Therefore, the variation of μ does not change the stability, i.e., there is no stationary bifurcation at $\mu = 0$. If $\lambda = 0$, the stability of the system depends on the value of the parameter. Different kinds of bifurcations occur in the performance. So, we focus on systems with $\lambda = 0$ in the following sections. Under a suitable linear change of coordinates and linear feedback, a system with a single uncontrollable mode $\lambda = 0$ can be transformed into one of the following forms (see [15] or [16]).

$$\begin{aligned} \dot{z} &= f_1(z, x, \mu) + g_1(z, x, \mu)u, \\ \dot{x} &= A_2 x + B_2 u + f_2(z, x, \mu) + g_2(z, x, \mu)u \end{aligned} \tag{2.2}$$

or

$$\begin{aligned} \dot{z} &= \mu + f_1(z, x, \mu) + g_1(z, x, \mu)u, \\ \dot{x} &= A_2 x + B_2 u + f_2(z, x, \mu) + g_2(z, x, \mu)u, \end{aligned} \tag{2.3}$$

where f_1, f_2, and their first derivatives equal zero at the origin $(z, x, \mu) = (0, 0, 0)$, g_1 and g_2 equal zero at the origin. The pair (A_2, B_2) is in the following Brunovsky form:

$$A_2 = \begin{bmatrix} 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 1 & \cdots & 0 \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ 0 & 0 & 0 & \cdots & 1 \\ 0 & 0 & 0 & \cdots & 0 \end{bmatrix}_{(n-1)\times(n-1)}, \quad B_2 = \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 1 \end{bmatrix}. \tag{2.4}$$

A feedback

$$u = \alpha(z, x, \mu) \tag{2.5}$$

for bifurcation control is a smooth function of (z, x, μ) such that $\alpha(0, 0, 0) = 0$. The linearization of a feedback is in the following form:

$$\alpha(z, x, \mu) = a_z z + \sum_{i=1}^{n-1} a_i x_i + a_\mu \mu + O(z, x, \mu)^2. \tag{2.6}$$

Notice that the value of μ is not always available. The function $\alpha(z, x, \mu)$ involves the parameter μ for two reasons. (1) Introducing μ in $\alpha(z, x, \mu)$ makes the theory more general. Feedbacks independent of μ form a subset of the feedbacks defined by (2.5). (2) By transforming the original system into it normal form, the linearization of the last equations in (2.2) and (2.3) has only one term which is the input u. However, the original system may have the terms with μ in the last equation. In this case, the term μ is absorbed in (2.5). So, a_μ in (2.6) comes from the original model. It is not necessarily zero.

In this paper, a system is said to be *stable* at a given equilibrium point if it is locally asymptotically stable at this point. To achieve stability, feedbacks in this paper are assumed to satisfy the following assumption.

ASSUMPTION 2.3. *The state feedback* (2.6) *places the controllable poles in the left half plane, i.e., the eigenvalues of the matrix* $A_2 + B_2 \begin{bmatrix} a_1 & a_2 & \cdots & a_{n-1} \end{bmatrix}$ *are all in the left half plane.*

It is known that $(-1)^n a_1$ equals the multiplication of all eigenvalues of the matrix

$$A_2 + B_2 \begin{bmatrix} a_1 & a_2 \cdots a_{n-1} \end{bmatrix}.$$

From Assumption 2.3, these eigenvalues are on the left half plane. So we have the following lemma.

LEMMA 2.4. *If a feedback* (2.6) *satisfies Assumption* 2.3*, then* $a_1 < 0$.

3. Normal forms and invariants. In this section, nonlinear invariants are defined by the coefficients of resonant terms. Then quadratic normal forms in [14] and [15] are introduced without proof.

3.1. Resonant terms and invariants. In the classical theory of dynamical systems, a set of resonant terms was found for the homogeneous parts of nonlinear systems. The coefficients of resonant terms are invariant under homogeneous transformations. For systems with the Hopf bifurcation, the values of invariants determine the stability of the periodic solutions. For control systems, the invariants were introduced in [12] for linearly controllable systems. In this section, a set of invariants is found for systems which are not linearly controllable. It plays a key role in the stability analysis for control systems with bifurcations.

The quadratic and cubic terms in the Taylor expansion of vector fields are used in the proofs of many results. The homogeneous parts of degree d for f_i and g_i in (2.2) and (2.3) are denoted by $f_i^{[d]}$ and $g_i^{[d]}$. For instance, the quadratic terms in the Taylor expansion of $f_1 + g_1 u$ have the form $f_1^{[2]}(z, x, \mu) + g_1^{[1]}(z, x, \mu)u$. The components of $f_i^{[d]}$ and $g_1^{[d-1]}$ are homogeneous polynomials of degree d and $d-1$, respectively. A homogeneous transformation of degree d for control systems consists of the change of coordinates and state feedbacks in the form

$$\begin{aligned} &z = \bar{z} + \phi_1^{[d]}(\bar{z}, \bar{x}, \mu), \qquad x = \bar{x} + \phi_2^{[d]}(\bar{z}, \bar{x}, \mu), \\ &u = \bar{u} + \alpha^{[d]}(\bar{z}, \bar{x}, \mu) + \beta^{[d-1]}(\bar{z}, \bar{x}, \mu)\bar{u}, \end{aligned} \tag{3.1}$$

where $\bar{z}$ and $\bar{x}$ are the new coordinates and $\bar{u}$ is the new control input introduced by the regular feedback. A transformation of degree d does not change the terms of degree less than d in a control system. If $d = 2$, (3.1) is called a quadratic transformation. If $d = 3$, it is called a cubic transformation.

DEFINITION 3.1. *Consider* (2.2) *or* (2.3). *A homogeneous term in* $f_i^{[d]}(z, x, \mu)$ *or* $g_i^{[d-1]}(z, x, \mu)u$, *is called a resonant term if transformations of the form* (3.1) *leave the coefficient of the term invariant. The coefficient of a resonant term is called an invariant.*

For instance, if (3.1) is applied to (2.2), it can be proved that the term $\bar{z}^2$ in $\bar{f}_1^{[2]}$ of the resulting system has the same coefficient as the term z^2 in $f_1^{[2]}$ of (2.2). So z^2 in $f_1^{[2]}$ is resonant. Notice that the resonant terms of control systems are different from the resonant terms in the classical theory of dynamical systems. A resonant term in Definition 3.1 is invariant under both changes of coordinates and state feedbacks. However, the classical dynamic systems theory does not deal with any feedback. In the next theorem, resonant terms are found for (2.2) and (2.3). Define

$$\begin{aligned} &R^{[d]}(z, x_1, \mu) = f_1^{[d]}(z, x, \mu)|_{x_2=x_3=\cdots=x_{n-1}=0}, \\ &R_1^{[d]}(z, x_1) = R(z, x_1, 0), \end{aligned} \tag{3.2}$$

where $f_1^{[d]}(z, x, \mu)$ is the homogeneous vector field of degree d from the Taylor expansion of $f_1(z, x, \mu)$ in (2.2) and (2.3).

THEOREM 3.2. *In* (2.2), *all terms of* $R^{[d]}(z, x_1, \mu)$ *are resonant. In* (2.3), *all terms of* $R_1^{[d]}(z, x_1)$ *are resonant.*

Proof. Consider the system (2.2). Suppose that (2.2) is transformed into the following system by (3.1):

$$\begin{aligned} \dot{\bar{z}} &= \bar{f}_1(\bar{z}, \bar{x}, \mu) + \bar{g}_1(\bar{z}, \bar{x}, \mu)\bar{u}, \\ \dot{\bar{x}} &= A_2\bar{x} + B_2\bar{u} + \bar{f}_2(\bar{z}, \bar{x}, \mu) + \bar{g}_2(\bar{z}, \bar{x}, \mu)\bar{u}. \end{aligned}$$

It was proved in [13] that the homogeneous parts of f_1 and $\bar{f}_1$ satisfy the homological equation

$$\frac{\partial \phi_1^{[d]}(z, x, \mu)}{\partial x} A_2 x = f_1^{[d]}(z, x, \mu) - \bar{f}_1^{[d]}(z, x, \mu).$$

However,

$$\frac{\partial \phi_1^{[d]}}{\partial x} A_2 x = \frac{\partial \phi_1^{[d]}}{\partial x_1} x_2 + \cdots + \frac{\partial \phi_1^{[d]}}{\partial x_{n-2}} x_{n-1}. \tag{3.3}$$

Therefore, every term of $f_1^{[d]} - \bar{f}_1^{[d]}$ has at least one of the variables $x_2, \ldots, x_{n-1}$. The terms in $R(z, x_1, \mu)$ do not appear in $f_1^{[d]} - \bar{f}_1^{[d]}$. This implies that the function $R^{[d]}(z, x_1, \mu)$ in $f_1^{[d]}$ is invariant under (3.1).

Now consider the system (2.3). The homological equation for $f_1^{[d]}$ is

$$\frac{\partial \phi_1^{[d]}(z, x, \mu)}{\partial z} \mu + \frac{\partial \phi_1^{[d]}(z, x, \mu)}{\partial x} A_2 x = f_1^{[d]}(z, x, \mu) - \bar{f}_1^{[d]}(z, x, \mu).$$

From (3.3), every nonzero term in $f_1^{[d]} - \bar{f}_1^{[d]}$ has at least one of the variables μ, $x_2, \ldots, x_{n-1}$. This implies that the coefficients in $R_1^{[2]}(z, x_1)$ are not changed by (3.1). ∎

In the following, the coefficients of resonant terms are denoted by γ with corresponding subindices. For example, the coefficient of z^2 in $R_1^{[2]}(z,x_1)$ or $R^{[2]}(z,x_1,\mu)$ is γ_{zz}, the coefficient of $zx_1\mu$ is $\gamma_{zx_1\mu}$, etc. Based on this theorem, the coefficients γ_{zz}, γ_{zx_1}, $\gamma_{z\mu}$, $\gamma_{x_1x_1}$, $\gamma_{x_1\mu}$, $\gamma_{\mu\mu}$ in (2.2) and the coefficients γ_{zz}, γ_{zx_1}, $\gamma_{x_1x_1}$ in (2.3) are called (quadratic) invariants. They are part of the quadratic invariants introduced in [12] and [15] using Lie brackets. The quadratic functions of resonant terms $R^{[2]}(z,x_1,\mu)$ and $R_1^{[2]}(z,x_1)$ determine two symmetric matrices,

$$Q=\begin{bmatrix} \gamma_{zz} & \frac{\gamma_{zx_1}}{2} & \frac{\gamma_{z\mu}}{2} \\ \frac{\gamma_{zx_1}}{2} & \gamma_{x_1x_1} & \frac{\gamma_{x_1\mu}}{2} \\ \frac{\gamma_{z\mu}}{2} & \frac{\gamma_{x_1\mu}}{2} & \gamma_{\mu\mu} \end{bmatrix}, \qquad Q_1=\begin{bmatrix} \gamma_{zz} & \frac{\gamma_{zx_1}}{2} \\ \frac{\gamma_{zx_1}}{2} & \gamma_{x_1x_1} \end{bmatrix}. \tag{3.4}$$

In this paper, the quadratic function defined by Q (or Q_1) is also denoted by $Q(x,y,z)$ (or $Q_1(x,y)$), i.e.,

$$Q(x,y,z)=\begin{bmatrix} x & y & z \end{bmatrix} Q \begin{bmatrix} x & y & z \end{bmatrix}^T, \quad Q_1(x,y)=\begin{bmatrix} x & y \end{bmatrix} Q_1 \begin{bmatrix} x & y \end{bmatrix}^T. \tag{3.5}$$

Equivalently,

$$Q(z,x_1,\mu)=R^{[2]}(z,x_1,\mu), \quad Q_1(z,x_1)=R_1^{[2]}(z,x_1).$$

3.2. Quadratic normal forms. Since systems with the same normal form have equivalent bifurcations, most proofs in this paper are given for quadratic normal forms. From [15] and [16], (2.2) and (2.3) can be transformed into a unique system in normal form by a suitable quadratic transformation of the form (3.1) with $d=2$.

For (2.2), the normal form is

$$\begin{aligned} \dot z &= \sum_{i=2}^{n-1}\gamma_{x_ix_i}x_i^2+Q(z,x_1,\mu)+O(z,x,\mu,u)^3, \\ \dot x &= A_2x+B_2u+\tilde f_2^{[2]}(x)+O(z,x,\mu,u)^3. \end{aligned} \tag{3.6}$$

For (2.3), the normal form is

$$\begin{aligned} \dot z &= \mu+\sum_{i=2}^{n-1}\gamma_{x_ix_i}x_i^2+\gamma_{x_1\mu}x_1\mu+Q_1(z,x_1)+O(z,x,\mu,u)^3, \\ \dot x &= A_2x+B_2u+\tilde f_2^{[2]}(x)+O(z,x,\mu,u)^3, \end{aligned} \tag{3.7}$$

where $\tilde f_2^{[2]}(x)$ is in the extended controller form of [12]. Details are omitted since it is not used in this paper. Before the end of this section, we introduce the following well-known result on stationary bifurcations (see [9]).

THEOREM 3.3. *Consider the following one-dimensional dynamical system with a parameter μ.*

$$\dot x=f(x,\mu), \qquad f(0,0)=0, \qquad x\in\mathbb{R}. \tag{3.8}$$

(i) *It has a saddle-node bifurcation at the origin if*

$$f_x(0,0)=0, \qquad f_\mu(0,0)\neq 0, \qquad f_{xx}(0,0)\neq 0. \tag{3.9}$$

(ii) *It has a transcritical bifurcation at the origin if*

$$\begin{aligned} &f_x(0,0)=0, \qquad f_\mu(0,0)=0,\\ &f_{xx}(0,0)\neq 0, \qquad f_{x\mu}^2(0,0)-f_{xx}(0,0)f_{\mu\mu}(0,0)>0. \end{aligned} \tag{3.10}$$

(iii) *It has a pitchfork bifurcation at the origin if*

$$f_x(0,0)=0, \quad f_\mu(0,0)=0, \quad f_{xx}(0,0)=0, \quad f_{x\mu}(0,0)\neq 0, \quad f_{xxx}(0,0)\neq 0. \tag{3.11}$$

If $f_{xxx}(0,0)<0$, *the pitchfork bifurcation is supercritical. If* $f_{xxx}(0,0)>0$, *it is subcritical.*

4. Bifurcations of system (2.2). In this section, all bifurcations of (2.2) with both quadratic and cubic degeneracies are addressed. An example is given to illustrate the feedback design technique of this section. In the example, the model of engine compressors with the transcritical bifurcation is studied. Feedbacks are designed to stabilize the equilibria on the positive side of the uncontrollable variable R.

Define a matrix $\bar{Q}$ from the linearization of a feedback and the quadratic invariants of (2.2)

$$\bar{Q}=\begin{bmatrix} a_1 & -a_z & 0\\ 0 & -a_\mu & a_1 \end{bmatrix} Q \begin{bmatrix} a_1 & 0\\ -a_z & -a_\mu\\ 0 & a_1 \end{bmatrix}, \tag{4.1}$$

where Q is defined by (3.4). The matrix $\bar{Q}$ is used in the next theorem to characterize the bifurcation. Following the notation introduced in section 2, E_c represents the set of closed-loop equilibrium points, i.e.,

$$E_c=\{(x,\mu)|f(x,\mu)+g(x,\mu)\alpha(x,\mu)=0\}. \tag{4.2}$$

THEOREM 4.1. *Consider a closed-loop system* (2.2)–(2.6) *satisfying Assumption* 2.3. *Suppose*

$$Q_1(a_1,-a_z)\neq 0. \tag{4.3}$$

(i) *If* $\bar{Q}$ *is sign definite, then* $(z,x,\mu)=(0,0,0)$ *is an isolated equilibrium point of the closed-loop system. It is unstable.*

(ii) *If* $\bar{Q}$ *is indefinite with full rank, then the closed-loop system has a transcritical bifurcation around the origin.*

(iii) *Assume that the feedback satisfies the condition in* (ii). *Given any* $(z,x,\mu)\in E_c$ *in a neighborhood of the origin, it is locally asymptotically stable if*

$$\begin{bmatrix} a_1 & -a_z & 0 \end{bmatrix} Q \begin{bmatrix} z & x_1 & \mu \end{bmatrix}^T>0. \tag{4.4}$$

The system is unstable if

$$\begin{bmatrix} a_1 & -a_z & 0 \end{bmatrix} Q \begin{bmatrix} z & x_1 & \mu \end{bmatrix}^T<0. \tag{4.5}$$

Remark. In $zx_1\mu$-space, the open loop equilibrium set E of (3.6) is approximately a cone

$$\begin{aligned} &Q(z,x_1,\mu)=0,\\ &x_i=0 \qquad \text{for } i=2,3,\dots,n-1, \end{aligned} \tag{4.6}$$

provided that Q is indefinite with full rank (see [15]). The third and higher degree terms are omitted in the approximation. Under the feedback

$$\alpha(z,x,\mu)=\sum a_i x_i + a_z z + a_\mu \mu + O(z,x,\mu)^2,$$

the closed-loop equilibrium set E_c is the intersection between E and the surface $\alpha(z,x,\mu)=0$. Therefore, E_c is approximated by the intersection between the plane $a_z z + a_1 x_1 + a_\mu \mu = 0$ and the cone (4.6). The geometry of E_c has two generic cases, which are (1) it has a unique point, and (2) the intersection consists of two lines. The first case is characterized by (i) of Theorem 4.1. The graph of E_c in the second case indicates a transcritical bifurcation, which is proved by (ii) of Theorem 4.1. □

Since (2.2) is equivalent to its normal form (3.6), it is important to find a center manifold and the reduced system on it for the normal form (3.6). This is obtained in the following lemma.

LEMMA 4.2. *Consider the quadratic normal form* (3.6). *Under the state feedback* (2.6), *the center manifold of the closed-loop system satisfies*

$$x_1 = -\frac{a_z}{a_1} z - \frac{a_\mu}{a_1}\mu + O(z,\mu)^2, \qquad x_i = O(z,\mu)^2 \quad for \;\; i = 2,\dots,n-1. \tag{4.7}$$

The reduced system on the center manifold satisfies

$$\dot z = \frac{1}{a_1^2} \begin{bmatrix} z & \mu \end{bmatrix} \bar Q \begin{bmatrix} z & \mu \end{bmatrix}^T + O(z,\mu)^2. \tag{4.8}$$

Proof. From [3], the center manifold is determined by a function $x=\pi(z,\mu)$, where $\pi(z,\mu)$ can be approximated by polynomials. The function $\pi(z,\mu)$ satisfies an equation of the following form:

$$A\pi(z,\mu) + Bu(z,\pi,\mu) + O(z,\pi,\mu)^2 = \frac{\partial \pi}{\partial z} O(z,\pi,\mu)^2.$$

Denote the linear part of $\pi(z,\mu)$ by $\pi^{[1]}(z,\mu)$. The linearization of this equation is

$$\pi_2^{[1]} = 0, \quad \pi_3^{[2]} = 0, \dots, \quad a_z z + a_1 \pi_1^{[1]}(z,\mu) + a_\mu \mu = 0. \tag{4.9}$$

It is easy to check that the linear part of (4.7) satisfies (4.9). The functions in (4.7) are equivalent to

$$\begin{bmatrix} z \\ x_1 \\ \mu \end{bmatrix} = \frac{1}{a_1} \begin{bmatrix} a_1 & 0 \\ -a_z & -a_\mu \\ 0 & a_1 \end{bmatrix} \begin{bmatrix} z \\ \mu \end{bmatrix} + O(z,\mu)^2, \quad x_i = O(z,\mu)^2 \;\text{ for }\; 2 \le i \le n-1.$$

Substituting this relation into the z dynamical equation in (3.6), we obtain (4.8) as the reduced system on the center manifold. □

The proof of Theorem 4.1. Since (2.2) can be transformed into its normal form (3.6) by a quadratic transformation, and since the conditions in the theorem are invariant under quadratic transformations, it is enough to prove the result for the quadratic normal form (3.6).

(i) The closed-loop system is equivalent to its reduced system (4.8) on the center manifold. Denote the right side of (4.8) by $f_c(z,\mu)$. If $\bar Q$ is sign definite, then $(z,\mu) = (0,0)$ is the unique local solution of $f_c(z,\mu)=0$. Therefore, the origin is an isolated equilibrium point. In this case, the reduced system (4.8) at $\mu = 0$ is

$$\dot z = \frac{1}{a_1^2} Q_1(a_1, -a_z) z^2 + O(z)^3.$$

Since $\det(\bar{Q}) > 0$ and since $Q_1(a_1, -a_z)$ is the first diagonal entry in $\bar{Q}$, we know that $Q_1(a_1, -a_z) \neq 0$. The system is unstable. Part (i) is proved.

(ii) Now, assume that $\det(\bar{Q}) < 0$. It is obvious that

$$\frac{\partial f_c}{\partial z}(0,0) = 0, \qquad \frac{\partial f_c}{\partial \mu}(0,0) = 0. \tag{4.10}$$

It is easy to check that

$$\frac{\partial^2 f_c}{\partial z^2}(0,0) = \frac{1}{a_1^2} Q_1(a_1, -a_z) \neq 0. \tag{4.11}$$

$$\frac{\partial^2 f_c}{\partial z^2}(0,0)\frac{\partial^2 f_c}{\partial \mu^2}(0,0) - \left(\frac{\partial^2 f_c}{\partial z \mu}(0,0)\right)^2 = 4\det(\bar{Q}) < 0. \tag{4.12}$$

Therefore, the conditions in (ii) of Theorem 3.3 are satisfied. This implies that the closed-loop system has a transcritical bifurcation.

(iii) The stability of the closed-loop system agrees with the reduced system on the center manifold. It is easy to check that

$$\frac{\partial f_c}{\partial z} = \frac{2}{a_1^2} \begin{bmatrix} 1 & 0 \end{bmatrix} \bar{Q} \begin{bmatrix} z & \mu \end{bmatrix}^T + O(z,\mu)^2. \tag{4.13}$$

If (z, x, μ) is in E_c, then (z, μ) is an equilibrium point on the center manifold,

$$\begin{bmatrix} z & \mu \end{bmatrix} \bar{Q} \begin{bmatrix} z & \mu \end{bmatrix}^T + O(z,\mu)^3 = 0, \qquad x_1 = -\frac{a_z}{a_1} z - \frac{a_\mu}{a_1}\mu + O(z,\mu)^2. \tag{4.14}$$

Therefore,

$$\begin{bmatrix} z & \mu \end{bmatrix} = \begin{bmatrix} z_0 & \mu_0 \end{bmatrix} t + O(t)^2, \qquad x_1 = \left(-\frac{a_z}{a_1} z_0 - \frac{a_\mu}{a_1}\mu_0\right) t + O(t)^2, \tag{4.15}$$

where $t \in \mathbb{R}$ and $(z_0, \mu_0) \neq (0,0)$ satisfies

$$\begin{bmatrix} z_0 & \mu_0 \end{bmatrix} \bar{Q} \begin{bmatrix} z_0 & \mu_0 \end{bmatrix}^T = 0. \tag{4.16}$$

From (4.1) and (4.15), we have

$$\frac{1}{a_1} \begin{bmatrix} a_1 & -a_z & 0 \end{bmatrix} Q \begin{bmatrix} z & x_1 & \mu \end{bmatrix}^T = \frac{1}{a_1^2} \begin{bmatrix} 1 & 0 \end{bmatrix} \bar{Q} \begin{bmatrix} z_0 & \mu_0 \end{bmatrix}^T t + O(t)^2. \tag{4.17}$$

If we can prove that

$$\begin{bmatrix} 1 & 0 \end{bmatrix} \bar{Q} \begin{bmatrix} z_0 & \mu_0 \end{bmatrix}^T \neq 0, \tag{4.18}$$

then the sign of $\frac{1}{a_1^2} \begin{bmatrix} 1 & 0 \end{bmatrix} \bar{Q} \begin{bmatrix} z_0 & \mu_0 \end{bmatrix}^T t$, which agrees with that of $\frac{\partial f_c}{\partial z}$, is opposite to the sign of the number given by

$$\begin{bmatrix} a_1 & -a_z & 0 \end{bmatrix} Q \begin{bmatrix} z & x_1 & \mu \end{bmatrix}^T,$$

because $a_1 < 0$. Therefore, (4.4) implies that $\frac{\partial f_c}{\partial z} < 0$ at the point in E_c around zero. The closed-loop system is locally asymptotically stable. Similarly, (4.5) implies that the system is unstable.

Now we prove (4.18) by contradiction. Suppose

$$\begin{bmatrix} 1 & 0 \end{bmatrix} \bar{Q} \begin{bmatrix} z_0 & \mu_0 \end{bmatrix}^T = 0. \tag{4.19}$$

Because $\det(\bar{Q}) \neq 0$, and because $(z_0, \mu_0) \neq (0,0)$, we have that $\bar{Q} \begin{bmatrix} z_0 & \mu_0 \end{bmatrix}^T \neq 0$. Equation (4.16) implies that $\begin{bmatrix} z_0 & \mu_0 \end{bmatrix} \bar{Q} \begin{bmatrix} z_0 & \mu_0 \end{bmatrix}^T = 0$. Comparing this equation with (4.19), we have that $\mu_0 = 0$ and $z_0 \neq 0$. Therefore, (4.19) implies $q_{1\,1} = 0$, where $q_{1\,1}$ is the entry in $\bar{Q}$ at the upper-left corner. However, $q_{1\,1} = Q(a_1, -a_z)$. It contradicts (4.3). Therefore, (4.18) is true.

Remark. From (4.13), it is obvious that the system is locally asymptotically stable at $(z, x, \mu) \in E_c$ if $\begin{bmatrix} 1 & 0 \end{bmatrix} \bar{Q} \begin{bmatrix} z & \mu \end{bmatrix} < 0$. This is another method of stability testing.

Theorem 4.1 deals with bifurcation control under the assumption $Q_1(a_1, -a_z) \neq 0$. In the following, we study the case where $Q_1(a_1, -a_z) = 0$. For this purpose, we need cubic invariants. The function of cubic resonant terms, $R^{[3]}(z, x_1)$, of the normal form (3.6) is denoted by $C(z, x_1)$, which is

$$C(z, x_1) = R_1^{[3]}(z, x_1) = f_1^{[3]}(z, x, \mu)|_{x_2 = x_3 = \cdots = x_{n-1} = \mu = 0}, \tag{4.20}$$

where $f_1^{[3]}$ represents the cubic part in the first equation of (3.6). The state feedback for bifurcation control is

$$\begin{aligned} &u = \alpha(z, x, \mu), \\ &\alpha(z, x, \mu) = a_z z + a_1 x_1 + \cdots + a_{n-1} x_{n-1} + a_\mu x_\mu + \alpha^{[2]}(z, x, \mu) + O(z, x, \mu)^3, \end{aligned} \tag{4.21}$$

where $\alpha^{[2]}$ is a quadratic homogeneous polynomial. The coefficients in $\alpha^{[2]}$ are denoted by a_{zz}, $a_{z\mu}$, $a_{x_1 x_1}$, $a_{x_1 \mu}$, etc. The following quadratic function from $\alpha^{[2]}$ is useful:

$$\alpha_{z x_1}^{[2]}(z, x_1) = a_{zz} z^2 + a_{z x_1} z x_1 + a_{x_1 x_1} x_1^2, \tag{4.22}$$

i.e., $\alpha_{z x_1}^{[2]}$ is the restriction of $\alpha^{[2]}$ to the $z x_1$-plane. To simplify the notation, we define

$$D = a_1 C(a_1, -a_z) + (2 a_z \gamma_{x_1 x_1} - a_1 \gamma_{z x_1}) \alpha_{z x_1}^{[2]}(a_1, -a_z), \tag{4.23}$$

where $\gamma_{x_1 x_1}$ and $\gamma_{z x_1}$ are quadratic invariants in (3.4).

THEOREM 4.3. *Consider a closed-loop system* (3.6)–(4.21) *satisfying Assumption* 2.3.

(i) *Suppose*

$$Q_1(a_1, -a_z) = 0. \tag{4.24}$$

Then the closed-loop system has a pitchfork bifurcation at the origin provided

$$\begin{aligned} &D \neq 0, \\ &\begin{bmatrix} a_1 & -a_z & 0 \end{bmatrix} Q \begin{bmatrix} 0 & -a_\mu & a_1 \end{bmatrix}^T \neq 0. \end{aligned} \tag{4.25}$$

(ii) *The pitchfork bifurcation is supercritical if $D < 0$. It is subcritical if $D > 0$.*

Proof. The center manifold of (3.6)–(4.21) is the graph of $x = \pi(z, \mu)$, which satisfies

$$\frac{\partial \pi}{\partial z} \left(\sum_{i=2}^{n-1} \gamma_{x_i x_i} \pi_i^2 + Q(z, \pi_1, \mu) \right) = A_2 \pi + B_2 \alpha(z, \pi, \mu) + \tilde{f}_2^{[2]}(\pi) + O(z, \pi, \mu)^3. \tag{4.26}$$

The Taylor expansion of $\pi(z,\mu)$ is $\pi(z,\mu)=\pi^{[1]}(z,\mu)+\pi^{[2]}(z,\mu)+\cdots$. Solving (4.26) to the second degree, we have

$$
\begin{aligned}
&\pi_1^{[1]}(z,x)=-\frac{a_z}{a_1}z-\frac{a_\mu}{a_1}\mu, \qquad \pi_i^{[1]}(z,\mu)=0 \ \text{ for } \ i=2,\dots,n-1,\\
&\pi_1^{[2]}(z,\mu)=-\frac{1}{a_1}\begin{bmatrix} z & \pi_1^{[1]} & \mu \end{bmatrix}\Big(-\frac{a_z a_2}{a_1}Q+Q_{fb}\Big)\begin{bmatrix} z & \pi_1^{[1]} & \mu \end{bmatrix}^T,\\
&\pi_2^{[2]}(z,\mu)=-\frac{a_z}{a_1}Q(z,\pi_1^{[1]},\mu),
\end{aligned}
\tag{4.27}
$$

where Q_{fb} represents the matrix of the quadratic function of z, x_1, and μ in the feedback (4.21). More specifically,

$$
Q_{fb}=\begin{bmatrix} a_{zz} & \dfrac{a_{zx_1}}{2} & \dfrac{a_{z\mu}}{2}\\ \dfrac{a_{zx_1}}{2} & a_{x_1x_1} & \dfrac{a_{x_1\mu}}{2}\\ \dfrac{a_{z\mu}}{2} & \dfrac{a_{x_1\mu}}{2} & a_{\mu\mu} \end{bmatrix}.
$$

The closed-loop system has the same bifurcation as its reduced system on the center manifold. The reduced system is $\dot z=f_c(z,\mu)$, where $f_c(z,\mu)$ is obtained by substituting $x=\pi(z,\mu)$ into the z dynamical equation in (3.6)–(4.21). It has the following approximation:

$$
f_c(z,\pi,\mu)=\begin{bmatrix} z & \pi_1 & \mu \end{bmatrix}Q\begin{bmatrix} z & \pi_1 & \mu \end{bmatrix}^T+f_1^{[3]}(z,\pi^{[1]},\mu)+O(z,\mu)^4. \tag{4.28}
$$

From (4.27), we have

$$
\begin{bmatrix} z & \pi_1^{[1]} & \mu \end{bmatrix}Q\begin{bmatrix} z & \pi_1^{[1]} & \mu \end{bmatrix}^T=\frac{1}{a_1^2}\begin{bmatrix} z & \mu \end{bmatrix}\bar Q\begin{bmatrix} z & \mu \end{bmatrix}^T. \tag{4.29}
$$

Therefore,

$$
\begin{aligned}
&\frac{\partial f_c}{\partial z}(0,0)=0, \quad \frac{\partial f_c}{\partial \mu}(0,0)=0,\\
&\frac{\partial^2 f_c}{\partial z^2}(0,0)=\frac{2}{a_1^2}Q_1(a_1,-a_z),\\
&\frac{\partial^2 f_c}{\partial z\mu}(0,0)=\frac{1}{a_1^2}\begin{bmatrix} a_1 & -a_z & 0 \end{bmatrix}Q\begin{bmatrix} 0 & -a_\mu & a_1 \end{bmatrix}^T.
\end{aligned}
\tag{4.30}
$$

From (4.27) and (4.24), we have

$$
\begin{bmatrix} z & \pi_1^{[1]}(z,0) & 0 \end{bmatrix}Q_{fb}\begin{bmatrix} z & \pi_1^{[1]}(z,0) & 0 \end{bmatrix}^T=\alpha_{zx_1}^{[2]}\Big(z,-\frac{a_z}{a_1}z\Big)=\frac{1}{a_1^2}\alpha_{zx_1}^{[2]}(a_1,-a_z)z^2,
$$

$$
\begin{bmatrix} z & \pi_1^{[1]}(z,0) & 0 \end{bmatrix}Q\begin{bmatrix} z & \pi_1^{[1]}(z,0) & 0 \end{bmatrix}^T=Q_1\Big(z,-\frac{a_z}{a_1}z\Big)=\frac{1}{a_1^2}Q_1(a_1,-a_z)z^2=0,
$$

$$
f_1^{[3]}(z,\pi^{[1]}(z,0),0)=C\Big(z,-\frac{a_z}{a_1}z\Big)=\frac{1}{a_1^3}C(a_1,-a_z)z^3. \tag{4.31}
$$

Equations (4.31) and (4.27) imply

$$
\pi_1(z,0)=-\frac{a_z}{a_1}z-\frac{1}{a_1^3}\alpha_{zx_1}^{[2]}(a_1,-a_z)z^2. \tag{4.32}
$$

In order to find the coefficient of z^3 in $f_c(z, \pi(z,0),0)$, we substitute the formulae of $f_1^{[3]}(z, \pi^{[1]}(z,0),0)$ and $\pi_1(z,0)$ in (4.31) and (4.32) into (4.28). The substitution yields

$$\frac{\partial^3 f_c}{\partial z^3}(0,0) = \frac{6}{a_1^4}\Big(a_1 C(a_1,-a_z) + (2a_z\gamma_{x_1x_1} - a_1\gamma_{zx_1})\alpha^{[2]}_{zx_1}(a_1,-a_z)\Big) = \frac{6}{a_1^4}D. \tag{4.33}$$

From (4.30), (4.33), (4.24), and (4.25), we have

$$\frac{\partial^2 f_c}{\partial z^2}(0,0) = 0, \qquad \frac{\partial^2 f_c}{\partial z\mu}(0,0) \neq 0, \qquad \frac{\partial^3 f_c}{\partial z^3}(0,0) \neq 0. \tag{4.34}$$

Equations (4.34) and (4.30) imply that the reduced dynamical system on the center manifold satisfies (3.11). Therefore, the system has a pitchfork bifurcation. Furthermore, the bifurcation is supercritical if $D < 0$, and it is subcritical if $D > 0$. □

Since $\alpha^{[2]}_{zx_1}$ is in the feedback, its coefficients are adjustable. There always exist suitable quadratic functions $\alpha^{[2]}_{zx_1}$ which render the pitchfork bifurcation supercritical, provided

$$2a_z\gamma_{x_1x_1} - a_1\gamma_{zx_1} \neq 0. \tag{4.35}$$

This condition is related to the rank of Q_1. From $Q_1(a_1,-a_z) = 0$, we have

$$\begin{bmatrix} a_1, -a_z \end{bmatrix} \begin{bmatrix} \gamma_{zz}a_1 - \frac{\gamma_{zx_1}}{2}a_z & \frac{\gamma_{zx_1}}{2}a_1 - \gamma_{x_1x_1}a_z \end{bmatrix}^T = 0. \tag{4.36}$$

Therefore, if $2a_z\gamma_{x_1x_1} - a_1\gamma_{zx_1} = 0$, (4.36) implies $a_1(\gamma_{zz}a_1 - \frac{\gamma_{zx_1}}{2}a_z) = 0$. Since $a_1 \neq 0$, we have $Q_1 \begin{bmatrix} a_1 & -a_z \end{bmatrix}^T = \begin{bmatrix} 0 & 0 \end{bmatrix}^T$. So, $\operatorname{rank}(Q_1) < 2$. This is equivalent to saying that $\operatorname{rank}(Q_1) = 2$ implies (4.35). Therefore, we have the following.

COROLLARY 4.4. *Suppose that* (3.6)–(4.21) *satisfies* (4.24) *and* (4.25). *If* Q_1 *has full rank, then there exists a quadratic function* $\alpha^{[2]}_{zx_1}(z,x_1)$ *for the nonlinear feedback such that the closed-loop system has a supercritical pitchfork bifurcation.*

Remark. If $C(a_1,-a_z) > 0$, then $\alpha^{[2]}_{zx_1}(z,x_1) = 0$ implies $D < 0$. In this case, a linear feedback renders the pitchfork bifurcation supercritical.

Now we introduce the following model of engine compressors as an example of Theorem 4.1. The system exhibits various bifurcation phenomena (see, for instance, [23], [20], [18], and [6]). The compressor control has been studied by many authors ([19], [20], and [18]). The example introduced here is not for the purpose of proving new results for compressor control. We use the model to illustrate some ideas of feedback design based on the results of this section. In the following, the results obtained in [18] are proved using Theorem 4.1. The simplest model that describes the system is a three-state ODE in Moore and Greitzer [24],

$$\begin{aligned} \dot{R} &= \sigma R(-2\phi - \phi^2 - R), \\ \dot{\phi} &= -\psi - \frac{3}{2}\phi^2 - \frac{1}{2}\phi^3 - 3R\phi - 3R, \\ \dot{\psi} &= \phi - \sqrt{\psi+\psi_0}\Big(\frac{2}{\sqrt{\psi_0}} + \mu + u\Big) + 2, \end{aligned} \tag{4.37}$$

where $R \geq 0$ is the normalized stall cell squared amplitude, ϕ is the mass flow, ψ is the pressure rise, and ψ_0 and σ are constant positive numbers. The control input u can be changed by varying the throttle opening. The system has an uncertain parameter

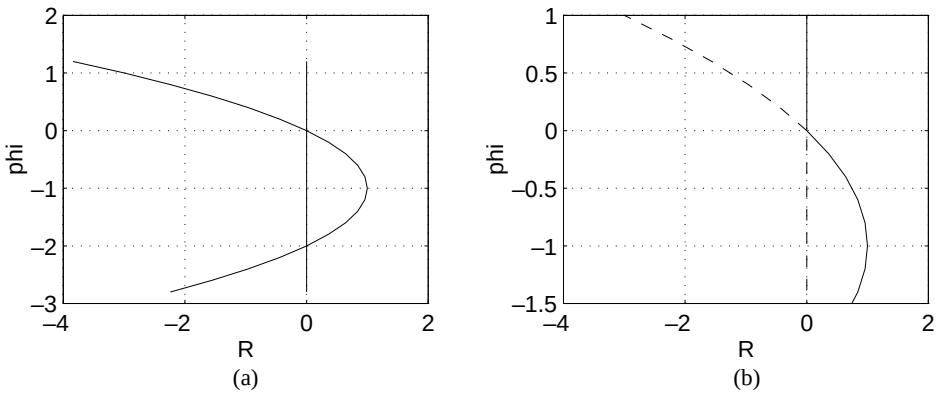


FIG. 4.1. *The equilibrium set of* (4.38).

μ. The values of ϕ, ψ, and u are shifted by a constant [19] so that the origin is the focal bifurcation point. A simple linear change of coordinates $z = R$, $x_1 = \phi$, and $x_2 = -\psi - 3R$ transforms the dynamics into the following system in the form of (2.3).

$$\begin{aligned} \dot{z} &= -2\sigma z x_1 - \sigma z^2 - \sigma z x_1^2, \\ \dot{x}_1 &= x_2 - \frac{3}{2}x_1^2 - \frac{1}{2}x_1^3 - 3zx_1, \\ \dot{x}_2 &= -\frac{3}{\psi_0}z - x_1 - \frac{1}{\psi_0}x_2 + \sqrt{\psi_0}(\mu + u) + O(z, x_1, x_2, \mu, u)^2. \end{aligned} \tag{4.38}$$

Although the quadratic part of the system is not in normal form, the invariant matrices Q and Q_1 are found from the resonant terms

$$Q = \begin{bmatrix} Q_1 & 0 \\ 0 & 0 \end{bmatrix}, \qquad Q_1 = \begin{bmatrix} -\sigma & -\sigma \\ -\sigma & 0 \end{bmatrix}.$$

So Q_1 is indefinite. In zx_1-plane (or equivalently $R\phi$-plane), the graph of the equilibrium set E is shown in Figure 4.1a. The state of the real system always stays in the region $R \geq 0$. Therefore, it is desired to find state feedbacks which render the system asymptotically stable at equilibrium points with $R > 0$. The equilibrium points with $R < 0$ are meaningless. Consider a state feedback

$$u = k_1 R + k_2 \phi + k_3 \psi + O(R, \phi, \psi)^2. \tag{4.39}$$

In the coordinates (z, x_1, x_2), (4.39) is

$$u = (k_1 - 3k_3)z + k_2 x_1 - k_3 x_2 + O(z, x_1, x_2)^2. \tag{4.40}$$

Substituting (4.40) into (4.38), the closed-loop system satisfies

$$a_z = -\frac{3}{\psi_0} + \sqrt{\psi_0}(k_1 - 3k_3), \quad a_1 = -1 + \sqrt{\psi_0}k_2, \quad a_2 = -\frac{1}{\psi_0} - \sqrt{\psi_0}k_3, \quad a_\mu = \sqrt{\psi_0}. \tag{4.41}$$

It is easy to check that

$$Q_1(a_1, -a_z) = -\sigma a_1(a_1 - 2a_z), \qquad \bar{Q} = \sigma \begin{bmatrix} a_1(a_1 - 2a_z) & -a_1 a_\mu \\ -a_1 a_\mu & 0 \end{bmatrix}. \tag{4.42}$$

Equations in (4.42) imply that the conditions in (ii) of Theorem 4.1 are satisfied if $a_1 \neq 2a_z$. Therefore, the bifurcation is transcritical. In the following, we use the result (iii) of Theorem 4.1 to find all feedbacks which stabilize the system at the equilibrium points with $R > 0$. From (4.38), the equilibrium points with $z > 0$ satisfy $z = -(2x_1 + x_1^2)$, and

$$\begin{bmatrix} a_1 & -a_z \end{bmatrix} Q_1 \begin{bmatrix} z & x_1 \end{bmatrix}^T = \sigma\left((a_1 - 2a_z)x_1 + (a_1 - a_z)x_1^2\right), \qquad x_1 < 0. \tag{4.43}$$

So (4.43) and Theorem 4.1 imply that, to guarantee the stability of the system at the equilibrium points with $z > 0$, we need $a_z > \dfrac{a_1}{2}$. Substituting (4.40) into this inequality, we have

$$2k_1 - k_2 - 6k_3 > (6 - \psi_0)/\psi_0^{\frac{3}{2}}. \tag{4.44}$$

Another branch of equilibrium points satisfies $z = 0$, so $\begin{bmatrix} a_1 & -a_z \end{bmatrix} Q_1 \begin{bmatrix} z & x_1 \end{bmatrix}^T = -\sigma a_1 x_1$. On this curve, the closed-loop system is locally asymptotically stable if $x_1 > 0$. In summary, if the feedback (4.39) satisfies Assumption 2.3 and (4.44), then the closed-loop system has a transcritical bifurcation. It is locally asymptotically stable at the equilibrium points where $R > 0$ or $R = 0$ and $\phi > 0$. The closed-loop equilibrium points are shown in Figure 4.1b. The system is locally stable on the solid curve, and unstable on the dotted curve.

If the scaled amplitude of the rotating stall cell is adopted as a state to replace R, the system model has a pitchfork bifurcation at the critical point. Its normal form and invariants satisfy the conditions in Corollary 4.4. Stabilization feedbacks obtained in [20] and [18] for the pitchfork bifurcation can be derived based on the invariants and Corollary 4.4. Details are omitted for the reason of space.

5. Bifurcations of system (2.3). In this section, bifurcations of (2.3) with both quadratic and cubic degeneracy are addressed. The results are summarized in Table 1. The bifurcations are classified based on the number $Q_1(a_1, -a_z)$. It was proved in [15] that the equilibrium set E of (2.3) is approximately a paraboloid or a saddle defined by

$$\mu + Q_1(z, x_1) = 0. \tag{5.1}$$

The set of closed-loop equilibrium points, E_c, is approximately the intersection between (5.1) and the plane

$$a_z z + a_1 x_1 + a_\mu \mu = 0, \tag{5.2}$$

where a_z, a_1, and a_μ are the coefficients in (2.6). Therefore, E_c is a quadratic curve. The following theorem proves that the closed-loop system has a saddle-node bifurcation. The stability of the system is characterized by the location of the closed-loop equilibrium points. Define two subsets, E_- and E_+, of the open loop equilibrium set E by

$$E_- = \{(z, x, \mu) \in E | Q_1(z, x_1)z < 0\}, \qquad E_+ = \{(z, x, \mu) \in E | Q_1(z, x_1)z > 0\}. \tag{5.3}$$

The following theorem proves that $E_c \cap E_-$ consists of stable equilibrium points and $E_c \cap E_+$ consists of unstable equilibrium points. The origin is the border between the two parts.

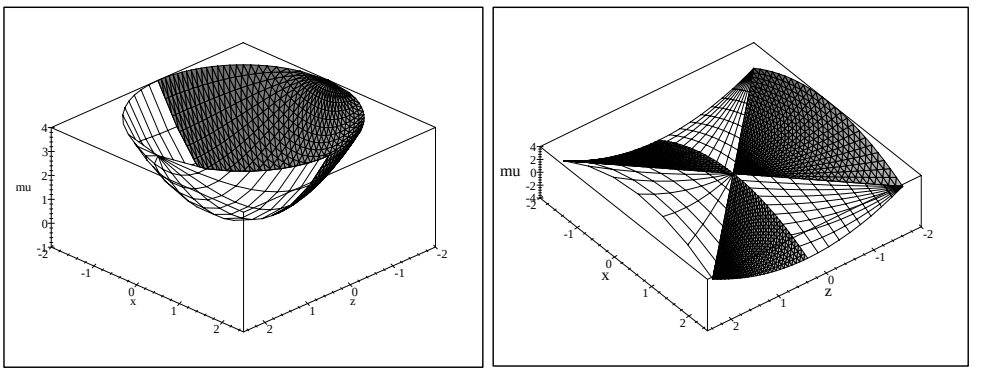


FIG. 5.1. *The gridded area is E_-. The shaded area is E_+.*

THEOREM 5.1. *Consider a closed-loop system* (2.3)–(2.6) *satisfying Assumption* 2.3. *Suppose*

$$Q_1(a_1, -a_z) \neq 0. \tag{5.4}$$

(i) *The system has a saddle-node bifurcation at the origin. The set E_c satisfies*

$$\begin{aligned} x_1 &= -\frac{a_z}{a_1} z + O(z)^2, \qquad & \mu &= -\frac{1}{a_1^2} Q(a_1, -a_z) z^2 + O(z)^3, \\ x_i &= O(z)^2, & i &= 2, \ldots, n-1. \end{aligned} \tag{5.5}$$

(ii) *The closed-loop system is locally asymptotically stable at the points in $E_c \cap E_-$ and the system is unstable at the points in $E_c \cap E_+$. In a neighborhood of the origin,*

$$E_c = (E_c \cap E_-) \cup (E_c \cap E_+) \cup \{(z, x, \mu) = (0, 0, 0)\}. \tag{5.6}$$

Condition (5.4) is always true if Q_1 is sign definite. If Q_1 is indefinite, E is approximately a saddle. Condition (5.4) implies that the plane (5.2) does not meet (5.1) on a line in the zx_1-plane. If (5.4) is false, E_c is not a quadratic curve of μ. This case is addressed later in this section.

Remark. Geometrically, the sets E_- and E_+ are simple. In [15] and [16], it was proved that the open loop equilibrium set E is approximately either a paraboloid or a saddle given by (5.1). In the case of a paraboloid, Q_1 is sign definite. So E_- and E_+ are simply the half paraboloid given by $z > 0$ and $z < 0$ (Figure 5.1a). In the case of a saddle, the set $Q_1(z, x)z = 0$ consists of three planes in $zx_1\mu$-space given by $Q_1(z, x) = 0$ and $z = 0$. They divide E into six parts. Three of them form E_- and the others form E_+. A point is in $E_c \cap E_-$ (or $E_c \cap E_+$) if and only if $\mu z > 0$ (or $\mu z < 0$) (Figure 5.1b).

LEMMA 5.2. *If $n \geq 3$, the center manifold of a system* (3.7)–(2.6) *has the form*

$$x_1 = -\frac{a_z}{a_1} z - \frac{1}{a_1}\left(a_\mu - \frac{a_2 a_z}{a_1}\right)\mu + O(z, \mu)^2, \qquad x_2 = -\frac{a_z}{a_1}\mu + O(z, \mu)^2, \tag{5.7}$$

and $x_i = O(z, \mu)^2$ for $i \geq 3$. The reduced system on center manifold is defined by

$$\begin{aligned} f_c(z, \mu) = \mu &+ Q_1\left(z, -\frac{a_z}{a_1} z - \frac{1}{a_1}\left(a_\mu - \frac{a_2 a_z}{a_1}\right)\mu\right) \\ &+ \gamma_{x_2 x_2} \frac{a_z^2}{a_1^2}\mu^2 + \gamma_{x_1 \mu}\left(-\frac{a_z}{a_1} z - \frac{1}{a_1}\left(a_\mu - \frac{a_2 a_z}{a_1}\right)\mu\right)\mu + O(z, \mu)^3. \end{aligned} \tag{5.8}$$

Proof. The center manifold $x = \pi(z, \mu)$ satisfies the equation

$$A\pi(z,\mu) + Bu(z,\pi,\mu) + O(z,\pi,\mu)^2 = \frac{\partial \pi}{\partial z}(\mu + O(z,\pi,\mu)^2).$$

Formula (5.7) is obtained by solving the linearization of this equation. The reduced dynamical system on the center manifold is derived by substituting (5.7) into the z dynamical equation in (3.7). □

Proof. The proof of Theorem 5.1. (i) Since (5.4) and (5.5) are invariant under (3.1), we prove Theorem 5.1 for systems in the normal form (3.7). To find E_c, we set the dynamics in the closed-loop system to be zero. Solving the linear and quadratic parts of the equations yields

$$x_1 = -\frac{a_z}{a_1} z - \frac{a_\mu}{a_1}\mu + O(z,\mu)^2, \qquad x_i = O(z,\mu)^2, \;\; i = 2, \ldots, n-1, \tag{5.9}$$

where a_1 is not zero (Lemma 2.4). Substitute this into the following equation:

$$\mu + \sum_{i=2}^{n-1} \gamma_{x_i x_i} x_i^2 + \gamma_{x_1 \mu} x_1 \mu + Q_1(z, x_1) + O(z, x, \mu, u)^3 = 0.$$

Therefore, $\mu = -\dfrac{1}{a_1^2} Q_1(a_1, -a_z) z^2 + O(z)^3$. This relation and (5.9) imply the condition (5.5).

To study the bifurcation, we focus on the center manifold. From (5.8), the following relation can be derived.

$$\frac{\partial f_c}{\partial z}(0,0) = 0, \qquad \frac{\partial f_c}{\partial \mu}(0,0) = 1 \neq 0, \qquad \frac{\partial^2 f_c}{\partial z^2}(0,0) = \frac{2}{a_1^2} Q(a_1, -a_z) \neq 0. \tag{5.10}$$

From (ii) of Theorem 3.3, the closed-loop system has a saddle-node bifurcation.

(ii) We prove (5.6) first. It is known that the points in E_c satisfy (5.5). Substituting (5.5) into $Q_1(z, x_1)z$, we have

$$Q_1(z, x_1) z = \frac{1}{a_1^2} Q_1(a_1, -a_z) z^3 + O(z)^4. \tag{5.11}$$

Notice that $Q_1(a_1, -a_z) \neq 0$. This implies that locally the points in E_c satisfy $Q_1(z, x_1)z \neq 0$ if $z \neq 0$. Therefore, (5.6) holds. To prove the stability properties claimed in Theorem 5.1, it is enough to show that (5.8) is locally stable at points in $E_c \cap E_-$ and unstable in $E_c \cap E_+$. From (5.5) and (5.8), it can be proved that at a point in E_c,

$$\frac{\partial f_c}{\partial z} = \frac{2}{a_1^2} Q_1(a_1, -a_z) z + O(z)^2. \tag{5.12}$$

Therefore, (5.8) is locally asymptotically stable when $Q_1(a_1, -a_z)z < 0$. From (5.11), the sign of $Q_1(a_1, -a_z)z$ is the same as the sign of $Q_1(z, x_1)z$ for points in E_c. Therefore, (5.8) is locally stable at any point in $E_c \cap E_-$. Similarly, (5.8) is unstable at any point in $E_c \cap E_+$. □

Remark. From (5.12), it is obvious that the closed-loop system is asymptotically stable at $(z, x, \mu) \in E_c$ if $Q_1(a_1, -a_z)$ and z have opposite sign. This is another way to test the stability.

The feedback in Theorem 5.1 is smooth. A saddle-node bifurcation occurs. The system has unstable equilibrium points around the origin. Furthermore, there is no closed-loop equilibrium point for either $\mu > 0$ or $\mu < 0$. In the following, C^0 feedbacks are employed under which the closed-loop system has a unique equilibrium point for any value of μ, no matter if μ is positive or negative. And the C^0 feedback renders the system locally stable at all the equilibrium points. In fact, a continuous (but not differentiable at $z = 0$) feedback can place all the closed-loop equilibrium points around the origin inside E_-. The C^0 feedback is

$$\alpha(z, x, \mu) = a_z|z| + a_1x_1 + \cdots + a_{n-1}x_{n-1} + a_\mu x_\mu + O(z, x, \mu)^2, \tag{5.13}$$

where $|z|$ is not differentiable, and the rest of the function is C^k for some $k \geq 1$.

THEOREM 5.3. *Consider* (2.3)–(5.13) *satisfying Assumption* 2.3. *Suppose* a_z *satisfies*

$$Q_1(a_1, -a_z) < 0, \quad Q_1(a_1, a_z) > 0. \tag{5.14}$$

Then, in a neighborhood of $(z, x, \mu) = (0, 0, 0)$, (2.3)–(5.13) *has a unique equilibrium point for every value of* μ. *The equilibrium point is stable if* $\mu \neq 0$.

The quadratic form $Q_1(x, y)$ has both positive and negative values if it is indefinite. The region in the xy-plane defined by $Q_1(x, y) > 0$ has two connected pieces. If the region is not symmetric with respect to the x-axis, there always exists an a_z satisfying (5.14). If $Q_1(x, y)$ is positive or negative definite, E is a paraboloid (Figure 5.1a). It is impossible to obtain equilibrium points in E_c for both $\mu < 0$ and $\mu > 0$, no matter what kind of feedback is used.

Remark. The theorem does not claim the stability at $\mu = 0$. In fact, the unique equilibrium at $\mu = 0$ is $(z, x) = (0, 0)$. However, the system is not smooth at $z = 0$. Therefore, the theory of center manifold is not applicable. The justification of the stability at $\mu = 0$ requires tools of nonsmooth vector fields. For some special cases, the stability at $\mu = 0$ is obvious. For example, if $f_1(z, x, \mu) = 0$ and $g_1(z, x, \mu) = 0$ at all the points $(z, x, \mu) = (0, x, 0)$, then the subspace $z = 0$ is an invariant submanifold. It follows a standard argument that, on each side of $z = 0$, the performance of the system is equivalent to a smooth vector field. The stability is guaranteed by the center manifold theory.

Proof. The feedback is equivalent to

$$\alpha(z, x, \mu) = \begin{cases} a_z z + a_1x_1 + \cdots + a_{n-1}x_{n-1} + a_\mu x_\mu + O(z, x, \mu)^2 & \text{if } z \geq 0, \\ -a_z z + a_1x_1 + \cdots + a_{n-1}x_{n-1} + a_\mu x_\mu + O(z, x, \mu)^2 & \text{if } z < 0. \end{cases}$$

From (5.5), the closed-loop equilibrium points are

$$\begin{aligned} \mu &= \begin{cases} -\dfrac{1}{a_1^2}Q(a_1, -a_z)z^2 + O(z)^3 & \text{if } z \geq 0, \\ -\dfrac{1}{a_1^2}Q(a_1, a_z)z^2 + O(z)^3 & \text{if } z < 0, \end{cases} \\ x_1 &= -\frac{a_z}{a_1}|z| + O(z)^2, \quad x_i = O(z)^2, \ \ i = 2, \ldots, n-1. \end{aligned} \tag{5.15}$$

The condition (5.14) and the first equation of (5.15) imply that the sign of μ at any equilibrium point is the same as the sign of z. Therefore, there exist equilibrium points on both sides of $\mu = 0$. Since the first equation of (5.15) is locally a one-to-one correspondence, (2.3)–(5.13) has a unique equilibrium point for every value of μ.

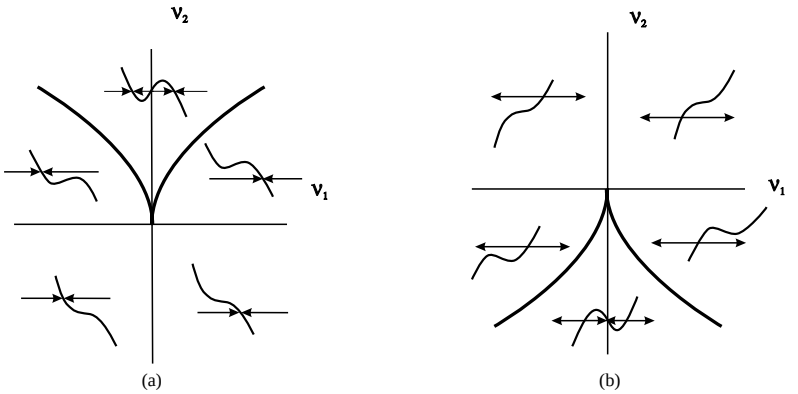


FIG. 5.2. *A cusp and some phase portraits of* $f(z, \nu_1, \nu_2)$. $D_3(0) < 0$ *in* (a); $D_3(0) > 0$ *in* (b).

To study the stability, the reduced system on the center manifold $\dot{z} = f_c(z, \mu)$ is used, where $f_c(z, \mu)$ is defined by (5.8). From (5.12), we have

$$\frac{\partial f_c}{\partial z}(z, \mu) = \begin{cases} \dfrac{2}{a_1^2} Q_1(a_1, -a_z) z + O(z)^2, & z > 0, \\ \dfrac{2}{a_1^2} Q_1(a_1, a_z) z + O(z)^2, & z < 0. \end{cases}$$

This relation and the condition (5.14) imply that $\dfrac{\partial f_c}{\partial z}(z, \mu) < 0$ at every closed-loop equilibrium point with $\mu \neq 0$. So the system is stable at equilibrium points with $\mu \neq 0$. □

The feedback in Theorems 5.1 and 5.3 satisfies $Q_1(a_1, -a_z) \neq 0$. If it does not hold, the quadratic part of a vector field is degenerate. We focus on this case in the rest of this section. Before the next theorem is introduced, it is necessary to review the bifurcation with a cusp equilibrium set. Consider a one-dimensional system with two parameters ν_1 and ν_2:

$$\begin{aligned} &\dot{z} = f(z, \nu_1, \nu_2) = \nu_1 + \nu_2 z + D_2(\nu_1, \nu_2) z^2 + D_3(\nu_1, \nu_2) z^3 + O(z)^4, \\ &D_2(0, 0) = 0, \qquad D_3(0, 0) \neq 0. \end{aligned} \tag{5.16}$$

The system has cubic degeneracy. Following [10], the set of degenerate equilibrium points is defined by $f(z, \nu_1, \nu_2) = 0$ and $\dfrac{\partial f}{\partial z}(z, \nu_1, \nu_2) = 0$. It is a cusp approximated by

$$\nu_2^3 = -27 D_3(0, 0) \nu_1^2 / 4. \tag{5.17}$$

Following [10], the stability around the cusp is illustrated in Figure 5.2. The solid cusp represents (5.17). The cubic curves in Figure 5.2 represent the graphs of $y = f(z, \nu_1, \nu_2)$ for a pair of fixed values (ν_1, ν_2) in different regions. In Figure 5.2a, $D_3(0, 0)$ is less than zero. In Figure 5.2b, $D_3(0, 0)$ is positive. A system with $D_3(0, 0) < 0$ performs more stably than otherwise.

Now let's consider (3.7) with the feedback (4.21). It is proved that the system has a saddle-node bifurcation if $Q_1(a_1, -a_z) \neq 0$. If $Q_1(a_1, -a_z) = 0$, the center manifold

of the closed-loop system is $x_i = \pi_i(z,\mu)$ for $i = 1,\dots,n-1$. Its linear and quadratic parts are

$$\pi_1^{[1]}(z,\mu) = -\frac{a_z}{a_1}z - \frac{1}{a_1}\left(a_\mu - \frac{a_2 a_z}{a_1}\right)\mu, \pi_2^{[1]}(z,\mu) = -\frac{a_z}{a_1}\mu, \quad \pi_i^{[1]}(z,\mu) = 0 \text{ for } i \geq 3,$$

$$\pi_1^{[2]}(z,0) = -\frac{1}{a_1^3}\alpha_{zx_1}^{[2]}(a_1,-a_z)z^2, \qquad \pi_2^{[2]}(z,0) = 0,\dots,\pi_{n-1}^{[2]}(z,0) = 0, \tag{5.18}$$

in which $\alpha_{zx_1}^{[2]}$ is defined by (4.21) and (4.22). We only consider the case $n \geq 3$, which is slightly different from $n = 2$. The argument for $n = 2$ is similar. Substituting $x = \pi(z,\mu)$ into the uncontrollable part in (3.7)–(4.21), the reduced system on the center manifold is

$$\begin{aligned} \dot{z} &= f_c(z,\mu), \\ f_c(z,\mu) &= \mu + \sum_{i=2}^{n-1}\gamma_{x_i x_i}\pi_i^2 + \gamma_{x_1\mu}\pi_1\mu + Q_1(z,\pi_1) \\ &\quad + f_1^{[3]}(z,\pi,\mu) + g_1^{[2]}(z,\pi,\mu)\alpha(z,\pi,\mu) + O(z,\pi,\mu)^3. \end{aligned} \tag{5.19}$$

Substitute (5.18) into f_c. The expansion of f_c in z is

$$f_c(z,\mu) = D_0(\mu) + D_1(\mu)z + D_2(\mu)z^2 + D_3(\mu)z^3 + \cdots, \tag{5.20}$$

$$D_0(\mu) = \mu + O(\mu)^2, \ \ D_1(\mu) = O(\mu), \ \ D_2(\mu) = \frac{1}{a_1^2}Q_1(a_1,-a_z) + O(\mu) = O(\mu), \tag{5.21}$$

$$D_3(0) = \frac{1}{a_1^3}C(a_1,-a_z) + \frac{1}{a_1^4}(2a_z\gamma_{x_1x_1} - a_1\gamma_{zx_1})\alpha_{zx_1}^{[2]}(a_1,-a_z) = D, \tag{5.22}$$

where $C(a_1,-a_z)$ is defined by (4.20). The notation D is introduced in (4.23). In the following, we introduce the unfolding of (5.20). Let ν_1 and ν_2 be the parameters which replace the coefficients of lower degree terms in (5.20). The unfolding system is

$$\dot{z} = \tilde{f}_c(z,\nu_1,\nu_2) = \nu_1 + \nu_2 z + \tilde{D}_2(\nu_1)z^2 + \tilde{D}_3(\nu_1)z^3 + \cdots, \tag{5.23}$$

$$\nu_1 = D_0(\mu), \quad \nu_2 = D_1(\mu). \tag{5.24}$$

where $\tilde{D}_2(\nu_1) = D_2(D_0^{-1}(\nu_1))$ and $\tilde{D}_3(\nu_1) = D_3(D_0^{-1}(\nu_1))$. This system is in the form of (5.16). Its bifurcation diagram is illustrated in Figure 5.2. From (5.21), we have $\nu_2 = O(\nu_1)$. Therefore, the curve (5.24) and the cusp intersect at the origin. The curve (5.24) is transversal to the ν_2-axis. Therefore, the curve does not get inside the cusp area because both curves of the cusp are tangent to the ν_2-axis (see Figure 5.3). Around the origin, (5.19) has a unique equilibrium point for every value of μ. From (5.22), the system is always stable if $D < 0$.

THEOREM 5.4. *Consider* (3.7)–(4.21) *satisfying Assumption* 2.3. *Suppose*

$$\begin{aligned} &Q_1(a_1,-a_z) = 0, \\ &D \neq 0. \end{aligned} \tag{5.25}$$

(i) *The system has no bifurcation. There exists a neighborhood of* $(z,x,\mu) = (0,0,0)$ *in which* (3.7)–(4.21) *has a unique closed-loop equilibrium point for any value of* μ.

(ii) *The system* (3.7)–(4.21) *is stable if* $D < 0$. *It is unstable if* $D > 0$.

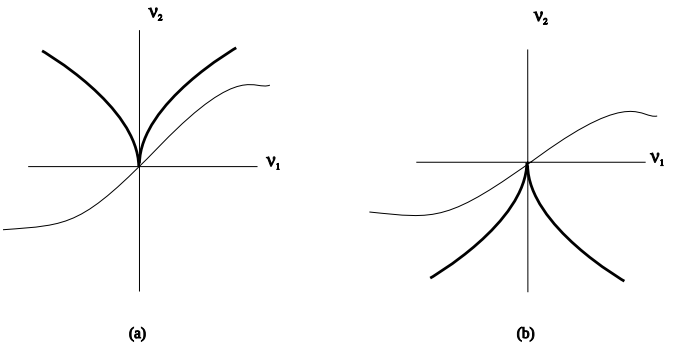


FIG. 5.3. *The picture of the cusp* (5.17) *and the curve* (5.25).

Remark. Notice that the inequality $D < 0$ is also a condition in Theorem 4.3. From the proof of Corollary 4.4, stabilizing feedbacks can be found for the system (3.7)–(4.21) around the cusp, provided that Q_1 has full rank.

The result in Theorem 5.4 can be proved without embedding it into the cusp bifurcation. However, the embedding becomes critical when more complicated cases are considered. For example, if the control $\alpha(z, x, \mu)$ is not zero at the origin, a two-dimensional diagram is required for the analysis as shown in the next section.

6. Closed-loop system with a perturbed feedback. In the previous sections, all feedbacks are assumed to be zero at the origin. However, bifurcations can be qualitatively changed if the feedback does not vanish at the origin. A feedback which is not zero at the origin is treated as a perturbed feedback,

$$u = \alpha(z, x, \mu) + \nu, \tag{6.1}$$

where $\alpha(z, x, \mu)$ satisfies $\alpha(0, 0, 0) = 0$ and Assumption 2.3. The parameter ν is a perturbation.

6.1. System (2.3). It is proved in section 4 that (2.2)–(6.1) has a transcritical bifurcation at $\nu = 0$ if $\det(\bar{Q}) < 0$. From the imperfection theory, two saddle-node bifurcations are generated at the critical point if the symmetry in the system is broken.

THEOREM 6.1. *Consider a control system* (2.2) *with* $\det(Q) \neq 0$. *Suppose that* $\alpha(z, x, \mu)$ *satisfies* (ii) *of Theorem* 4.1. *For any fixed value of* ν *around* $\nu = 0$, (2.2)–(6.1) *is a system with a single parameter* μ. *It has two saddle-node bifurcations around the origin if* $\nu \neq 0$. *The bifurcation diagram of* (2.2)–(6.1) *in* $z\mu$*-plane is shown in Figure* 6.1.

Proof. Given the normal form (3.6) with the feedback (6.1), its center manifold satisfies

$$x_1 = -\frac{a_z}{a_1} z - \frac{a_\mu}{a_1}\mu - \frac{1}{a_1}\nu + O(z, \mu, \nu)^2, \qquad x_i = O(z, \mu, \nu)^2 \quad \text{for } i \geq 2. \tag{6.2}$$

Denote

$$P = \begin{bmatrix} 1 & 0 & 0 \\ -a_z/a_1 & -a_\mu/a_1 & 1/a_1 \\ 0 & 1 & 0 \end{bmatrix}, \qquad P^T Q P = \begin{bmatrix} \bar{Q} & Q_{12} \\ Q_{21} & Q_{22} \end{bmatrix},$$

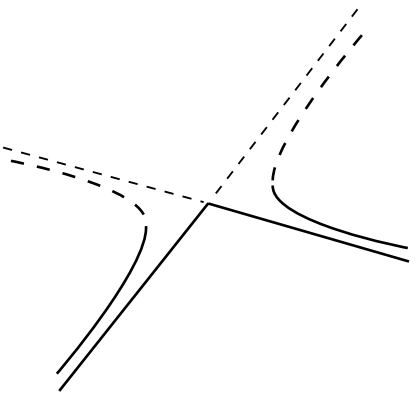


FIG. 6.1. *The bifurcation diagram of the closed-loop system for a fixed value of ν. At $\nu = 0$, it is a transcritical bifurcation. At $\nu \neq 0$, it consists of two saddle-node bifurcations.*

where $\bar{Q}$ is defined by (4.1), $Q_{12}^T = Q_{21}$, and the reduced system on the center manifold is

$$\dot{z} = \begin{bmatrix} z & \mu & \nu \end{bmatrix} P^T Q P \begin{bmatrix} z & \mu & \nu \end{bmatrix}^T + O(z,\mu,\nu)^3. \tag{6.3}$$

Define $\begin{bmatrix} \bar{z} & \bar{\mu} \end{bmatrix} = \begin{bmatrix} z & \mu \end{bmatrix} + Q_{21}\bar{Q}^{-1}\nu$; then

$$\dot{\bar{z}} = (Q_{22} - Q_{21}\bar{Q}^{-1}Q_{12})\nu^2 + \begin{bmatrix} \bar{z} & \bar{\mu} \end{bmatrix} \bar{Q} \begin{bmatrix} \bar{z} & \bar{\mu} \end{bmatrix}^T + O(\bar{z},\bar{\mu},\bar{\nu})^3. \tag{6.4}$$

Since Q has full rank, it is easy to check that $Q_{22} - Q_{21}\bar{Q}^{-1}Q_{12} \neq 0$. The bifurcation of (6.4) at $\nu \neq 0$ is the ν perturbation of a transcritical bifurcation. In the imperfection theory (see, for instance, [26] or [9]), it was proved that two saddle-node bifurcations are generated around the critical point (Figure 6.1). □

6.2. System (3.7). Consider (3.7). If the invariant quadratic form $Q_1(z, x_1)$ is positive or negative definite, the equilibrium set of (3.7) is a paraboloid. Its intersection with the surface $\nu + \alpha(z, x, \mu) = 0$ is the set of closed-loop equilibrium points. It is approximately a quadratic curve for any value of ν around zero. Furthermore, the closed-loop system has a saddle-node bifurcation for small values of μ. Therefore, the perturbation does not qualitatively change the bifurcation. Another case is more interesting in which $Q_1(z, x_1)$ is not sign definite and $Q_1(a_1, -a_z) = 0$. In this case, it is proved in the following that the system's performance can be dramatically changed by the perturbation. On one side of $\nu = 0$, the system has no bifurcation. And on the other side, the system exhibits the hysteresis phenomenon. Since the case for $n = 2$ is similar to the case $n \geq 3$, we only prove the result for $n \geq 3$. The center manifold of (3.7)–(6.1) and the reduced system satisfy

$$\begin{aligned} x_1 &= -\frac{a_z}{a_1} z - \frac{a_1 a_\mu - a_2 a_z}{a_1^2}\mu - \frac{1}{a_1}\nu + O(z,\mu)^2, \\ x_2 &= -\frac{a_z}{a_1}\mu + O(z,\mu)^2,\ x_i = O(z,\mu)^2 \text{ for } i \geq 3, \\ \dot{z} &= f_c(z,\mu,\nu) = \nu_1 + \nu_2 z + D_2(\mu,\nu)z^2 + Dz^3 + \cdots, \end{aligned}$$

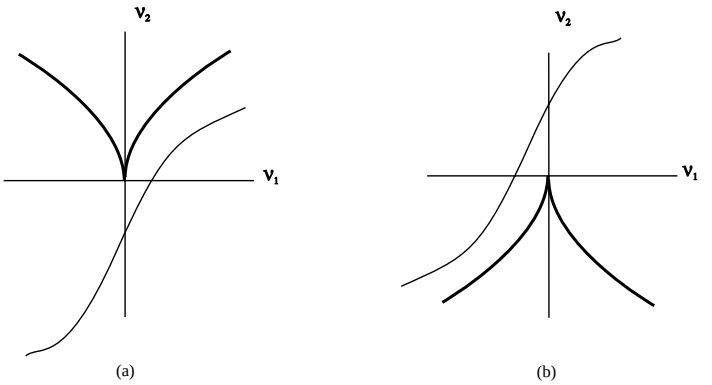


FIG. 6.2. *The figure of* $(\nu_1, \nu_2) = (D_1(\mu, \nu), D_2(\mu, \nu))$ *for a fixed value of* ν.

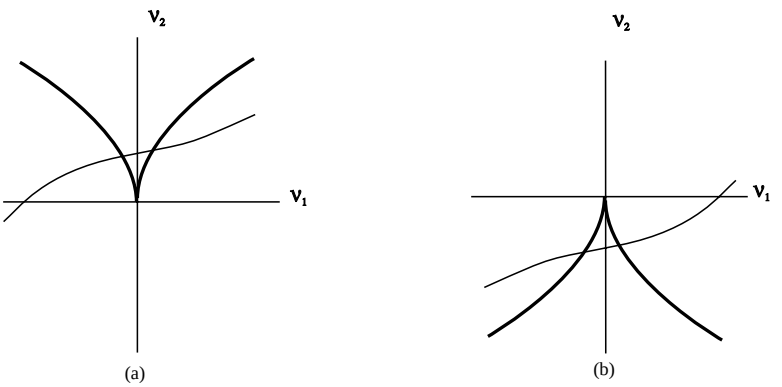


FIG. 6.3. *The figure of* $(\nu_1, \nu_2) = (D_1(\mu, \nu), D_2(\mu, \nu))$ *for a fixed value of* ν.

where D is defined by (4.23), and

$$
\begin{aligned}
\nu_1 &= \mu + O(\mu, \nu)^2, \\
\nu_2 &= \left(\frac{2\gamma_{x_1x_1} a_z - \gamma_{zx_1} a_1}{a_1^2} \left(a_\mu - \frac{a_2 a_z}{a_1} \right) - \frac{\gamma_{x_1\mu} a_z}{a_1} \right) \mu \\
&\quad + \frac{2a_z \gamma_{x_1x_1} - a_1 \gamma_{zx_1}}{a_1^2} \nu + \cdots.
\end{aligned} \tag{6.5}
$$

For any fixed value of ν, (6.5) represents a curve in the $\nu_1\nu_2$-plane. The bifurcation of (3.7)–(6.1) is determined by this curve and the cusp (5.17). If $(2a_z\gamma_{x_1x_1} - a_1\gamma_{zx_1})\nu$ and D have the same signs, then (6.5) and (5.17) have no intersection around the origin (see Figure 6.2). From the cusp bifurcation diagram (Figure 5.2), we know that no bifurcation occurs around the origin. If $(2a_z\gamma_{x_1x_1} - a_1\gamma_{zx_1})\nu$ and D have opposite sign, then the curve (6.5) meets with (5.17) at two different points (see Figure 6.3). Suppose that (6.5) intersects (5.17) at $\mu_1 < 0$ and $\mu_2 > 0$. Then the system has one equilibrium point for $\mu < \mu_1$ or $\mu > \mu_2$. There exist three equilibrium points if $\mu_1 < \mu < \mu_2$. The bifurcation diagram is shown in Figure 6.4, which is called a hysteresis [26]. We summarize the results in the following theorem.

THEOREM 6.2. *Consider* (3.7)–(6.1) *in which* $\alpha(z, x, \mu)$ *satisfies Assumption* 2.3 *and* (5.25). *Fix any value for* ν *around* $\nu = 0$, *and* (3.7)–(6.1) *has the following*

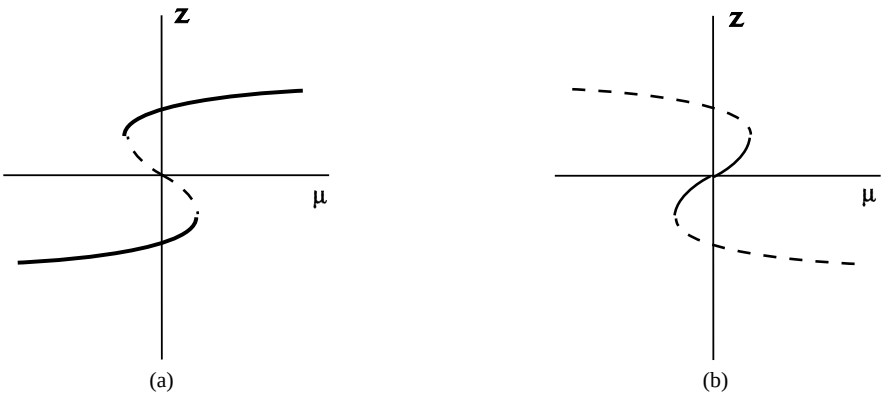


FIG. 6.4. *The hysteresis. In* (a), $D < 0$. *In* (b), $D > 0$.

properties.

(i) *If* $(2a_z\gamma_{x_1x_1} - a_1\gamma_{zx_1})D\nu > 0$, (3.7)–(4.21) *has no bifurcation around the origin and* $\mu = 0$. *There exists a neighborhood of* $(z, x, \mu) = (0, 0, 0)$ *in which the system has a unique closed-loop equilibrium point for every value of* μ. *Furthermore, the system is stable if* $D < 0$, *and the system is unstable if* $D > 0$.

(ii) *If* $(2a_z\gamma_{x_1x_1} - a_1\gamma_{zx_1})D\nu < 0$, *the system exhibits hysteresis. Its stability is determined by the value of* D *as shown in Figure* 6.4.

7. Conclusion. In this paper, the problem of bifurcation control by nonlinear state feedbacks is studied thoroughly. Based on normal forms and invariants, all bifurcations exhibited in the performance of nonlinear control systems with a zero uncontrollable mode are studied. Feedbacks can be designed based on the results to achieve the stability or to achieve the desired bifurcation pattern. It is proved in section 4 that systems having the normal form (3.6) exhibit a transcritical bifurcation or a pitchfork bifurcation. The transcritical bifurcation can be converted into a supercritical pitchfork bifurcation by using state feedbacks. In section 5, it is proved that systems with the normal form (3.7) have a saddle-node bifurcation. It can be locally removed by a C^0 state feedback, provided that the equilibrium set is a saddle. If the closed-loop system satisfies the cubic degeneracy condition $(Q_1(a_1, -a_z) = 0, D \neq 0)$, the saddle-node bifurcation is also locally removable by feedbacks. Its closed-loop equilibrium set can be embedded into the diagram of a cusp. Under feedbacks which are not zero at the critical point, it is shown in section 6 that a transcritical bifurcation of (3.6) is bifurcated into two saddle-node bifurcations. For the system (3.7) with cubic degeneracy, it has either a hysteresis or no bifurcation, depending on the value of the feedback at the critical point. All the conditions on bifurcations and their stability are characterized by invariants and the coefficients in the feedback. The results obtained in sections 4 and 5 draw a complete picture of bifurcations under smooth feedbacks which are zero at the critical point. This is summarized in Table 1.

The approach based on control system normal forms can certainly be used for the study of other bifurcations which are not addressed here. For instance, the normal form of control systems with a pair of imaginary uncontrollable modes or with a double zero uncontrollable mode are available [7]. The control of the Hopf bifurcation and the double zero bifurcation based normal form is part of our future research.

REFERENCES

[1] E. H. Abed and J.-H. Fu, *Local feedback stabilization and bifurcation control,* I–II. *Stationary bifurcation*, Systems Control Lett., 8 (1987), pp. 467–473.

[2] S. Behtash and S. Sastry, *Stabilization of nonlinear systems with uncontrollable linearization*, IEEE Trans. Automat. Control, 33 (1988), pp. 585–591.

[3] J. Carr, *Application of Center Manifold Theory*, Springer-Verlag, New York, 1981.

[4] G. Chen, *Control and Synchronization of Chaotic Systems (a Bibliography)*, ECE Dept., University of Houston, TX. Available via ftp from ftp.egr.uh.edu/pub/TeX/chaos.tex.

[5] G. Chen and J. L. Moiola, *An overview of bifurcation, chaos and nonlinear dynamics in control systems*, J. Franklin Inst. B, 331 (1994), pp. 819–858.

[6] K. M. Eveker and C. N. Nett, *Control of compression system surge and rotating stall: A laboratory-based "hands-on" introduction*, Proceedings of the American Control Conference, Baltimore, MD, International Federation of Automatic Control, 1994, pp. 1307–1311.

[7] O. E. Fitch, *The Control of Bifurcations with Engineering Applications*, Ph.D. Dissertation, Naval Postgraduate School, Monterey, CA, 1997.

[8] R. Genesio, A. Tesi, H. O. Wang, and E. H. Abed, *Control of period doubling bifurcations using harmonic balance*, Proceedings of the IEEE Conference on Decision and Control, San Antonio, Texas, 1993, pp. 492–497.

[9] P. Glendinning, *Stability, Instability and Chaos: An Introduction to the Theory of Nonlinear Differential Equations*, Cambridge University Press, Cambridge, UK, 1994.

[10] J. Hale and H. Koçak, *Dynamics and Bifurcations*, Springer-Verlag, New York, 1991.

[11] G. Gu, X. Chen, A. Sparks, and S. Banda, *Bifurcation Stabilization with Local Output Feedback*, Department of ECE, Louisiana State University, Baton Rouge, LA, preprint.

[12] W. Kang and A. J. Krener, *Extended quadratic controller normal form and dynamic state feedback linearization of nonlinear systems*, SIAM J. Control Optim., 30 (1992), pp. 1319–1337.

[13] W. Kang, *Extended Controller Normal Form, Invariants and Dynamic Feedback Linearization of Nonlinear Control Systems*, Ph.D. Dissertation, University of California at Davis, 1991.

[14] W. Kang, *Bifurcation and normal form of nonlinear control systems—part* I, SIAM J. Control Optim., 36 (1998), pp. 193–212.

[15] W. Kang, *Bifurcation and normal form of nonlinear control systems—part* II, SIAM J. Control Optim., 36 (1988), pp. 213–232.

[16] W. Kang, *Bifurcation and topology of equilibrium sets*, Proceedings of the IEEE Conference on Decision and Control, Kobe, Japan, 1996, pp. 2151–2155.

[17] W. Kang and F. Papoulias, *Bifurcation and normal forms of dive plane reversal of submersible vehicles*, Proceedings of the International Offshore and Polar Engineering Conference, Honolulu, Hawaii, International Society of Offshore and Polar Engineers, 1997, pp. 62–68.

[18] A. J. Krener, *The Feedbacks which Soften the Primary Bifurcation of MG* 3, Department of Mathematics, University of California at Davis, Davis, CA, preprint.

[19] M. Krstic, J. M. Protz, J. D. Paduano, and P. V. Kokotovic, *Backstepping designs for jet engine stall and surge control*, Proceedings of the IEEE Conference on Decision and Control, New Orleans, LA, 1995, pp. 3049–3055.

[20] D.-C. Liaw and E. H. Abed, *Stability analysis and control of rotating stall*, Proceedings of the International Federation of Automatic Control Nonlinear Control Systems Design Symposium, Bordeaux, France, 1992.

[21] M. A. Pinsky and B. Essary, *Normal forms, averaging and resonance control of flexible structures*, J. Dynam. Systems, Measurement and Control, 116, N3 (1994), pp. 357–366.

[22] M. A. Pinsky and I. Shumyatsky, *Feedback stabilization of bifurcation phenomena and its application to the control of voltage instabilities and collapse*, in Proceedings of the International Federation of Automatic Control Nonlinear Control Systems Design Symposium, Lake Tahoe, CA, 1995, pp. 58–63.

[23] F. E. McCaughan, *Bifurcation analysis of axial flow compressor stability*, SIAM J. Appl. Math., 50 (1990), pp. 1232–1253.

[24] F. K. Moore and E. M. Greitzer, *A theory of post-stall transients in axial compression systems–part* I*: Development of equations*, ASME J. Engineering for Gas Turbines and Power, 108 (1986), pp. 68–76.

[25] H. O. Wang and E. H. Abed, *Bifurcation control of a chaotic system*, Automatica J. IFAC, 31 (1995), pp. 1213–1226.

[26] S. Wiggins, *Introduction to Applied Nonlinear Dynamical Systems and Chaos*, Springer-Verlag, New York, 1990.

SIAM J. CONTROL OPTIM.
Vol. 38, No. 5, pp. 1453–1467

SPECTRAL ANALYSIS OF FOKKER–PLANCK AND RELATED OPERATORS ARISING FROM LINEAR STOCHASTIC DIFFERENTIAL EQUATIONS*

DANIEL LIBERZON† AND ROGER W. BROCKETT‡

Abstract. We study spectral properties of certain families of linear second-order differential operators arising from linear stochastic differential equations. We construct a basis in the Hilbert space of square-integrable functions using modified Hermite polynomials, and obtain a representation for these operators from which their eigenvalues and eigenfunctions can be computed. In particular, we completely describe the spectrum of the Fokker–Planck operator on an appropriate invariant subspace of rapidly decaying functions. The eigenvalues of the Fokker–Planck operator provide information about the speed of convergence of the corresponding probability distribution to steady state, which is important for stochastic estimation and control applications. We show that the operator families under consideration can be realized as solutions of differential equations in the double bracket form on an operator Lie algebra, which leads to a simple expression for the flow of their eigenfunctions.

Key words. modified Hermite polynomials, linear stochastic differential equation, Fokker–Planck operator, double bracket equation

AMS subject classifications. 35P05, 35Q58, 93E03

PII. S0363012998338193

1. Introduction.

Given a system of stochastic differential equations, one can associate with it a (deterministic) partial differential equation which describes the evolution of the probability density with time. This so-called *Fokker–Planck equation* takes the form

$$\frac{\partial \rho}{\partial t} = L\rho, \tag{1}$$

where L is a second-order linear differential operator known as the *Fokker–Planck operator*. If $g_0, g_1, g_2, \ldots$ are the eigenfunctions of L corresponding to distinct eigenvalues $\lambda_0, \lambda_1, \lambda_2, \ldots$, then the solution of (1) with initial condition

$$\rho(0,x) = \sum_{i=0}^{\infty} \alpha_i g_i(x), \qquad \alpha_i \in \mathbb{R},$$

is given by

$$\rho(t,x) = \sum_{i=0}^{\infty} \alpha_i e^{\lambda_i t} g_i(x).$$

*Received by the editors May 4, 1998; accepted for publication (in revised form) October 18, 1999; published electronically May 11, 2000.
http://www.siam.org/journals/sicon/38-5/33819.html

†Department of Electrical Engineering, Yale University, New Haven, CT 06520-8267 (liberzon@sysc.eng.yale.edu). The research of this author was supported in part by ARO grant DAAH04-95-1-0114, NSF grant ECS 9634146, and AFOSR grant F49620-97-1-0108.

‡Division of Engineering and Applied Sciences, Harvard University, Cambridge, MA 02138 (brockett@hrl.harvard.edu). The research of this author was supported in part by Army DAAH 04-96-1-0445, Army DAAG 55-97-1-0114, and Army DAAL 03-92-G-0115.

Thus the eigenvalues of the Fokker–Planck operator L, particularly the one with the smallest magnitude, provide information about the speed of convergence of the probability distribution to steady state (when one exists), which is important in stochastic filtering and control applications. For a discussion along these lines and examples, see [3].

In the paper by Holley, Kusuoka, and Stroock [14], and more recently in [7], [8], [9], spectral properties of Fokker–Planck operators associated with certain types of nonlinear stochastic systems were investigated with the view towards applications to function minimization procedures. In this paper we confine our attention to Fokker–Planck operators that correspond to *linear* stochastic differential equations. An understanding of their spectral properties, besides being of interest in its own right, under certain circumstances helps shed some light on the nonlinear case (see [16, p. 88]). As in [14], we consider these operators as acting on a dense subspace of $L^2(\mathbb{R}^n)$ (rather than $L^1(\mathbb{R}^n)$ which might seem more natural from the probabilistic point of view). We apply a standard gauge transformation technique to convert them to self-adjoint operators, which greatly facilitates the analysis.

The paper starts with the one-dimensional case. Motivated by the explicit form of the steady-state probability density, we modify the classical Hermite polynomials by introducing one additional parameter σ (in our context, σ corresponds to the steady-state variance). This construction leads to an orthonormal basis for $L^2(\mathbb{R})$ with respect to which the operators under consideration take a particularly transparent form. The representation thus obtained allows us to compute their eigenvalues and eigenfunctions directly. As a result, we are able to provide a complete description of the spectrum of the Fokker–Planck operator on an appropriate invariant subspace of rapidly decaying functions. We then show that the essential features of this analysis carry over to the multidimensional case and enable us to obtain information about eigenvalues of Fokker–Planck operators in a more general setting.

Moreover, we observe that the operator families parameterized by σ can be described by differential equations on an operator Lie algebra which take the so-called *double bracket* form $\frac{dL}{d\sigma} = [L,[L,M]]$. This leads to a simple expression for the flow of the corresponding eigenfunctions. The study of differential equations in the double bracket form on finite-dimensional Lie algebras was initiated in [2] and [6] in connection with integrable gradient flows and numerical algorithms. It was shown, in particular, that such equations give rise to isospectral flows. In this paper we present what seems to be a new framework in which double bracket equations appear. The corresponding flows on an operator Lie algebra preserve the eigenvalues (actually, the entire spectrum in the self-adjoint case). This property is supported by probabilistic intuition.

The paper is organized as follows. In section 2 we construct an orthonormal basis in $L^2(\mathbb{R})$ using modified Hermite polynomials. In section 3 we study second-order differential operators arising from scalar linear stochastic differential equations. In section 4 we treat the multidimensional case, giving generalizations of the previous results. In section 5 we discuss double bracket differential equations on an operator Lie algebra and indicate connections with some known results on completely integrable gradient flows.

2. Orthonormal bases in $L^2(\mathbb{R})$. It is well known (see, e.g., [15, p. 121]) that the *Hermite functions*

$$u_k(x) = h_k(x)e^{-x^2/2}, \qquad k = 0, 1, \ldots, \tag{2}$$

where $h_k(x) = \frac{1}{\pi^{1/4}\sqrt{2^k k!}}\, e^{x^2}\, \frac{d^k e^{-x^2}}{dx^k}$ are the *Hermite polynomials*, form an orthonormal basis for $L^2(\mathbb{R})$. We consider here the *modified Hermite polynomials*

$$h_k(x,\sigma) := \frac{\sqrt{\sigma^k}}{(\sigma\pi)^{1/4}\sqrt{2^k k!}} e^{x^2/\sigma} \frac{d^k e^{-x^2/\sigma}}{dx^k}\,, \qquad k = 0, 1, \ldots,$$

where $\sigma > 0$ is a real parameter, and introduce the *modified Hermite functions*

$$u_k(x,\sigma) := h_k(x,\sigma) e^{-x^2/2\sigma} = c_k(\sigma) e^{x^2/2\sigma} \frac{d^k e^{-x^2/\sigma}}{dx^k} \tag{3}$$

with constants $c_k(\sigma)$ given by the relations

$$c_k(\sigma) = \frac{\sqrt{\sigma^k}}{(\sigma\pi)^{1/4}\sqrt{2^k k!}}. \tag{4}$$

The functions (3) reduce to those given by (2) for $\sigma = 1$. Various modifications of the classical Hermite polynomials, analogous to (and more general than) the one considered here, can be found in the literature [10], [13].

LEMMA 1. *For any $\sigma > 0$, the functions* (3) *form an orthonormal basis for $L^2(\mathbb{R})$.*

Proof. We have

$$\begin{aligned}\langle u_k(x,\sigma), u_l(x,\sigma)\rangle &= c_k(\sigma)c_l(\sigma) \int_{-\infty}^{\infty} e^{x^2/\sigma} \frac{d^k e^{-x^2/\sigma}}{dx^k} \frac{d^l e^{-x^2/\sigma}}{dx^l} dx \\ &= c_k(1)c_l(1) \int_{-\infty}^{\infty} e^{y^2} \frac{d^k e^{-y^2}}{dy^k} \frac{d^l e^{-y^2}}{dy^l} dy,\end{aligned}$$

where we have made the change of variable $x = \sqrt{\sigma}\, y$. The statement of the lemma follows from the fact that the Hermite functions (2) form an orthonormal basis for $L^2(\mathbb{R})$. □

3. Fokker–Planck operators in $L^2(\mathbb{R})$. Let us consider the linear stochastic differential equation in the Itô sense

$$dx = -ax\, dt + b\, dw, \qquad a > 0, \tag{5}$$

where $x \in \mathbb{R}$ and w is a standard Wiener process. The reader may consult [11] for basic concepts of the theory of stochastic differential equations. The equation for the steady-state probability density that corresponds to (5) is $L(a,b)\rho(x) = 0$, where

$$L(a,b)\rho := \frac{b^2}{2}\rho_{xx} + ax\rho_x + a\rho \tag{6}$$

and ρ_x and ρ_{xx} denote the first and the second derivatives of ρ, respectively. The operator $L(a,b)$ is the *Fokker–Planck operator* associated with (5). Define

$$\sigma = \frac{b^2}{2a}. \tag{7}$$

The steady-state probability density is then given by $\bar{\rho}(x) = Ne^{-x^2/2\sigma}$, where $N > 0$ is a normalization constant. Dividing the Fokker–Planck operator $L(a,b)$ by a, we are led to studying a one-parameter family of differential operators, L_σ, defined by

$$L_\sigma\rho := \frac{1}{a} L(a, \sqrt{2a\sigma})\rho = \sigma\rho_{xx} + x\rho_x + \rho\,, \qquad \sigma > 0. \tag{8}$$

Before proceeding, we need to specify the domain of the above operators. It is easy to see that $L_\sigma u_k(x,\sigma) \in L^2(\mathbb{R})$ for each k, and L_σ is well defined by the formula (8) on the dense subspace U of $L^2(\mathbb{R})$ consisting of finite linear combinations of the functions $u_k(x,\sigma)$. We then define L_σ to be the minimal closed linear operator in $L^2(\mathbb{R})$ such that $L_\sigma\rho$ is given by (8) whenever $\rho \in C^2(\mathbb{R}) \cap L^2(\mathbb{R})$ and $\sigma\rho_{xx} + x\rho_x + \rho \in L^2(\mathbb{R})$. We thus obtain an operator $L_\sigma : \mathcal{D}_{L_\sigma} \to L^2(\mathbb{R})$, where $\mathcal{D}_{L_\sigma}$ is a dense subspace of $L^2(\mathbb{R})$ that contains U. Throughout the paper, unless specified otherwise, all differential operators are to be interpreted in the above sense.[1] For details on defining differential operators in this way, see [12].

The analysis of the operators L_σ is complicated by the fact that they are not self-adjoint. There is a standard technique which allows one to convert these operators to self-adjoint ones (this is sometimes referred to as *gauge*, or *ground state, transformation*). In our case, write $\rho = vf$, where the function v is to be fixed. We have

$$L_\sigma(vf) = \sigma v_{xx} f + 2\sigma v_x f_x + \sigma v f_{xx} + x v_x f + x v f_x + v f.$$

We see that in order for the first-order derivatives to disappear, v must satisfy the equation $v_x = -\dfrac{x}{2\sigma} v$. Letting

$$v = e^{-x^2/4\sigma} \tag{9}$$

we obtain $v^{-1} L_\sigma(vf) = \sigma f_{xx} + (\frac{1}{2} - \frac{x^2}{4\sigma}) f$.

Motivated by the above discussion, we define a new operator family, T_σ, by the formula

$$T_\sigma \rho := \sigma\rho_{xx} + \Big(\frac{1}{2} - \frac{x^2}{4\sigma}\Big)\rho, \qquad \sigma > 0. \tag{10}$$

For any positive σ, the operator T_σ is closed and self-adjoint, its domain being a dense subspace $\mathcal{D}_{T_\sigma}$ of $L^2(\mathbb{R})$ (defined as explained before).

We know that $L_\sigma u_0(x,\sigma) = L_\sigma c_0 e^{-x^2/2\sigma} = 0$, i.e., $e^{-x^2/2\sigma}$ is an eigenfunction with the eigenvalue zero. To investigate the spectral properties of the operators L_σ and T_σ, it seems natural to use the basis given by the modified Hermite functions (3) (with the same value of σ). We first carry out direct calculations for the family L_σ, setting the stage for the multidimensional case. We will then see that the analysis of the self-adjoint operators T_σ is more straightforward and allows one to obtain precise information about the spectrum of the original Fokker–Planck operator on an appropriate space of rapidly decaying functions.

PROPOSITION 2. *The spectrum of the operator $L_\sigma : \mathcal{D}_{L_\sigma} \to L^2(\mathbb{R})$ is independent of σ. For any $\sigma > 0$, the eigenvalues of L_σ are all numbers in the half-plane $\{\lambda \in \mathbb{C} : \operatorname{Re}\lambda < 1/2\}$.*

[1] Alternatively, $C^2(\mathbb{R})$ here could be replaced by the space of functions $\rho : \mathbb{R} \to \mathbb{R}$ such that ρ_x exists and is absolutely continuous, i.e., the space of twice weakly differentiable functions for which the differential expression (8) is defined almost everywhere.

Proof.[2] Straightforward computations give

$$L_\sigma u_k(x,\sigma) = c_k(\sigma)\left[\left(2+\frac{2x^2}{\sigma}\right)e^{x^2/2\sigma}\frac{d^k e^{-x^2/\sigma}}{dx^k} + 3xe^{x^2/2\sigma}\frac{d^{k+1}e^{-x^2/\sigma}}{dx^{k+1}} + \sigma e^{x^2/2\sigma}\frac{d^{k+2}e^{-x^2/\sigma}}{dx^{k+2}}\right]. \tag{11}$$

We introduce the notation $d_k(x,\sigma) = e^{x^2/2\sigma}\dfrac{d^k e^{-x^2/\sigma}}{dx^k}$, so that (11) becomes

$$L_\sigma u_k = c_k\left(\left(2+\frac{2x^2}{\sigma}\right)d_k + 3xd_{k+1} + \sigma d_{k+2}\right). \tag{12}$$

To obtain recurrence relations on d_k, notice that by Newton's binomial formula we have

$$d_{k+1} = e^{x^2/2\sigma}\left(-\frac{2x}{\sigma}\frac{d^k e^{-x^2/\sigma}}{dx^k} - \frac{2k}{\sigma}\frac{d^{k-1}e^{-x^2/\sigma}}{dx^{k-1}}\right)$$

which in the new notation becomes

$$d_{k+1} = -\frac{2x}{\sigma}d_k - \frac{2k}{\sigma}d_{k-1}\,. \tag{13}$$

From (13) we obtain

$$xd_k = -\frac{\sigma}{2}d_{k+1} - kd_{k-1}, \qquad k = 1,2,\ldots, \tag{14}$$

and also (multiplying both sides of (13) by x and then using (14))

$$\frac{2x^2}{\sigma}d_k = (2k+1)d_k + \frac{\sigma}{2}d_{k+2} + \frac{2k(k-1)}{\sigma}d_{k-2}. \tag{15}$$

Combining (12)–(15) gives

$$L_\sigma u_k = c_k\left(-kd_k + \frac{2k(k-1)}{\sigma}d_{k-2}\right) \tag{16}$$

and we see that the terms containing d_{k+2} disappear. Moreover, notice that we have

$$\frac{2k(k-1)}{\sigma}\,c_k = \sqrt{k(k-1)}\,c_{k-2}. \tag{17}$$

The formulas (16) and (17) imply that with respect to the basis (3) the operator L_σ takes the upper triangular form as given by

$$L_\sigma u_k(x,\sigma) = -ku_k(x,\sigma) + \sqrt{k(k-1)}u_{k-2}(x,\sigma). \tag{18}$$

From (18) it immediately follows that the spectrum of L_σ is independent of σ. Moreover, it is easy to see that the nonpositive integers are eigenvalues of L_σ. The

[2]The proofs in this section are given in sketched form; full details can be found in [16].

corresponding eigenfunctions are finite linear combinations of the basis elements u_k and thus belong to $C^\infty(\mathbb{R}) \cap L^2(\mathbb{R})$. They do not, however, form a complete set of eigenfunctions. The formula (18) implies that the existence of an eigenfunction of L_σ with an eigenvalue λ is equivalent to the convergence of at least one of the series

$$\sum_{n=1}^{\infty} \frac{\lambda^2(\lambda+2)^2 \cdots (\lambda+2n-2)^2}{(2n)!}$$

and

$$\sum_{n=1}^{\infty} \frac{(\lambda+1)^2(\lambda+3)^2 \cdots (\lambda+2n-1)^2}{(2n+1)!}.$$

Using Gauss' test for convergence (see, e.g., [18]), one can show in a straightforward manner that each series converges if $\mathrm{Re}\lambda < 1/2$ and diverges if $\mathrm{Re}\lambda \geq 1/2$. □

We can gain more insight into the spectral properties of the operator L_σ from its probabilistic interpretation. Recall that L_σ was defined in terms of the Fokker–Planck operator $L(a,b)$ via the formula (8). It follows from Proposition 2 that the eigenvalues of $L(a,b)$ are all numbers in the half-plane $\{\lambda \in \mathbb{C} : \mathrm{Re}\,\lambda < a/2\}$. (However, it can be deduced from (1) that any eigenfunction of $L(a,b)$ that is nonnegative and belongs to $L^1(\mathbb{R})$ must be proportional to the steady-state probability density, which corresponds to the eigenvalue zero.) The fact that the spectrum of $L(a,b)$ does not depend on the noise coefficient b should not be surprising if we notice that we can change b by simply rescaling x; i.e., substituting $y = px$ in (5) for an arbitrary $p \in \mathbb{R}$ gives $\dot{y} = -ay + pb\dot{w}$. It is easy to check that the spectrum of the Fokker–Planck operator associated with (5) is not affected by such changes of variable.

We now turn our attention to the family of self-adjoint operators T_σ defined by (10). It is well known that Hermite polynomials appear frequently in expressions for eigenfunctions of self-adjoint linear second-order differential operators. The next proposition shows that the eigenfunctions of T_σ are given by the modified Hermite functions (3) and is to be considered as a preparation for a more general result to be presented in the next section. For $\sigma = 1/2$, the statement reduces to a standard result involving the classical Hermite functions (see, e.g., [1, p. 256]).

PROPOSITION 3. *For any $\sigma > 0$, the spectrum of the operator $T_\sigma : \mathcal{D}_{T_\sigma} \to L^2(\mathbb{R})$ consists of the nonpositive integers, all of which are eigenvalues. The corresponding eigenfunctions are the functions $u_k(x, 2\sigma)$, i.e., $T_\sigma u_k(x, 2\sigma) = -k\, u_k(x, 2\sigma)$.*

Proof. For $\rho = e^{x^2/4\sigma} \frac{d^k e^{-x^2/2\sigma}}{dx^k}$ one can verify that

$$T_\sigma \rho = e^{x^2/4\sigma} \frac{d^k e^{-x^2/2\sigma}}{dx^k} + x e^{x^2/4\sigma} \frac{d^{k+1} e^{-x^2/2\sigma}}{dx^{k+1}} + \sigma e^{x^2/4\sigma} \frac{d^{k+2} e^{-x^2/2\sigma}}{dx^{k+2}}$$

which in our previous notation becomes

$$T_\sigma d_k(x, 2\sigma) = d_k(x, 2\sigma) + x d_{k+1}(x, 2\sigma) + \sigma d_{k+2}(x, 2\sigma). \tag{19}$$

Replacing σ by 2σ in (14) and substituting into (19), we arrive at

$$T_\sigma d_k(x, 2\sigma) = -k\, d_k(x, 2\sigma).$$

This immediately implies the second part of the statement. The first part of the statement follows from this, since we have found an orthonormal basis in $L^2(\mathbb{R})$ consisting of eigenfunctions of T_σ. □

As a consequence, the eigenvalues of the original operator L_σ restricted to the space of functions of the form $\rho = vf$, where v is given by (9) and $f \in \mathcal{D}_{T_\sigma}$, are the nonpositive integers. The eigenfunction that corresponds to the eigenvalue $-k$ is given by $e^{-x^2/4\sigma}u_k(x, 2\sigma) = c_k(2\sigma)\frac{d^k e^{-x^2/2\sigma}}{dx^k}$. This leads us to a complete characterization of the spectrum of the Fokker–Planck operator $L(a, b)$ restricted to an appropriate space of rapidly decaying functions. Namely, let us denote by $\mathcal{L}_\sigma$ the space of functions that can be represented by finite linear combinations of the form $\sum_{k=1}^m \alpha_k \frac{d^k e^{-x^2/2\sigma}}{dx^k}$, $\alpha_k \in \mathbb{R}$. From the definitions of L_σ and T_σ and from Proposition 3 we immediately obtain the following result.

COROLLARY 4. *The space $\mathcal{L}_\sigma$ is invariant with respect to the Fokker–Planck operator $L(a, b)$ associated with* (5). *The spectrum of the restriction of $L(a, b)$ to $\mathcal{L}_\sigma$ consists of the numbers* 0, $-a$, $-2a$, $-3a$, $\ldots$, *all of which are eigenvalues.*

Remark 1. The eigenfunctions of L_σ on $\mathcal{L}_\sigma$ found above form an orthonormal basis for the space $L^2(\mathbb{R}, e^{x^2/2\sigma}dx)$, on which the operator L_σ can be shown to be self-adjoint. If instead of $\mathcal{L}_\sigma$ we consider a dense subspace of $L^2(\mathbb{R}, e^{x^2/2\sigma}dx)$ containing $\mathcal{L}_\sigma$, which can be constructed as explained at the beginning of the section, the statement about the spectrum still applies. Clearly, this larger subspace is no longer invariant under the action of $L(a, b)$. The operator T_σ is convenient because it is self-adjoint with respect to the standard inner product on $L^2(\mathbb{R})$.

We see in view of (7) that as the value of a increases while the noise coefficient b stays constant, the rate of decay of functions in $\mathcal{L}_\sigma$ becomes more rapid and so does the convergence to steady state. If we fix one member of the family $\{T_\sigma : \sigma > 0\}$, say, $T_{1/2}$, then for any value of σ the operator T_σ can be expressed as $T_\sigma = \Theta_\sigma^{-1} T_{1/2} \Theta_\sigma$, where Θ_σ is the unitary operator defined by $\Theta_\sigma u_k(x, 2\sigma) = u_k(x, 1) = u_k(x)$. We will use this observation in section 5.

4. Fokker–Planck operators in $L^2(\mathbb{R}^n)$. Consider the system of linear stochastic differential equations

$$dx = Ax\,dt + B\,dw\,, \qquad x \in \mathbb{R}^n, \tag{20}$$

where w is a standard m-dimensional Wiener process and A and B are matrices of suitable dimensions. Recall that *separable functions*, i.e., functions that can be expressed as products $\rho_1(x_1)\cdots\rho_n(x_n)$, span a dense subspace of $L^2(\mathbb{R}^n)$. Thus we can construct an orthonormal basis for $L^2(\mathbb{R}^n)$ by taking products of the modified Hermite functions (3) for each variable. The analysis of the previous section now directly generalizes to those linear stochastic systems in $\mathbb{R}^n$ whose equations are completely decoupled. In this case, the Fokker–Planck operator decomposes into a sum of Fokker–Planck operators of the kind considered above for each variable. Our earlier results then imply, in particular, that the sums of the eigenvalues of the matrix A are eigenvalues of the corresponding Fokker–Planck operator, and that the corresponding eigenfunctions belong to the space $C^\infty(\mathbb{R}^n) \cap L^2(\mathbb{R}^n)$ and can be explicitly constructed.

Although the analysis for the general multidimensional system (20) is more complicated than in the scalar case, results that parallel most of our earlier developments can be obtained. Let us denote the Fokker–Planck operator associated with (20) by L_n and consider it as being a closed operator defined on a dense subspace $\mathcal{D}_{L_n}$ of

$L^2(\mathbb{R}^n)$ (cf. section 3). We have the following expression for L_n:

$$L_n\rho = \frac{1}{2}\sum_{i,j=1}^{n}(BB^T)_{ij}\,\rho_{x_ix_j} - \sum_{i,j=1}^{n}A_{ij}x_j\rho_{x_i} - \mathrm{tr}A\cdot\rho. \tag{21}$$

From this point on, let us make the following two assumptions with regard to the system (20):

(a) The eigenvalues of A have negative real parts.

(b) (A,B) is a controllable pair (i.e., $\mathrm{rank}(B, AB, \ldots, A^{n-1}B) = n$).

Under these assumptions, the steady-state variance equation

$$AQ + QA^T + BB^T = 0 \tag{22}$$

associated with (20) has a positive definite symmetric solution Q. After an appropriate change of coordinates in $\mathbb{R}^n$ we can have $Q = \frac{1}{2}I$, so that $A = \Omega - BB^T$ with Ω skew-symmetric. Such a coordinate transformation does not change the eigenvalues of the Fokker–Planck operator L_n. The steady-state probability density then becomes $\bar\rho(x) = Ne^{-x^Tx}$, $N > 0$, and this is an eigenfunction that corresponds to the eigenvalue zero of the Fokker–Planck operator.

Next let us determine all eigenfunctions of L_n that take the form

$$\rho(x) = (h_1x_1 + \cdots + h_nx_n)\bar\rho(x) = h^Tx\bar\rho(x), \qquad h \in \mathbb{R}^n. \tag{23}$$

LEMMA 5. *Suppose that $A = \Omega - BB^T$, where $\Omega = -\Omega^T$. Then the function (23) is an eigenfunction of the operator L_n with eigenvalue λ if and only if h is an eigenvector of the matrix A with the same eigenvalue λ.*

Proof. Let ρ be of the form (23). Taking into account that $L_n\bar\rho = 0$, we have

$$\begin{aligned} L_n\rho &= \sum_{i,j=1}^{n}(BB^T)_{ij}(h)_i(-2x_j)\bar\rho - \sum_{i,j}A_{ij}x_j(h)_i\bar\rho \\ &= -\bar\rho\sum_{i,j=1}^{n}(A_{ji} + 2(BB^T)_{ji})(h)_jx_i \\ &= \bar\rho\sum_{i,j=1}^{n}(\Omega_{ij} - (BB^T)_{ij})(h)_jx_i = \sum_{i=1}^{n}(Ah)_ix_i\bar\rho \end{aligned}$$

and this obviously equals $\lambda\rho = \lambda\sum_i h_ix_i\bar\rho$ if and only if $Ah = \lambda h$. ∎

Denote by $h_1, \ldots, h_k$ the eigenvectors of A and by $\lambda_1, \ldots, \lambda_k$ the corresponding eigenvalues ($k \le n$). Now let us see how L_n acts on functions of the form

$$\rho(x) = \bar\rho(x)\prod_{m\in J}h_m^Tx, \tag{24}$$

where the product is taken over some index set J whose elements are (not necessarily distinct) positive integers no greater than n. Using Lemma 5 and the fact that

$L_n\bar\rho = 0$, we have

$$\begin{aligned} L_n(\bar\rho \textstyle\prod_{m\in J} h_m^T x) &= \sum_{i,j=1}^{n} \sum_{m,l\in J} (BB^T)_{ij}(h_m)_i(h_l)_j \bar\rho \prod_{p\in J\setminus\{m,l\}} h_p^T x \\ &\quad + \sum_{i,j=1}^{n} \sum_{m\in J} (BB^T)_{ij}(h_m)_i \bar\rho_{x_j} \prod_{p\in J\setminus\{m\}} h_p^T x \\ &\quad - \sum_{i,j=1}^{n} \sum_{m\in J} A_{ij}x_j(h_m)_i \bar\rho \prod_{p\in J\setminus\{m\}} h_p^T x \\ &= \sum_{i,j=1}^{n} \sum_{m,l\in J} (BB^T)_{ij}(h_m)_i(h_l)_j \bar\rho \prod_{p\in J\setminus\{m,l\}} h_p^T x + \Big(\sum_{m\in J} \lambda_m\Big)\bar\rho \prod_{m\in J} h_m^T x. \end{aligned}$$

Thus functions of the form (24) for various index sets J form an invariant subspace under the action of L_n. It is not hard to see that $\sum_{m\in J}\lambda_m$ are eigenvalues of L_n. The corresponding eigenfunctions are finite linear combinations of functions of the form (24). Summarizing, we have the following theorem.

THEOREM 6. *The sums of the eigenvalues of the matrix A are eigenvalues of the Fokker–Planck operator $L_n : \mathcal{D}_{L_n} \to L^2(\mathbb{R}^n)$.*

Theorem 6 can probably be best appreciated in the following context. It is well known and easy to show that there are $N_n^p = \binom{n+p-1}{p}$ linearly independent monomials of degree p in n variables of the form $x_1^{p_1}\dots x_n^{p_n}$, where $\sum_{i=1}^n p_i = p$ and $p_i \ge 0$. The linear differential equation

$$\dot x = Ax, \qquad x \in \mathbb{R}^n,$$

gives rise to the equation

$$\frac{d}{dt}x^{[p]} = A_{[p]}x^{[p]}, \qquad x^{[p]} \in \mathbb{R}^{N_n^p}.$$

One of the basic properties of the matrix $A_{[p]}$ defined in this way is that its eigenvalues are the p-term sums of the eigenvalues of A. As is shown in [4], the matrices $A_{[p]}$ are directly related to the pth moment equations for the system (20).

Theorem 6 shows that the situation in the infinite-dimensional case is consistent with the one described in the previous paragraph in the following sense. Associated with the system (20) we have the Fokker–Planck equation for the probability density

$$\frac{\partial\rho(t,x)}{\partial t} = L_n\rho(t,x).$$

The operator L_n is well defined on a dense subspace of $L^2(\mathbb{R}^n)$. We know that the basis elements in $L^2(\mathbb{R}^n)$ can be taken to be polynomials of an arbitrary degree multiplied by Gaussians, and we have shown that the sums (with an arbitrary number of terms) of the eigenvalues of A are eigenvalues of the operator L_n.

In view of the results of section 3, it would be interesting to obtain conditions under which it is possible to convert the Fokker–Planck operator L_n to a self-adjoint operator by means of an appropriate gauge transformation. The following result provides such conditions, as well as an explicit formula for the function v to be used.

PROPOSITION 7. *Suppose that the matrix B is nondegenerate and that we have*

$$ABB^T = BB^TA^T. \tag{25}$$

If the function v is defined by the formula

$$v = e^{x^T (BB^T)^{-1} Ax/2}, \tag{26}$$

then the operator T_n given by

$$T_n \rho = v^{-1} L_n (v\rho)$$

is self-adjoint.

Proof. The first-order terms in the expression for T_n are

$$\frac{1}{2} \sum_{j,k=1}^{n} (BB^T)_{jk} \left(v_{x_j} \frac{\partial}{\partial x_k} + v_{x_k} \frac{\partial}{\partial x_j} \right) - \sum_{i,j=1}^{n} A_{ij} x_j v \frac{\partial}{\partial x_i}.$$

We see that the coefficient of $\dfrac{\partial}{\partial x_i}$ is

$$\sum_{j=1}^{n} (BB^T)_{ij} v_{x_j} - \sum_{j=1}^{n} A_{ij} x_j v$$

and we need this to be zero for each i. This is equivalent to having

$$(BB^T) \operatorname{grad} v = Ax\, v$$

or

$$\operatorname{grad} v = (BB^T)^{-1} Ax\, v. \tag{27}$$

Therefore, we must have

$$\begin{aligned} v_{x_i x_j} &= \frac{\partial}{\partial x_i} \left[\sum_{k=1}^{n} ((BB^T)^{-1} A)_{jk} x_k\, v \right] \\ &= ((BB^T)^{-1} A)_{ji}\, v + \sum_{k=1}^{n} ((BB^T)^{-1} A)_{jk} x_k\, v_{x_i} \\ &= ((BB^T)^{-1} A)_{ji}\, v + \sum_{k=1}^{n} ((BB^T)^{-1} A)_{jk} x_k \sum_{l=1}^{n} ((BB^T)^{-1} A)_{il} x_l\, v. \end{aligned}$$

The compatibility conditions $v_{x_i x_j} = v_{x_j x_i}$ now imply that the matrix $(BB^T)^{-1} A$ has to be symmetric:

$$(BB^T)^{-1} A = A^T (BB^T)^{-1}.$$

Multiplying both sides of this formula by BB^T, we arrive at (25). It is straightforward to show that the function given by (26) satisfies (27). ∎

Let us switch to coordinates in which $Q = \sigma I$ for some $\sigma > 0$ (in the language of statistical thermodynamics, these are coordinates in which the *equipartition of energy* property holds, and σ is the steady-state *temperature* of the system). Then $(A + A^T)\sigma = -BB^T$, and (25) can be rewritten as $A^2 = (A^T)^2$. This last condition is satisfied, for example, if A is symmetric. In this case (26) becomes

$$v = e^{-x^T x/4\sigma} \tag{28}$$

which is a constant multiple of the square root of the steady-state probability density. This is in accordance with our earlier results for the one-dimensional case.

Denote by ∇ the gradient with respect to the metric on $\mathbb{R}^n$ given by $G = (BB^T)^{-1}$. In other words, given a function $\phi : \mathbb{R}^n \to \mathbb{R}$, we define the vector $\nabla\phi$ by $(\nabla\phi)_i = \sum_{j=1}^n (BB^T)_{ij}\phi_{x_j}$. Assume that $A = A^T$ and that the positive definite solution of (22) is $Q = \sigma I$. Let $\phi(x) = \frac{1}{4\sigma}x^T x$. Then we have

$$(\nabla\phi)_i = 2\sigma \sum_{j=1}^n A_{ij}\phi_{x_j} = \sum_{j=1}^n A_{ij}x_j$$

so that the system (20) can be rewritten as

$$dx = -\nabla\phi(x)\,dt + B\,dw. \tag{29}$$

Systems of the general form (29) and the corresponding steady-state probability densities were studied in [8].

Under the present assumptions, the Fokker–Planck operator takes the form

$$L_{n,\sigma}\rho = -\sigma \sum_{i,j=1}^n A_{ij}\rho_{x_i x_j} - \sum_{i,j=1}^n A_{ij}x_j\rho_{x_i} - \operatorname{tr}A \cdot \rho.$$

Using Proposition 7, we can also construct a self-adjoint operator which in this case is given by

$$T_{n,\sigma}\rho = -\sigma \sum_{i,j=1}^n A_{ij}\rho_{x_i x_j} - \left(\frac{1}{2}\operatorname{tr}A - \frac{1}{4\sigma}\sum_{i,j=1}^n A_{ij}x_i x_j\right)\rho.$$

As we have done throughout the paper, we consider the above expression as defining a closed operator acting on a dense subspace of $L^2(\mathbb{R}^n)$ which we denote by $\mathcal{D}_{T_{n,\sigma}}$. The following result is to be viewed as a generalization of Proposition 3 to the case of the multidimensional system (20) written in equipartition coordinates as explained earlier ($Q = \sigma I$), under the assumption that in these coordinates the nonrandom part of the system is symmetric ($A = A^T$). As shown above, this system is of the gradient form (29) for an appropriate quadratic function ϕ and a suitable constant metric.

THEOREM 8. *For any $\sigma > 0$, the spectrum of the operator $T_{n,\sigma} : \mathcal{D}_{T_{n,\sigma}} \to L^2(\mathbb{R}^n)$ consists of eigenvalues which are the sums of the eigenvalues of the matrix A.*

Proof. The matrix A has n real negative eigenvalues $\lambda_1, \dots, \lambda_n$. There exists an orthogonal matrix R such that $RAR^T = D$, where $D = \operatorname{diag}(\lambda_1, \dots, \lambda_n)$. Making the change of variables $y = Rx$, we obtain an operator $\overline{T}_{n,\sigma}$ given by

$$\overline{T}_{n,\sigma}\rho = -\sigma\sum_{i=1}^n \lambda_i\rho_{y_i y_i} - \left(\frac{1}{2}\sum_{i=1}^n \lambda_i - \frac{1}{4\sigma}\sum_{i=1}^n \lambda_i y_i^2\right)\rho.$$

The spectrum of $\overline{T}_{n,\sigma}$ is the same as that of $T_{n,\sigma}$. We have $\overline{T}_{n,\sigma} = -\sum_{i=1}^n \lambda_i T_{\sigma,y_i}$, where T_{σ,y_i} are the operators considered in section 3 for each variable (cf. remarks made at the beginning of this section). To complete the proof, recall that by Proposition 3 the eigenvalues of T_{σ,y_i} are the nonpositive integers. The eigenfunctions of $\overline{T}_{n,\sigma}$ are given by the products of the functions $u_k(y_i, 2\sigma)$ for each variable; they form an orthonormal basis for $L^2(\mathbb{R}^n)$. ∎

As before, we conclude that the eigenvalues of the original operator $L_{n,\sigma}$ restricted to the space of functions of the form $\rho = vf$, where v is given by (28) and $f \in \mathcal{D}_{T_{n,\sigma}}$, are the sums of the eigenvalues of A. The corresponding eigenfunctions take the form $vg_k(x,\sigma)$, where $g_k(x,\sigma)$, $k = 0, 1, \dots$ are the eigenfunctions of $T_{n,\sigma}$ described in the proof of Theorem 8. Let $\mathcal{L}_{n,\sigma}$ denote the space of functions $\{\rho : v^{-1}\rho = \sum_{k=1}^{m} \alpha_k g_k(x,\sigma)\}$, $\alpha_k \in \mathbb{R}$. As a generalization of Corollary 4 we have the following statement.

COROLLARY 9. *The space $\mathcal{L}_{n,\sigma}$ is invariant with respect to the Fokker–Planck operator $L_{n,\sigma}$ associated with* (29)*. The spectrum of the restriction of $L_{n,\sigma}$ to $\mathcal{L}_{n,\sigma}$ consists of eigenvalues which are the sums of the eigenvalues of the matrix A.*

Remark 2. Under the change of variables described in the proof of Theorem 8 the operator $L_{n,\sigma}$ becomes a Fokker–Planck operator associated with a decoupled system. This makes the statement of Corollary 9 obvious in view of Corollary 4 and the discussion at the beginning of this section. In the case when A is a scalar multiple of the identity matrix, the spectrum (but not the eigenfunctions) of the operator $L_{n,\sigma}$ on $\mathcal{L}_{n,\sigma}$ can be obtained from the analysis of its adjoint presented in [19, section 7.5]. The eigenfunctions of $L_{n,\sigma}$ on $\mathcal{L}_{n,\sigma}$ found above form an orthonormal basis for the space $L^2(\mathbb{R}^n, e^{x^T x/4\sigma}dx)$. We could also consider $L_{n,\sigma}$ as acting on a larger dense subspace of $L^2(\mathbb{R}^n, e^{x^T x/4\sigma}dx)$, which would not change the spectrum—cf. Remark 1 in section 3.

It is interesting to notice that, given the original system (20), we can always find a basis in which $Q = \sigma I$ satisfies (22) and $A = A^T$ if A is allowed to depend on time. First, switch to an equipartition basis in which we have $Q = \sigma I$. Note that the last equality is preserved under orthogonal coordinate transformations. Let $\Omega = \frac{1}{2}(A - A^T)$. Making the change of variable $y = e^{-\Omega t}x$ in (20), we obtain

$$dy = \frac{1}{2}e^{-\Omega t}(A + A^T)e^{\Omega t}y\,dt + e^{-\Omega t}B\,dw \tag{30}$$

and the first term features a symmetric matrix as needed.

5. Double bracket equations. Consider the operators P_1, P_2, P_3, and P_4 acting on the space

$$\mathcal{D} = \{\rho \in C^2(\mathbb{R}) \cap L^2(\mathbb{R}) : \rho_{xx}, x\rho_x, x^2\rho \in L^2(\mathbb{R})\}$$

that are defined as follows:

$$P_1\rho = \rho_{xx}, \qquad P_2\rho = x\rho_x, \qquad P_3\rho = x^2\rho, \qquad P_4\rho = \rho.$$

It is easy to verify that the linear span of the above operators is closed under commutation with respect to the usual Lie bracket $[P_i, P_j] = P_iP_j - P_jP_i$. We will let $\mathfrak{g}$ denote the operator Lie algebra spanned by P_i, $i = 1, 2, 3, 4$. Such Lie algebras and their representations have been studied in the context of quantum mechanics and, more recently, estimation theory [5].

Observe that L_σ and T_σ can be realized as operators in $\mathfrak{g}$ because $\mathcal{D} \subset \mathcal{D}_{L_\sigma}$ for each $\sigma > 0$. More precisely, let us denote by $L(\sigma)$ and $T(\sigma)$ the restrictions of L_σ and T_σ to $\mathcal{D}$. Proposition 3 implies that $T(\sigma)$, $0 < \sigma < \infty$, is an isospectral family of operators in $\mathfrak{g}$. In fact, for any $\sigma > 0$ the spectrum of $T(\sigma)$ consists of eigenvalues which are the nonpositive integers. As we know from Proposition 2, the nonpositive integers are also eigenvalues of $L(\sigma)$ for each $\sigma > 0$ (because the corresponding eigenfunctions of L_σ belong to $\mathcal{D}$).

In this section we show that the families of operators $L(\sigma)$ and $T(\sigma)$ correspond to integral curves of differential equations in the double bracket form on $\mathfrak{g}$. We also obtain the corresponding dynamical representation for the family of modified Hermite functions defined by (3). The proofs are completely straightforward calculations and will not be given.

PROPOSITION 10. *Let M be an operator in $\mathfrak{g}$ defined by $M\rho = \frac{1}{4}\rho_{xx}$. Then $L(\sigma)$, $0 < \sigma < \infty$, is a solution of the differential equation*

$$\frac{dL}{d\sigma} = [L, [L, M]]. \tag{31}$$

The Fokker–Planck operator associated with (5) is uniquely determined by two parameters: σ, which corresponds to the steady-state variance, and a, which describes the speed of convergence to steady state. In making the transition to the operators L_σ we factored out the dependence on a. Thus the flow (31) can be thought of as evolving on the "slice" of Fokker–Planck operators with the same spectral properties but different steady states. As we will see, in the multidimensional case σ corresponds to the steady-state temperature of the system (defined in section 4).

To each of the operators $L(\sigma)$ there corresponds the steady-state probability density ρ_σ which satisfies the equation $L(\sigma)\rho_\sigma(x) = 0$. The flow (31) on the operator Lie algebra $\mathfrak{g}$ thus induces a flow on the manifold of Gaussian probability densities. For example, making the change of variable $\sigma = e^t$, we obtain a particular case of the gradient flow of Gaussians described by Nakamura in [17].

PROPOSITION 11. *Let N be an operator in $\mathfrak{g}$ defined by $N\rho = \frac{1}{2}\rho_{xx}$. Then $T(\sigma)$, $0 < \sigma < \infty$, is a solution of the differential equation*

$$\frac{dT}{d\sigma} = [T, [T, N]]. \tag{32}$$

In view of the remarks made at the end of section 3, we can write $T(\sigma) = \Theta^{-1}(\sigma)T(1/2)\Theta(\sigma)$, with the domain of $\Theta(\sigma)$ properly defined. Using the fact that for all $\sigma > 0$ the operator $\Theta(\sigma)$ is unitary and the operators $T(\sigma)$ and $N(\sigma)$ are self-adjoint, we arrive at the equation

$$\frac{d\Theta}{d\sigma} = T(1/2)\Theta N - \Theta N\Theta^{-1}T(1/2)\Theta = \Theta[T, N] \tag{33}$$

which describes the evolution of the eigenbasis for $T(\sigma)$ induced by the flow (32). This is the same equation as the one obtained in [6] for the finite-dimensional case.

We point out an interesting analogy between the results of Propositions 10 and 11 and the sorting algorithms described in [6]. If N is a real diagonal matrix with unrepeated eigenvalues, and if $H(0)$ is a suitably chosen symmetric matrix, then the solution of the double bracket equation $\dot{H} = [H, [H, N]]$ approaches a diagonal matrix $H(\infty)$ such that the diagonal elements of $H(\infty)$ and N are similarly ordered; since $H(\infty)$ is diagonal, it commutes with N. For large positive values of σ, the "principal term" of the operators $L(\sigma)$ and $T(\sigma)$ is $\sigma\frac{d^2}{dx^2}$, which is proportional to both M and N and thus commutes with them. Thus the double bracket equations (31) and (32) can be thought of as performing a task of "operator sorting."

We would like to generalize the above results to the multidimensional case. Consider the operators $P_{1,i,j}$, $P_{2,i,j}$, $P_{3,i,j}$, and P_4 acting on the space

$$\mathcal{D}_n = \{\rho \in C^2(\mathbb{R}^n) \cap L^2(\mathbb{R}^n) : \rho_{x_ix_j}, x_i\rho_{x_j}, x_ix_j\rho \in L^2(\mathbb{R}^n)\ \forall i, j = 1, 2, \ldots, n\}$$

that are defined as follows:

$$P_{1,i,j}\rho = \rho_{x_i x_j}, \qquad P_{2,i,j}\rho = x_i \rho_{x_j}, \qquad P_{3,i,j}\rho = x_i x_j \rho, \qquad P_4 \rho = \rho.$$

These operators span a Lie algebra which we denote by $\mathfrak{g}_n$. For each $\sigma > 0$, let $L_n(\sigma)$ and $T_n(\sigma)$ denote the restrictions of $L_{n,\sigma}$ and $T_{n,\sigma}$ to $\mathcal{D}_n$. Theorem 8 implies that $T_n(\sigma)$, $0 < \sigma < \infty$, is an isospectral family of operators in $\mathfrak{g}_n$.

PROPOSITION 12. *Let M be an operator in $\mathfrak{g}_n$ defined by*

$$M\rho = -\frac{1}{4} \sum_{i,j=1}^{n} (A^{-1})_{ij} \rho_{x_i x_j}.$$

Then $L_n(\sigma)$, $0 < \sigma < \infty$, is a solution of the differential equation

$$\frac{dL}{d\sigma} = [L, [L, M]].$$

PROPOSITION 13. *Let N be an operator in $\mathfrak{g}_n$ defined by*

$$N\rho = -\frac{1}{2} \sum_{i,j=1}^{n} (A^{-1})_{ij} \rho_{x_i x_j}.$$

Then $T_n(\sigma)$, $0 < \sigma < \infty$, is a solution of the differential equation

$$\frac{dT}{d\sigma} = [T, [T, N]].$$

As in the scalar case, we can define a unitary operator Θ_σ by $\Theta_\sigma g_k(x, \sigma) = g_k(x, 1)$, where $g_k(x, \sigma)$ are the eigenfunctions of $T_n(\sigma)$. This operator will then satisfy (33), which describes the flow of these eigenfunctions. Finally, note that Propositions 12 and 13 apply to the system (30) without any changes (except that now $T_n(\sigma)$ will also depend on t).

Acknowledgments. The first author would like to thank Mark Adler for helpful discussions and Daniel Stroock for calling his attention to the book [19].

REFERENCES

[1] G. BIRKHOFF AND G. C. ROTA, *Ordinary Differential Equations. Introductions to Higher Mathematics*, Ginn, Boston, 1962.

[2] A. M. BLOCH, R. W. BROCKETT, AND T. S. RATIU, *Completely integrable gradient flows*, Comm. Math. Phys., 147 (1992), pp. 57–74.

[3] R. W. BROCKETT, *Lie algebras and Lie groups in control theory*, in Geometric Methods in System Theory, D. Q. Mayne and R. W. Brockett, eds., Reidel Publishing Co., Dordrecht, The Netherlands, 1973, pp. 43–82.

[4] R. W. BROCKETT, *Parametrically stochastic linear differential equations*, Math. Programming Stud., 5 (1976), pp. 8–21.

[5] R. W. BROCKETT, *Nonlinear systems and nonlinear estimation theory*, in Stochastic Systems: The Mathematics of Filtering and Identification and Applications, M. Hazewinkel and J. C. Willems, eds., NATO Adv. Study Inst. Ser. C: Math. Phys. Sci. 78, Reidel, Dordrecht, Boston, 1981, pp. 441–477.

[6] R. W. BROCKETT, *Dynamical systems that sort lists, diagonalize matrices, and solve linear programming problems*, Linear Algebra Appl., 146 (1991), pp. 79–91.

[7] R. W. BROCKETT, *Oscillatory descent for function minimization*, in Current and Future Directions in Applied Mathematics, M. Alber et al., eds., Birkhäuser, Boston, 1997, pp. 65–82.

[8] R. W. Brockett, *Notes on stochastic processes on manifolds*, in Systems and Control in the Twenty-First Century, C. I. Byrnes et al, eds., Progr. Systems Control Theory 22 Birkhäuser, Boston, 1997, pp. 75–100.

[9] R. W. Brockett and D. Liberzon, *On explicit steady-state solutions of Fokker–Planck equations for a class of nonlinear feedback systems*, in Proceedings of the American Control Conference, Philadelphia, PA, 1998, pp. 264–268.

[10] R. Y. Chang and M. L. Wang, *The properties of modified Hermite polynomials and their applications to functional differential equations*, J. Chinese Inst. Engrs., 9 (1986), pp. 75–81.

[11] I. I. Gikhman and A. V. Skorokhod, *Stochastic Differential Equations*, Springer-Verlag, New York, 1972.

[12] S. Goldberg, *Unbounded Linear Operators*, McGraw-Hill, New York, 1966.

[13] H. W. Gould and A. T. Hopper, *Operational formulas connected with two generalizations of Hermite polynomials*, Duke Math. J., 29 (1962), pp. 51–63.

[14] R. A. Holley, S. Kusuoka, and D. W. Stroock, *Asymptotics of the spectral gap with applications to the theory of simulated annealing*, J. Funct. Anal., 83 (1989), pp. 333–347.

[15] L. V. Kantorovich and G. P. Akilov, *Functional Analysis*, Pergamon Press, Oxford, UK, 1982.

[16] D. Liberzon, *Asymptotic Properties of Nonlinear Feedback Control Systems*, Ph.D. Thesis, Brandeis University, Waltham, MA, 1998.

[17] Y. Nakamura, *Completely integrable gradient flows on the manifolds of Gaussian and multinomial distributions*, Japan J. Indust. Appl. Math., 10 (1993), pp. 179–189.

[18] E. D. Rainville, *Infinite Series*, Macmillan, New York, 1967 .

[19] D. W. Stroock, *Probability Theory, an Analytic View*, Cambridge University Press, Cambridge, UK, 1993 .

SIAM J. CONTROL OPTIM.
Vol. 38, No. 5, pp. 1468–1480

STABILIZATION OF LINEAR DIFFERENTIAL SYSTEMS VIA HYBRID FEEDBACK CONTROLS*

ELENA LITSYN†, YURII V. NEPOMNYASHCHIKH‡, AND ARCADY PONOSOV§

Abstract. We study so-called "hybrid feedback stabilizers" for an arbitrarily general system of linear differential equations. We prove that under assumptions of controllability and observability there exists a hybrid feedback output control which makes the system asymptotically stable. The control is designed by making use of a discrete automaton implanted into the system's dynamics. In general, the automaton has infinitely many locations, but it gives rise to a "uniform" (in some sense) feedback control. The approach we propose goes back to classical feedback control techniques combined with some ideas used in stability theory for equations with time-delay.

Key words. stabilization, hybrid feedback control, functional differential equations

AMS subject classification. 93D15

PII. S036301299833255X

1. Introduction.

Hybrid systems are those that combine both discrete and continuous dynamics. Many examples of hybrid systems can be found in manufacturing systems, intelligent vehicle highway systems, and various chemical plants. Hybrid systems also arise when there is a necessity to combine logical decision with the generation of continuous control laws. See also the "biological" Example 1.2 below.

We are interested in the question of how to stabilize a continuous control plant through its interaction with a discrete time controller (an automaton). Such a feedback complements the output feedback within the plant when the latter fails to stabilize the system or fails to give smooth stabilization. This, for example, may be the case if no complete information on the plant's dynamics is available.

The framework we use follows developments in hybrid systems, as described in Nerode and Kohn [12]. The stability and stabilization notions we use are classical (see, for example, [11], [14], [16], [17], where some relevant results can be found). Some results on stabilization of hybrid systems are available in the literature [1], [5], [9], [13]. In this paper we exploit ideas proposed by Artstein [2], who studied a possibility of stabilization by hybrid feedback controls via examples. The point of view given in [2] is that one has an underlying continuous plant, and the challenge is to stabilize it efficiently with a hybrid device.

Let us now consider some examples that give a partial motivation for our study. More examples can be found in [15].

Example 1.1 (the harmonic oscillator). A rather simple example of a linear system

*Received by the editors February 18, 1998; accepted for publication (in revised form) October 5, 1999; published electronically May 11, 2000.

http://www.siam.org/journals/sicon/38-5/33255.html

†Department of Theoretical Mathematics, The Weizmann Institute of Science, Rehovot 76100, Israel, and The Research Institute, College of Judea and Samaria, Ariel 44837, Israel (elenal@wisdom.weizmann.ac.il). The research of this author was partially supported by the Ministry of Science and the Ministry of Absorption, Center for Absorption in Science, Israel.

‡Department of Mechanics and Mathematics, Perm State University, Bukirev Street 15, 614600 Perm, Russia (yuvn@psu.ru). The research of this author was supported by a foundation grant for young scientists (Goscomvuz RF, 1997) and by grants 96-15-96195 and 99-01-01278 of the RFBR and of the research grant in natural sciences, St.-Petersburg.

§Institutt for Matematiske Fag, NLH, Postboks 5035, N-1432 Ås, Norway (matap@imf.nlh.no).

which cannot be stabilized by an ordinary output feedback is two-dimensional:

$$\frac{d\xi}{dt} = \eta, \qquad \frac{d\eta}{dt} = -\xi + u, \qquad y = \xi. \tag{1.1}$$

This is the harmonic oscillator where the control is an external force and the only measured quantity (output) is the position variable ξ.

Although this system is both controllable and observable, it cannot be stabilized by (even discontinuous and nonlinear) output feedback (see, e.g., [2]) since any control $u = f(\xi)$ makes curves of the form $\xi^2 + \eta^2 + 2\int_0^\xi f(\rho)d\rho = c$ invariant under the flow.

However, it was shown in [2] that there exists a *hybrid feedback control* under which system (1.1) becomes asymptotically stable.

Example 1.2 (predator–prey interactions). Our second example gives a "biological" motivation for studying hybrid feedback controls. Consider the well-known predator–prey model

$$\frac{x'}{x} = -b_1 + a_1 y, \qquad \frac{y'}{y} = b_2 - a_2 x, \tag{1.2}$$

where a_i, b_i are positive; $x(t)$ describes a population at time t of a species called *the predator* (fish, animals, insects, etc.); and $y(t)$ describes a population at time t of another species called *the prey*.

The *predator* lives on the *prey*; $\frac{x'}{x}$ and $\frac{y'}{y}$ mean the growth rates of the populations.

The linearization about the equilibrium state (e_1, e_2) gives the following linear system:

$$\dot{X} = e_1 a_1 Y, \qquad \dot{Y} = -e_2 a_2 X, \tag{1.3}$$

that is, the harmonic oscillator which is asymptotically *unstable*. Any attempt to stabilize (1.3) (i.e., (1.2) locally) by inserting a certain control into the right-hand side of (1.3) fails. Such a control, usually called a "harvesting strategy," would give

$$\frac{x'}{x} = -b_1 - u + a_1 y, \qquad \frac{y'}{y} = b_2 - v - a_2 x. \tag{1.4}$$

The linearization of (1.4) about the new equilibrium (e_1^v, e_2^u) state will again give an asymptotically unstable system

$$\dot{X} = e_1^v a_1 Y \equiv e_1 a_1 Y - \frac{v a_1}{a_2} Y, \qquad \dot{Y} = -e_2^u a_2 X \equiv -e_2 a_2 X - \frac{u a_2}{a_1} X.$$

The problem becomes more complicated if it is not allowed to harvest the prey. Then $v = 0$, and we obtain nothing but the controlled oscillator (1.1) up to a change of variables.

We also remark that (1.2) describes one of the simplest models of biological interactions. If we take more species and more complicated interactions, then we shall arrive at more general control problems.

The goal of the paper. We look at a general linear system of differential equations which is asymptotically unstable. We also assume that no complete information about solutions is available and that we cannot continuously influence the solutions. Nevertheless, we ask whether it is still possible to stabilize the system.

In Artstein [2] two conjectures have been mentioned: (1) whether there exists a kind of elementary hybrid stabilizer $u(\cdot)$ driven by an automaton with finitely many locations and (2) whether there is a possibility for expediency by using more general hybrid stabilizers.

In the present paper, we give a partial affirmative answer to Artstein's questions. Namely, we describe here how one can explicitly design a hybrid feedback stabilizer for a general linear system. To do so, we introduce a new class of hybrid feedback control strategies and automata with infinitely many locations, this being a natural generalization of the elementary hybrids used in [2]. Our approach is based on the classical stabilization technique as well as on some recent results in the theory of *functional differential equations* [3], [4], [7].

2. Formulation of the main result. The following system of linear differential equations

$$\begin{aligned} \dot{x} &= Ax + Bu, \\ y &= Cx \end{aligned} \tag{2.1}$$

is under consideration. Here $x \in \mathbf{R}^n$ is a physical state of the system, $y \in \mathbf{R}^m$ is the output, and the control u is an element of $\mathbf{R}^\ell$. A, B, C are given real matrices of the sizes $n \times n$, $n \times \ell$, $m \times n$, respectively. *The control $u(\cdot)$ is assumed to depend on the output $y(\cdot)$ only.* The nature of this dependence is specified below.

The following problem is being studied: find (if there is a possibility) a (hybrid feedback) control u under which system (2.1) becomes asymptotically stable.

To specify our setup, we need some definitions and auxiliary results (see also [6], [8], [14], [16]).

DEFINITION 2.1. *Denoting by Z the set of all admissible controls, we say that the system* (2.1) *is Z-stabilizable if there exists a control $u \in Z$ under which* (2.1) *is u-stabilizable.*

The following technical theorem will be used.

THEOREM 2.2. *The pair (A, B) is controllable if and only if for any set Λ of n complex numbers which is symmetric with respect to the real axis, there exists a real $\ell \times n$-matrix G such that $\sigma(A + BG) = \Lambda$.*

In that which follows, for the sake of simplicity we will write $[A]_B$ instead of the $n \times \ell n$-matrix $[B\ AB\ \dots\ A^{n-1}B]$ and ${}_C[A]$ instead of the $mn \times n$-matrix $([A^\top]_{C^\top})^\top$.

The next statement is well known.

THEOREM 2.3 (Kalman's criterion). 1. *The pair (A, B) is controllable if and only if* $\operatorname{rank}([A]_B) = n$. 2. *The pair (A, C) is observable if and only if* $\operatorname{rank}({}_C[A]) = n$.

We will always assume in this paper that the pair (A, B) is controllable, and the pair (A, C) is observable.

Denote by $\mathcal{L}_0$ the class of controls $u = Gy$, where G is a constant $\ell \times m$-matrix. The following statement is of wide use in control theory.

LEMMA 2.4. *If* $\operatorname{rank} C = n$ *then the system* (2.1) *is $\mathcal{L}_0$-stabilizable.*

Proof. According to Theorem 2.2 there exists a real $\ell \times n$-matrix G_0 such that the matrix $A + BG_0$ is stable. Since $\operatorname{rank} C = n \le m$, there exists the left inverse matrix C_l^{-1}. The matrix $G = G_0 C_l^{-1}$ then provides a control which stabilizes the system (2.1). □

If $\operatorname{rank} C < n$, then this may imply an absence of a stabilizer in the class $\mathcal{L}_0$. If in addition one assumes that it may be desirable in applications to have discrete stabilizers, one immediately arrives at the concept of a *hybrid system* consisting of a discrete automaton coupled with a system to be stabilized (a plant).

Below we describe the concept of an automaton, basically following the approach presented in [2]. Our setup is a bit more general, however, since we also admit automata with infinitely many locations, so that our definitions differ from those in [2]. We have in mind a typical automaton, described by a triple $\mathcal{A} = (Q, I, M)$, where

(i) Q is a set of all possible automaton states (locations);

(ii) the set I contains the input alphabet;

(iii) the transition map $M : Q \times I \to Q$ indicates the location after a transition time, based on the previous location q and input i at the time of transition.

We also let our automaton follow solutions of the system (2.1). This can be done by assuming there is one more triplet $\mathcal{B} = (T, i, q_0)$. Here

(iv) $T : Q \to (0, \infty)$ is a mapping which sets a period $T(q)$ between transition times;

(v) $i : \mathbf{R}^m \to I$ is a function providing the element $i(y)$ of the alphabet I for any output y of the system (2.1);

(vi) and $q_0 = q(\tau_0)$ is a state of the automaton at the initial time τ_0.

We will later assume, without loss of generality, that $\tau_0 = 0$.

DEFINITION 2.5. *By an automaton we mean a* 6*-tuple* $\Delta = (Q, I, M, T, i, q_0)$.

For arbitrary sets X, Y (topological spaces X, Y) we denote by $\mathbf{P}(X, Y)$ (resp., $\mathbf{C}(X, Y)$) the set of all functions (resp., continuous functions) from X to Y.

Now, for any automaton Δ satisfying (i)–(vi) we define by induction a Volterra operator $F_\Delta : \mathbf{C}([0, \infty), \mathbf{R}^m) \to \mathbf{P}([0, \infty), Q)$. For each $y \in \mathbf{C}([0, \infty), \mathbf{R}^n)$, F_Δ is given by the following:

1. $(F_\Delta y)(0) = q(0); \quad \tau_1 = T(q(0)); \quad (F_\Delta y)(t) \equiv q(0), \quad t \in [0, \tau_1)$.
2. $(F_\Delta y)(\tau_1) = M(q(0), i(y(\tau_1))) := q(\tau_1); \quad \tau_2 = \tau_1 + T(q(\tau_1))$;
   $(F_\Delta y)(t) = q(\tau_1), \quad t \in [\tau_1, \tau_2)$.

3. If $\tau_0, \tau_1, \ldots, \tau_k$ and the values $(F_\Delta y)(t)$ for $t \in [0, \tau_k)$ are already known, then τ_{k+1} and $(F_\Delta y)(t)$ for $t \in [\tau_k, \tau_{k+1})$ are defined by the equalities

$$(F_\Delta y)(\tau_k) = M(q(\tau_{k-1}), i(y(\tau_k))) := q(\tau_k); \quad \tau_{k+1} = \tau_k + T(q(\tau_k));$$
$$(F_\Delta y)(t) \equiv q(\tau_k), \quad t \in [\tau_k, \tau_{k+1}).$$

DEFINITION 2.6. *A control* $u(\cdot)$ *of the type*

$$u(t) = \varphi(y(t), (F_\Delta y)(t)), \tag{2.2}$$

where $\varphi : \mathbf{R}^m \times Q \to \mathbf{R}^\ell$ *is a certain function, will be addressed as a* hybrid feedback control (HFC).

Denote by $\tilde{\mathcal{H}}$ a class of all HFC in the sense of Definition 2.6.

Let ω be finite or infinite cardinal. A class of HFC (2.2) generated by an automaton $\Delta = (Q, I, M, T, i, q_0)$, where $\operatorname{card} Q \le w$, $\operatorname{card} I \le w$ ($\operatorname{card} \mathcal{D}$ stands for the cardinality of a set $\mathcal{D}$), will be denoted by $\mathcal{H}_\omega$.

DEFINITION 2.7. *A hybrid feedback control* $u(\cdot)$ *belonging to the class* $\mathcal{H}_\omega$ *will be called* ω*-HFC.*

Some examples. 1. The class $\mathcal{H}_0$ consists of ordinary (nonlinear) feedback controls which are of the type $u = f(y)$ for some function $f : \mathbf{R}^m \to \mathbf{R}^\ell$. Clearly, $\mathcal{L}_0 \subset \mathcal{H}_0$ ($\mathcal{L}_0$ was introduced before Lemma 2.4, above).

2. It is also evident that $\mathcal{H}_1 = \mathcal{H}_0$, i.e., in case Q degenerates into a singleton, any HFC is given by a feedback control of the type $u = f(y)$ for a function $f : \mathbf{R}^m \to \mathbf{R}^\ell$.

3. An elementary hybrid system is that with a finite number of locations $Q = \{q_1, \ldots, q_n\}$ (see [2]). An elementary hybrid system gives rise to an elementary HFC.

In our notation, an elementary HFC is nothing but an n-HFC (or HFC of the class $\mathcal{H}_n$) for some natural number n. A typical elementary (or, more generally, discrete) hybrid system's dynamic is continuous, and the solution satisfies (2.1) with $u = \varphi(y, q_i)$ on the time intervals $(\tau_i, \tau_{i+1}]$ which cover $\mathbf{R}_+$. More specific examples of 2- and 3-HFCs are presented in [2].

The class $\cup_{n=1}^{\infty}\mathcal{H}_n$ of all elementary hybrids will hereafter be denoted by $\mathcal{H}_e$.

Let σ and ω_1 be the cardinalities of a countable set and a continuum set, respectively. Then

$$\mathcal{H}_0 = \mathcal{H}_1 \subset \mathcal{H}_2 \subset \cdots \subset \bigcup_{n=1}^{\infty} \mathcal{H}_n = \mathcal{H}_e \subset \mathcal{H}_\sigma \subset \mathcal{H}_{\omega_1} \subset \cdots \subset \tilde{\mathcal{H}}.$$

For systems in finite-dimensional spaces (when $y \in \mathbf{R}^m$) we always have $\mathcal{H}_{\omega_1} = \tilde{\mathcal{H}}$. The classes $\mathcal{H}_\omega$ for $\omega > \omega_1$ can be useful for more general dynamical systems, for example, those in abstract metric spaces or other topological spaces (some examples can be found in [12]).

Now we are able to define the class of HFCs which we will use to find a stabilizer to a system (2.1).

DEFINITION 2.8. *An HFC $u(\cdot)$ given by* (2.2) *is called* uniform *if*

(1) *T is a constant function;*

(2) *the function φ does not depend on the first argument, i.e., $u(t) = \varphi((F_\Delta y)(t))$.*

The set of all uniform HFCs will hereafter be denoted by $\mathcal{H}$. Clearly, $\mathcal{H} \subset \mathcal{H}_\sigma$.

The main result of our paper follows.

THEOREM 2.9. *The system* (2.1) *is $\mathcal{H}$-stabilizable under assumptions of controllability of (A, B) and observability of (A, C).*

3. The proof of the main result. We find it convenient to use the following notation:

– $\mathcal{M}_{j,k}$ is the set of all real $j \times k$-matrices;
– $\mathcal{M}_k = \mathcal{M}_{k,k}$ is the set of all real $k \times k$-matrices;
– $\mathcal{L}(\mathbf{R}^j, \mathbf{R}^k)$ is the set of all linear operators from $\mathbf{R}^j$ to $\mathbf{R}^k$;
– $\mathcal{L}(\mathbf{R}^k) = \mathcal{L}(\mathbf{R}^k, \mathbf{R}^k)$ is the set of all linear operators in the space $\mathbf{R}^k$;
– $|\cdot|$ is the Euclidean norm in the spaces $\mathbf{R}^k$ (for any k);
– $\|\cdot\|$ is the corresponding operator-norm in the spaces $\mathcal{L}(\mathbf{R}^j, \mathbf{R}^k)$;
– I_k is the identity $k \times k$-matrix;
– χ_E is the characteristic function of a set E;
– $\mathbf{C}$ is the space of all continuous functions from $[0, \infty)$ to $\mathbf{R}^n$;
– $\mathbf{L}_0$ is the space of all Lebesgue measurable functions from $[0, \infty)$ to $\mathbf{R}^n$.

We need two lemmas to prove Theorem 2.9.

LEMMA 3.1. *There exists an $(\ell \times n)$-matrix G such that the matrix $A + BG$ is stable and the matrix pair $(A + BG, C)$ is observable.*

Proof. According to Theorem 2.2 we choose a matrix $G_0 \in \mathcal{M}_{\ell,n}$, for which $A + BG_0$ is stable. Clearly, there exists $\varepsilon > 0$ such that for any $t \in (1 - \varepsilon, 1 + \varepsilon)$ a matrix $A + tBG_0$ is stable. Consider a function $\psi : \mathbf{R} \to \mathbf{R}$, which to each t associates the sum of squares of all nth order minors of the matrix ${}_C[A + tBG_0]$.

Due to Kalman's criterion $\mathcal{P} := \{\, t \in \mathbf{R} \mid (A + tBG, C) \text{ is observable} \,\} = \psi^{-1}(\mathbf{R} \setminus \{0\})$. Evidently, ψ is a polynomial, and moreover, $\psi \not\equiv 0$. The last inequality follows immediately from the observability of the pair (A, C), which implies $\psi(0) \neq 0$. Thus ψ has a finite number of zeros and, therefore there exists $t^* \in \mathcal{P} \bigcap (1 - \varepsilon, 1 + \varepsilon)$. Hence, the matrix $G = t^* G_0$ satisfies the required conditions. ∎

Our next lemma is a version of the Lyapunov stability theorem for a class of *functional differential equations*. Some similar, but not exactly similar, results are available in the literature (see, e.g., [3], [7], and references therein).

The Cauchy formula we will use is classical (see, e.g., [3] or [4]):

$$x(t) = \mathcal{X}(t)x(0) + \int_0^t \mathcal{C}(t,s)f(s)ds, \qquad t \geq 0. \tag{3.1}$$

Here $\mathcal{C}(t,s)$ is the Cauchy matrix of the equation

$$Lx \equiv \dot{x}(t) - \sum_{k=1}^{n} \tilde{a}_k(t)x[h_k(t)] = f(t), \qquad t > 0, \tag{3.2}$$

and $\mathcal{X}(t)$ is the fundamental matrix of the corresponding homogeneous equation.

LEMMA 3.2. *Let $\tilde{a}_k : [0,\infty) \to \mathcal{L}(\mathbf{R}^n)$ be locally Lebesgue-integrable and $h_k : [0,\infty) \to \mathbf{R}$ Lebesgue measurable functions. Assume that there exist $N, \beta, \tau > 0$ such that $\zeta(t) \leq h_k(t) \leq t$, $t \in [0,\infty)$, $k = 1, \dots, n$, where $\zeta(t) = \max\{0, t-\tau\}$ and*

$$\|\mathcal{C}(t,s)\| \leq Ne^{-\beta(t-s)}. \tag{3.3}$$

Then there exists $\varepsilon > 0$ such that for any operator $V : \mathbf{C} \to \mathbf{L}_0$ satisfying

$$|(Vx)(t)| \leq \varepsilon \max_{\zeta(t)\leq s\leq t} |x(s)|, \qquad t \in [0,\infty), \quad x \in \mathbf{C}, \tag{3.4}$$

the zero solution of the perturbed equation

$$Lx = Vx \tag{3.5}$$

is globally exponentially stable:

$$|x(t)| \leq N_0 e^{-\beta_0 t}|x(0)|, \qquad t \geq 0, \tag{3.6}$$

for any solution $x(t)$ of (3.6) and for some positive constants N_0, β_0 independent of $x(t)$.

Proof. From the Cauchy formula (3.1) as well as from (3.3) and (3.4) it immediately follows that

$$|x(t)| \leq Ne^{-\beta t}|x(0)| + N\varepsilon \int_0^t e^{-\beta(t-s)} \max_{\zeta(s)\leq\xi\leq s} |x(\xi)|ds, \qquad t \geq 0.$$

Putting $z(t) := \max_{\zeta(t)\leq s\leq t}|x(t)|e^{\beta t}$ we obtain the estimate

$$z(t) \leq N_0|x(0)| + N_0\varepsilon \int_0^t z(s)ds, \qquad t \geq 0,$$

with $N_0 = Ne^{\beta\tau}$.

By the Gronwall–Bellman inequality, $z(t) \leq N_0|x(0)|e^{N_0\varepsilon}$, $t \geq 0$, so that $|x(t)| \leq N_0 e^{(N_0\varepsilon-\beta)t}|x(0)|$, $t \geq 0$. Then for any $\varepsilon < \beta/N_0$ we obtain the estimate (3.6). ∎

Now we are ready to start proving our central result.

I. By Lemma 3.1, we find a matrix $G \in \mathcal{M}_{m,n}$ such that the matrix $A + BG$ is stable and the pair $(A + BG, C)$ is observable.

Let us fix an arbitrary $\tau > 0$ and put

$$A(t) = \chi_{[0,\tau]}A_0 + \chi_{[\tau,\infty)}A_1, \qquad \tau \in [0,\infty),$$

where $A_0 = A, \quad A_1 = A + BG$.

Denote by $X(t,s)$ the Cauchy matrix of the ordinary differential equation $\dot{x} = A(t)x$. We do not assume in that which follows that $t \geq s$ (defining $X(s,t) = X^{-1}(t,s)$ if necessary). Evidently,

$$\begin{array}{ll} X(t,s) = e^{A_1(t-s)}, & t \geq \tau, \quad s \geq \tau, \\ X(t,s) = X(t,\tau-0)X(\tau,s) = e^{A_0(t-\tau)}e^{A_1(\tau-s)}, & 0 \leq t \leq \tau \leq s. \end{array} \tag{3.7}$$

We do not describe all possible cases here, restricting ourselves to those formulae which we are going to use in the course of the proof.

II. Consider the following equation for the unknown matrix-valued function $X(t) := [X_1(t), \ldots, X_n(t)] \in \mathcal{M}_{n,mn}$ with $X_k : [\tau, \infty) \to \mathcal{M}_{n,m}$:

$$\sum_{k=1}^{n} X_k(t) C X(p\tau + kh, t) = I_n, \quad t \in [(p+1)\tau, (p+2)\tau) := E_p. \tag{3.8}$$

Here $h = \tau/n$ and $p = 0, 1, 2, \ldots$. Our next aim is to prove existence of a solution $X(t)$ to this equation.

(a) Let us first assume that $t \in E_0$. Then $0 \leq kh \leq \tau \leq t, \quad k = 1, \ldots, n$, and due to (3.7)

$$X(kh, t) = e^{A_0(kh-\tau)}e^{A_1(\tau-t)} = T_0^{k-1}U_0, \quad k = 1, \ldots, n, \tag{3.9}$$

where $T_0 = e^{A_0 h}, \quad U_0 = e^{A_0(h-\tau)}e^{A_1(\tau-t)}$.

(b) Let t belong to $E_p, \quad p = 1, 2, \ldots$. Then $0 < \tau \leq p\tau + kh \leq t, \quad k = 1, \ldots, n$, and according to (3.7)

$$X(p\tau + kh, t) = T_1^{k-1}U_p, \quad k = 1, \ldots, n, \tag{3.10}$$

where $T_1 = e^{A_1 h}, \quad U_p = e^{A_1(p\tau+h-t)}$. Notice that from the observability of the pairs (A_i, C) and Kalman's criterion it follows that $\operatorname{rank}({}_C[A_i]) = n$, $i = 0, 1$. Hence, for generic h (see, e.g., [10]),

$$\operatorname{rank}({}_C[T_i]) = n \tag{3.11}$$

as well. In what follows we shall always assume, without loss of generality, that our h satisfies conditions (3.11).

For an arbitrary ordered n-tuple $\mathcal{K} = \{k_1, \ldots k_n\} \subset \{1, 2, \ldots, mn\}$ ($k_j < k_{j+1}$, $j = 1, \ldots, n-1$) we now define two operators $R_{\mathcal{K}} : \mathcal{M}_{mn,n} \to \mathcal{M}_n$, $S_{\mathcal{K}} : \mathcal{M}_n \to \mathcal{M}_{mn,n}$ in the following manner:

– jth row in the matrix $R_{\mathcal{K}}Z$ coincides with k_jth row of the matrix Z;

– k_jth column of the matrix $S_{\mathcal{K}}Z$ coincides with jth column of the matrix Z;

– the columns with numbers $k \notin \mathcal{K}$ of the matrix $S_{\mathcal{K}}Z$ are all null-columns.

From (3.11) it follows that for some ordered n-tuples $\mathcal{K}_i \subset \{1, 2, \ldots, mn\}$ the matrices $R_{\mathcal{K}_i}({}_C[T_i])$ $(i = 1, 2)$ are nonsingular. Due to (3.9), (3.10), it is straightforward that the matrix $X(t)$, defined by

$$X(t) = S_{\mathcal{K}_i}\left(U_p^{-1}\left(R_{\mathcal{K}_i}({}_C[T_i])^{-1}\right)\right), \tag{3.12}$$

satisfies (3.9), where it is assumed that $i = p = 0$ for $t \in E_0$ and $i = 1$ for $t \in E_p$, $p = 1, 2, \ldots$.

III. Put

$$u_1(t) = \begin{cases} 0, & 0 \le t < \tau, \\ G\left(\sum_{k=1}^{n} X_k(t) y(p\tau + kh)\right), & t \in E_p, \quad p = 0, 1, 2, \dots, \end{cases} \tag{3.13}$$

where $y = Cx$ and $X(t) := [X_1(t), \dots, X_n(t)]$ is defined by (3.12). Due to (3.8),

$$u_1(t) = G\left(\sum_{k=1}^{n} X_k(t) C x(p\tau + kh)\right)$$
$$= G\left(\sum_{k=1}^{n} X_k(t) C X(p\tau + kh, t) x(t)\right) = Gx(t),$$
$$t \in E_p, \quad p = 0, 1, 2, \dots.$$

Hence, (2.1) coincides with the exponentially stable linear equation $\dot{x} = A_1 x$ in the interval $t \in [\tau, \infty)$. Therefore, the system (2.1) is u_1-stabilizable.

But u_1 is not yet a hybrid feedback control in the sense of Definition 2.8.

According to the formula (2.2) we now need to find suitable discrete approximations for $A_i(t)$ and $Cx(\tau h + kh)$. To do so, we will use a technique based on the preservation of the asymptotic stability for functional differential equations with respect to small perturbations both in coefficients and time-delays [4].

IV. Let us first rewrite the system (2.1) involving the control u_1 in the form of the following delay-equation:

$$\dot{x}(t) - \sum_{k=0}^{n} a_k(t) x[h_k(t)] = 0, \quad t \in [0, \infty). \tag{3.14}$$

Here

$$a_0(t) = \chi_{[0,\tau)}(t) A, \qquad a_k(t) = \chi_{[\tau,\infty)}(t) BGX_k(t) C, \quad k = 1, \dots, n,$$
$$h_0(t) = t, \qquad h_k(t) = \sum_{p=0}^{\infty} \chi_{E_p}(t)(p\tau + kh), \quad k = 1, \dots, n.$$

As shown in III, (3.14) is exponentially stable. From (3.12), it also follows that

$$a_k(t) = BG\, S_{\mathcal{K}_i}\left(U_p^{-1}\left(R_{\mathcal{K}_i}\,({}_C[T_i])^{-1}\right)\right) \Lambda_k C, \tag{3.15}$$
$$k = 1, \dots, n, \quad t \in E_p, \quad p = 0, 1, 2, \dots,$$

where the block matrices $\Lambda_k \in \mathcal{M}_{mn,n}$ are defined by

$$\Lambda_k = (\underbrace{\theta \dots \theta}_{m(k-1)}\ I_m \underbrace{\theta \dots \theta}_{m(n-k)})^{\top}, \qquad k = 1, \dots, n$$

(θ is an m-dimensional zero vector-column), and $i = 0$ for $p = 0$, $i = 1$ for $p > 0$.

The representation (3.15) implies that each of the functions $a_k : [0, \infty) \to \mathcal{L}(\mathbf{R}^n)$ is piecewise continuous with possible jumps at the points $p\tau$, $p = 1, 2, \dots$, only.

Since

$$\|a_k(t)\| \le \max\left\{\|A\|, \|BG\| \cdot \|C\| \cdot \max_{i=1,2}\left\|(S_{\mathcal{K}_i}({}_C[T_i]))^{-1}\right\|\right.$$
$$\left.\times \max(1, \|e^{-A_0\tau}\|) \cdot \|e^{A_1\tau}\|\right\} < \infty, \qquad t \in [\tau, \infty),$$

the functions a_k are bounded. Then, according to Corollary 2 in [4, p. 173], there exists $\sigma > 0$ such that for all locally Lebesgue-integrable functions $\tilde{a}_k : [0,\infty) \to \mathcal{L}(\mathbf{R}^n)$, satisfying

$$\max_k \limsup_{t\to\infty} \|a_k(t) - \tilde{a}_k(t)\| < \sigma, \tag{3.16}$$

the Cauchy matrix $\mathcal{C}(t,s)$ of (3.2) has the exponential estimate (3.3).

Equalities (3.15) and (3.10) imply a periodicity of a_k with the period τ for $t \geq 2\tau$. Now let us approximate $X(t)$ by a step function on $[2\tau, 3\tau]$. For the sake of convenience, we may assume that the points $2\tau + kh$ are included in the set of possible jump-points of the step function. Then we extend this function τ-periodically to the interval $[3\tau,\infty)$. Finally, we approximate $X(t)$ on $[\tau, 2\tau]$ by a suitable step function. Let us notice that such approximations can be found with a prescribed accuracy. Our output will be a function $\tilde{X}(t) = [\tilde{X}_1(t), \dots, \tilde{X}_n(t)]$ of the form

$$\tilde{X}_k(t) = \sum_{j=0}^{J-1} \chi_{E_{1j}}(t)\tilde{c}_{kj} + \sum_{p=1}^{\infty}\sum_{j=0}^{J-1} \chi_{E_{pj}}(t)c_{kj}, \quad t \in E_{pj}, \tag{3.17}$$

where J is a natural constant, $\delta = \tau/J$, $E_{pj} = [(p+1)\tau + j\delta, (p+1)\tau + (j+1)\delta)$, and $\tilde{c}_{kj}, c_{kj} \in \mathcal{M}_{n,m}$ are some matrices as well. Here $k = 1, \dots, n$, $j = 0, 1, \dots, J-1$, and $p = 0, 1, \dots$. By construction, $\|\tilde{X}_k(t) - X_k(t)\| < \sigma \cdot (\|BG\| \cdot \|C\|)^{-1}$.

We now set $a_0(t) \equiv A$, $\tilde{a}_k(t) = BG\tilde{X}_k(t)C$, $k = 1, \dots, n$. Then

$$\max_k \sup_{t\in[0,\infty)} \|\tilde{a}_k(t) - a_k(t)\| < \sigma,$$

so that (3.16) holds.

Consequently, we obtain that the Cauchy matrix of (3.2) admits the exponential estimate (3.3).

V. According to Lemma 3.2 one can choose a positive ε such that for any (as general as necessary) nonlinear operator $V : \mathbf{C} \to \mathbf{L}_0$ satisfying the condition (3.4) every solution x of the perturbed equation (3.5) has the exponential estimate (3.6) with certain positive constants N_0, β_0 (independent of x).

By (3.17), $\varepsilon_1 := \varepsilon[\,\|BG\| \cdot \|C\| \cdot \sup_{t\in[0,\infty)} \sum_{k=1}^n \|\tilde{X}_k(t)\|\,]^{-1} > 0$. A multivalued function $\Phi : \mathbf{R}^m \to 2^{\mathbf{Q}^m}$ ($\mathbf{Q}$ is the set of rational numbers) defined by

$$\Phi(v) = \{r \in \mathbf{Q}^m \,|\, |v - r| \leq \varepsilon_1 |v|\}$$

then has nonempty images: $\Phi(v) \neq \emptyset$, $v \in \mathbf{R}^m$. We take an arbitrary selector ϱ of the multivalued function Φ and define V as follows:

$$(Vx)(t) = \sum_{k=1}^n BG\tilde{X}_k(t)\left(\varrho(Cx[h_k(t)]) - Cx[h_k(t)]\right). \tag{3.18}$$

It is easy to check that V acts from $\mathbf{C}$ to $\mathbf{L}_0$ and satisfies the inequality (3.4). Hence (3.5) with the operator V just defined becomes asymptotically stable, and the stability is uniform with respect to compact subsets.

By construction, (3.5) has the form

$$\dot{x}(t) = Ax(t) + \sum_{k=1}^n BG\tilde{X}_k(t)\varrho(Cx[h_k(t)]).$$

This equation is, in turn, equivalent to the original system (2.1), where the corresponding control u is defined by

$$(3.19)\quad u(t)=\begin{cases} 0, & t\in[0,\tau),\\ G\left(\sum\limits_{k=1}^{n}\tilde{c}_{kj}\varrho(y(kh))\right), & t\in E_{0j},\quad j=0,1,\dots,J-1,\\ G\left(\sum\limits_{k=1}^{n}c_{kj}\varrho(y(p\tau+kh))\right), & t\in E_{pj},\quad j=0,1,\dots,J-1,\\ & \qquad\quad p=1,2,\dots. \end{cases}$$

Recall that $y=Cx$ is the output.

The system (2.1) controlled by u from (3.19) is therefore asymptotically stable, and the stability is uniform with respect to any compact subset of $\mathbf{R}^n$.

VI. We have not yet shown that $u\in\mathcal{H}$.

Put $N_J=\{0,1,\dots,J-1\}$ and denote by Ω a subset of $\mathcal{M}_{\ell,mn}$ consisting of the block matrices

$$\tilde{c}_j=(G\tilde{c}_{1j},\dots,G\tilde{c}_{nj}),\quad c_j=(Gc_{1j},\dots,Gc_{nj}),\quad \bar{z}=(z,\dots,z),\quad j=1,\dots,J-1,$$

where $z\in\mathcal{M}_{\ell,mn}\setminus\{G\tilde{c}_{kj},Gc_{kj}\mid k=1,\dots,n,\quad j=1,\dots,J-1\}$ is arbitrary. Put $\mathcal{Y}\subset\Omega\times N_J$. The elements of $\mathcal{Y}$ are $(\bar{z},j)$, $(\tilde{c}_j,j)$, and (c_j,j). Let us define a mapping $M_0:\mathcal{Y}\to\mathcal{Y}$ by

$$\begin{aligned} &M_0(\bar{z},j)=(\bar{z},j+1), && j=1,\dots,J-2, && M_0(z,J-1)=(\tilde{c}_0,0),\\ &M_0(\tilde{c}_j,j)=(\tilde{c}_{j+1},j+1), && j=0,1,\dots,J-2, && M_0(\tilde{c}_{J-1},J-1)=(c_0,0),\\ &M_0(c_j,j)=(c_{j+1},j+1), && j=0,1,\dots,J-2, && M_0(c_{J-1},J-1)=(c_0,0). \end{aligned}$$

Let P be the set of all row-vectors of the form $q=(q_1,\dots,q_n)$, where $q_k\in\mathbf{Q}^m$. We extend now M_0 to the set $\Omega\times N_J$ and define two mappings $M_-:P\times N_J\times P\to P$ and $M_+:P\times P\times N_J\to P$ as follows:

$$M_-(q,j,i)=\begin{cases} q & \text{if } j\neq J-1,\ j\neq\frac{kJ}{n},\ k=1,\dots,n-1,\\ (q_1,\dots,q_{k-1},i,q_{k+1},\dots,q_n) & \text{if } j=\frac{kJ}{n},\ k=1,\dots,n-1,\\ (q_1,\dots,q_{n-1},i) & \text{if } j=J-1, \end{cases}$$

where $q=(q_1,\dots,q_n)$,

$$M_+(q^+,q^-,j)=\begin{cases} q^+ & \text{if}\quad j\neq 0,\\ q^- & \text{if}\quad j=0. \end{cases}$$

We first describe an automaton $\Delta=(Q,I,M,T,i,q_0)$. Let the (countable) set $Q=P\times P\times(\Omega\times N_J)$ contain 4-tuples $(q^+,q^-,(c,j))$ and the (countable) alphabet I be equal to $\mathbf{Q}^m$. The mapping $M:Q\times I\to Q$ is then defined by

$$M(q^+,q^-,(c,j),i)=(M_+(q^+,M_-(q^-,j,i),j),M_-(q^-,j,i),M_0(c,j)).$$

Assume that $T\equiv\delta$ and define $i:\mathbf{R}^m\to I$ and q_0 by

$$i(s)=\varrho(s),\qquad q_0=(\Theta,\Theta,(\bar{z},0)),$$

where Θ is zero in P.

Finally, we choose a function $\varphi : Q \to \mathbf{R}^\ell$ as follows:

$$\varphi((q_1, \dots, q_n), q^-, ((d_1, \dots, d_n), j)) = \begin{cases} 0 & \text{if} \quad (d_1, \dots, d_n) = \bar{z}, \\ \sum\limits_{k=1}^{n} d_k q_k & \text{if} \quad (d_1, \dots, d_n) \neq \bar{z}. \end{cases}$$

By construction, the control u given by (3.19) is of the form $u(\cdot) = \varphi((F_\Delta y)(\cdot))$ and belongs to the class $\mathcal{H}$. The control u is therefore a uniform HFC in the sense of Definition 2.8. Moreover according to part V of the proof, the control u stabilizes the system (2.1). The proof of Theorem 2.9 is completed.

4. An example. Here we show how the proposed algorithm can be used to design an explicit hybrid strategy to stabilize particular systems.

Example 4.1. Consider the linear control system

$$\dot{\xi} = \eta + u, \qquad \dot{\eta} = \xi, \qquad u(t) = \varphi(\xi(t), (F_\Delta \xi)(t)). \tag{4.1}$$

This system has the following two properties:

1. There is no linear HFC stabilizing the system (4.1).

2. The pair (A, B) is controllable, the pair (A, C) is observable, so that according to Theorem 2.9 the system (4.1) is $\mathcal{H}$-stabilizable. The corresponding hybrid control strategy can be written explicitly.

To prove 1, one may observe that for an arbitrary $\eta_0 > 0$ the set $(\xi(t), \eta(t)) \in [0, \infty) \times [\eta_0, \infty) := \mathcal{D}$ is flow-invariant for any linear control strategy. This excludes asymptotic stability.

Now, in order to design an *explicit* HFC $u \in \mathcal{H}$ stabilizing the system (4.1) one should follow the algorithm from Theorem 2.9. For instance, if we take $G = [\,-2, \, -3\,]$, then the eigenvalues of the matrix $A + BG$ will be $-1 \pm i$, so that $A + BG$ is stable, and G can be used to construct an HFC. We put also $h = \frac{\pi}{2}$. Evidently,

$$e^{A_0 t} = \begin{bmatrix} \cosh t & \sinh t \\ \sinh t & \cosh t \end{bmatrix}, \qquad e^{A_1 t} = e^{-t} \cdot \begin{bmatrix} \cos t - \sin t & -2 \sin t \\ \sin t & \cos t + \sin t \end{bmatrix}. \tag{4.2}$$

The fundamental matrix defined in (3.7) satisfies

$$\|X(t, s)\| \leq \sqrt{7}\, e^{2\pi - t + s}, \qquad t \geq s \geq 0.$$

Now, from this and from simple estimates on $\|BG\|$, $\|C\|$ we can deduce that any piecewise approximation $\tilde{X}(t)$ of the matrix function $X(t)$ satisfying

$$\|\tilde{X}_k(t) - X_k(t)\| < \frac{e^{-4\pi}}{8\sqrt{7}}, \qquad k = 1, 2, \quad t \in [\pi, \infty), \tag{4.3}$$

will stabilize the system (4.1).

To obtain this estimate one can, e.g., define $\tilde{X}_k(t)$, $k = 1, 2$, by putting

$$\tilde{X}_k(t) = \sum_{p=0}^{\infty} \sum_{j=0}^{J-1} \chi_{E_{pj}}(t) X_k(2\pi(p+1) + \pi(2j+1)/(2J)), \quad t \in E_{pj}, \tag{4.4}$$

where $J = 5 \times 10^7$, $E_{pj} = [2\pi(p+1) + \pi j/J, 2\pi(p+1) + \pi(j+1)/J)$, $j = 0, 1, \dots, J-1$, and $p = 0, 1, \dots$.

Finally, for some $\varepsilon_1 \in (0, \frac{e^{-4\pi}}{40\sqrt{7}})$ we define a function $\varrho : \mathbf{R} \to \mathbf{Q}$ by $\varrho(v) = \big[10^K v\big] \cdot 10^{-K}$, where $K = \max\{1,\ 1 + [\lg \frac{2}{\varepsilon|v|}]\}$, $[w]$ being the integral part of $w \in \mathbf{R}$.

The resulting hybrid strategy will be as follows.

We set $u(t) = 0$ when $t \in [0, 2\pi)$ and define

$$\begin{aligned} u(t) = &-2X_{11}(2\pi(p+1) + \pi(2j+1)/(2J)) \cdot \varrho(\xi(\pi p + \pi/2)) \\ &-3X_{21}(2\pi(p+1) + \pi(2j+1)/(2J)) \cdot \varrho(\xi(\pi p + \pi/2)) \\ &-2X_{12}(2\pi(p+1) + \pi(2j+1)/(2J)) \cdot \varrho(\xi(\pi p + \pi)) \\ &-3X_{22}(2\pi(p+1) + \pi(2j+1)/(2J)) \cdot \varrho(\xi(\pi p + \pi)) \end{aligned}$$

when $t \in E_{pj},\ \ j = 0, 1, \ldots, J-1,\ \ p = 0, 1, \ldots$.

Remark. All the estimates in this example are rather rough. Using mathematical software will give more suitable constants if required.

5. Conclusion. We do not have answers to the following questions.

(1) Is an arbitrary system (2.1) $\mathcal{H}_e$-stabilizable under assumptions of (A, B)-controllability and (A, C)-observability? In other words, it is not clear to what extent one may use HFCs with finitely many locations in order to stabilize (2.1).

(2) Examples show that the assumptions of (A, B)-controllability and (A, C)-observability are not necessary for the hybrid feedback stabilization (just take a trivial example of an exponentially stable linear system with $B = C = 0$). The second open problem is to find better conditions for the hybrid feedback stabilization in terms of matrices A, B, C.

Acknowledgments. We would like to thank Professor Zvi Artstein for introducing us to the problem and for his very helpful comments. We are also grateful to the anonymous referee who carefully read the manuscript and suggested a number of improvements.

REFERENCES

[1] R. J. Antsaklis, J. A. Stiver, and M. Lemmon, *Hybrid system modeling and autonomous control systems*, in Hybrid Systems, Lecture Notes in Comput. Sci. 736, R. L. Grossman, A. Nerode, A. P. Ravn, and H. Rischel, eds., Springer-Verlag, Berlin, 1993, pp. 366–392.

[2] Z. Artstein, *Example of stabilization with hybrid feedback*, in Hybrid Systems III: Verification and Control, Lecture Notes in Comput. Sci. 1066, Springer-Verlag, Berlin, 1996, pp. 173–185.

[3] N. V. Azbelev, V. P. Maksimov, and L. F. Rakhmatullina, *Introduction to the Theory of Functional Differential Equations*, World Federation Publishers Inc., Atlanta, 1996.

[4] L. Berezansky and E. Braverman, *Preservation of the exponential stability under perturbation of linear delay impulsive differential equations*, Z. Anal. Anwendungen, 14 (1995), pp. 157–174.

[5] M. S. Branicky, *Studies in Hybrid Systems: Modeling, Analysis and Control*, Ph.D. Thesis, Massachusetts Institute of Technology, Cambridge, MA, 1995.

[6] R. W. Brokett, *Finite Dimensional Linear Systems*, John Wiley, New York, 1970.

[7] E. N. Chukwu, *Stability and Time-Optimal Control of Hereditary Systems*, Math. Sci. Engrg. 188, Academic Press, San Diego, 1992.

[8] P. A. Fuhrmann, *A Polynomial Approach to Linear Algebra*, Springer-Verlag, New York, 1996.

[9] J. Guckenheimer, *A robust hybrid stabilization strategy for equilibria*, IEEE Trans. Automat. Control, 40 (1995), pp. 321–326.

[10] G. Kreisselmeier, *On sampling without loss of observability/controllability*, IEEE Trans. Automat. Control, 44 (1999), pp. 1021–1025.

[11] H. Kwakernaak and R. Sivan, *Linear Optimal Control Systems*, Wiley-Interscience, New York, London, Sydney, Toronto, 1972.

[12] A. Nerode and W. Kohn, *Models for hybrid systems: Automata, topologies, controllability, observability*, in Hybrid Systems, Lecture Notes in Comput. Sci. 736, R. L. Grossman, A. Nerode, A. P. Ravn, and H. Rischel, eds., Springer-Verlag, Berlin, 1993, pp. 317–356.
[13] E. Sontag, *Nonlinear regulation: The piecewise linear approach*, IEEE Trans. Automat. Control, 26 (1981), pp. 346–358.
[14] E. Sontag, *Mathematical Control Theory*, Springer-Verlag, New York, 1989.
[15] S. H. Strogatz, *Nonlinear Dynamics and Chaos: With Applications to Physics, Biology, Chemistry, and Engineering*, Addison-Wesley, Reading, MA, 1994.
[16] W. A. Wolovich, *Linear Multivariable Systems*, Springer-Verlag, New York, Heidelberg, Berlin, 1974.
[17] W. M. Wonham, *Linear Multivariable Control: A Geometric Approach*, Springer-Verlag, New York, Heidelberg, Berlin, 1979.

SIAM J. CONTROL OPTIM.
Vol. 38, No. 5, pp. 1481–1500

MINIMAX CONTROLS FOR UNCERTAIN PARABOLIC SYSTEMS*

NADIR ARADA†, MAÏTINE BERGOUNIOUX‡, AND JEAN-PIERRE RAYMOND§

Abstract. We consider systems governed by a nonlinear parabolic equation with a distributed control and a disturbance in the initial condition. We prove the existence of solutions to a corresponding minimax problem, and we obtain necessary optimality conditions in the form of Pontryagin's principles.

Key words. uncertain systems, minimax, semilinear parabolic equations, existence of optimal solutions, necessary optimality conditions, Pontryagin's principle

AMS subject classifications. 93C73, 93C20, 93C10, 49K20, 49J20

PII. S0363012998345603

1. Introduction.

In this paper we consider an uncertain system described by the parabolic equation

$$\frac{\partial y}{\partial t} + Ay + \Phi(x,t,y) = u \text{ in } Q, \qquad y = \psi \text{ on } \Sigma, \tag{1.1}$$
$$y(x,0) = y_o(x) + g(x) \text{ in } \Omega,$$

where $Q = \Omega\times]0,T[$, Ω is a bounded domain in $\mathbb{R}^N$, $\Sigma = \Gamma\times]0,T[$, Γ is the boundary of Ω, and A is a second-order elliptic operator, $\psi \in L^\infty(\Sigma)$. The function u is a control variable; the initial condition is not completely known. We only suppose that g belongs to G_{ad}, where G_{ad} is a closed convex subset of $L^\infty(\Omega)$ (not necessarily reduced to a unique element). For this reason, system (1.1) is called an "uncertain system." Many physical systems may be described by equations involving disturbances, noises, or uncertainties. Here we suppose that the disturbance only appears through the initial condition, but the results of this paper may be extended to other classical situations. We can, for example, consider systems with a disturbance or a control in the boundary condition (see [2, 3]). Let us denote by $y(u,g)$ the solution of (1.1) corresponding to (u,g), and for a given $g \in G_{ad}$, consider the problem

$$(\mathcal{P}_g) \qquad \inf \{ \mathcal{I}(y(u,g),u,g) \mid u \in U_{ad} \},$$

where U_{ad} is a given control set, and $\mathcal{I}$ is a cost functional that we make explicit hereafter. We denote by $\text{Argmin}(\mathcal{P}_g)$ the set of solutions to $(\mathcal{P}_g)$, and we set $J(u,g) = \mathcal{I}(y(u,g),u,g)$. The control problem that we consider is the following.

$$(\mathcal{P}) \qquad \begin{cases} \text{Find } \bar{u} \in U_{ad} \text{ such that } \bar{u} \in \text{Argmin } (\mathcal{P}_{\bar{g}}) \text{ for some } \bar{g} \in G_{ad}, \text{ and} \\ J(u_g,g) \le J(\bar{u},\bar{g}) \text{ for all } g \in G_{ad} \text{ and all } u_g \in \text{ Argmin } (\mathcal{P}_g). \end{cases}$$

*Received by the editors October 27, 1998; accepted for publication (in revised form) November 1, 1999; published electronically May 19, 2000.
http://www.siam.org/journals/sicon/38-5/34560.html
†UMR-CNRS 5640, UFR MIG, Université Paul Sabatier, 31062 Toulouse Cedex 4, France (arada@mip.ups-tlse.fr).
‡UMR-CNRS 6628, Université d'Orléans, U.F.R. Sciences, B.P. 6759, F-45067 Orléans Cedex 2, France (Maitine.Bergounioux@labomath.univ-orleans.fr).
§UMR-CNRS 5640, UFR MIG, Université Paul Sabatier, 31062 Toulouse Cedex 4, France (raymond@mip.ups-tlse.fr).

This problem may be expressed in an equivalent form as

$$\begin{cases} \text{Find } \bar{g} \in G_{ad} \text{ and } u_{\bar{g}} \in \text{ Argmin } (\mathcal{P}_{\bar{g}}) \text{ such that} \\ J(u_g, g) \leq J(u_{\bar{g}}, \bar{g}) \text{ for all } g \in G_{ad} \text{ and all } u_g \in \text{ Argmin } (\mathcal{P}_g), \end{cases}$$

or

$$\max_{g \in G_{ad}} \min_{u \in U_{ad}} J(u, g) \ .$$

The cost functional that we consider is defined as follows:

$$\mathcal{I}(y, u, g) = \int_Q (F(x, t, y) + H(x, t, u)) \, dx \, dt + \int_\Omega (\ell(x, y(T)) + L(x, g)) \, dx,$$

where H is convex with respect to u, and L is concave with respect to g. We prove the existence of optimal solutions to $(\mathcal{P})$ (Theorem 4.3), and we establish optimality conditions in the form of Pontryagin's principles (Theorems 2.1 and 2.2). The proof of Pontryagin's principle for an optimal solution $\bar{g}$ is derived via Taylor's expansions of the state variable and the cost functional for diffuse perturbations of $\bar{g}$ (see Theorem 5.2). This is the main part of the paper. When the state equation is linear and when F and ℓ are convex, the proof is much simpler (see Theorem 2.3 and its proof in section 6.3).

We may relate this kind of problem to the concept of robustness, since considering a min-max problem is equivalent to finding the best control which takes into account the "worst" disturbance in the initial value. For linear equations, such problems have been studied by Lions [11]. The notion of "least regret controls" in [11] corresponds to our definition of robust controls.

Problem (P) is also connected to H_∞-control problems. Indeed for a quadratic functional of the form

$$\mathcal{I}(y, u, g) = \int_Q |y - y_d|^2 \, dx \, dt + \int_Q u^2 \, dx \, dt - \gamma \int_\Omega g^2 \, dx \qquad (\gamma > 0),$$

for linear equations ($\Phi \equiv 0$), and for the control sets $U_{ad} = L^2(Q)$, $G_{ad} = L^2(\Omega)$, robust controls that we consider are suboptimal solutions to some H_∞-control problems (see [5, p. 218]). However the terminology of "robust control" is not adapted here since we always deal with open-loop systems.

A series of papers has been widely devoted to uncertain systems [14], [1]. To analyze the different original contributions and to compare our results with the previous ones, we must distinguish the nature of the system under consideration (i.e., the state equation) and the definition of optimal solutions. First examine the second point. Ahmed and Xiang [1] and Mordukhovitch and Zhang [12] establish optimality conditions for saddle points, that is optimal strategies $(\bar{u}, \bar{g})$ satisfying

$$\sup_{g \in G_{ad}} \inf_{u \in U_{ad}} J(u, g) = \inf_{g \in G_{ad}} \sup_{u \in U_{ad}} J(u, g) = J(\bar{u}, \bar{g}).$$

The inequality

$$\sup_{g \in G_{ad}} \inf_{u \in U_{ad}} J(u, g) \leq \inf_{g \in G_{ad}} \sup_{u \in U_{ad}} J(u, g)$$

is always satisfied, and it is well known that for linear equations and convex cost functionals, the equality holds [6]. For nonlinear equations, the results presented here

are new to our knowledge. Thus our optimal solutions are different from the ones of Ahmed and Xiang. Indeed, for a saddle point strategy, that is for $(\bar{u}, \bar{g}) \in U_{ad} \times G_{ad}$ satisfying

$$J(\bar{u}, g) \leq J(\bar{u}, \bar{g}) \leq J(u, \bar{g}) \qquad \text{for all } u \in U_{ad}, \text{ all } g \in G_{ad},$$

optimality conditions for $\bar{u}$ and $\bar{g}$ can be obtained by separately considering two optimization problems [1]. Such an approach is not applicable for the problem that we consider. Mordukhovitch and Zhang [12] study minimax problems in the presence of pointwise state constraints. Optimality conditions are established for saddle point solutions satisfying the state constraints. Taking advantage of the linear structure of the state equation, the saddle point problem is split into two optimization problems. Still in this case, optimality conditions for $\bar{u}$ and $\bar{g}$ can be obtained separately. To our knowledge, the optimality conditions established in Theorem 2.2 for a minimax problem of the form

$$\sup_{g \in G_{ad}} \inf_{u \in U_{ad}} J(u, g)$$

are a completely new result. Papageorgiou [14] and Ahmed and Xiang [1] consider uncertain systems where the noise is modeled by parametrized measures. These authors consider a control problem where the uncertain term appears as a *distributed* measure. Our purpose is to consider an initial perturbation because in many situations, for example, in data assimilation, the initial condition is not well known. In this case, the disturbance can be considered as an impulsive disturbance, and the analysis is more complicated (see the Taylor's expansions stated in Theorem 5.2 and the role played by the parameter τ_ρ).

The systems considered in [1] are more general than the one considered here, but the results in [1] cannot be applied to problems with uncertain initial conditions. Since the case of convection-diffusion equations considered in section 5 in [1] is interesting from the point of view of applications, we explain in section 7 how to extend our results to these kinds of equations.

Finally, we mention that the existence of solutions for minimax control problems governed by variational inequalities is proven in [14].

The paper is organized as follows. After setting the problem, we present the main assumptions and results in section 2. Section 3 is devoted to the study of solutions for the state and adjoint equations. We prove existence results for the control problems in section 4. In section 5, we establish some Taylor expansions that are used to perform the proofs of the main results in section 6. Extensions to convection-diffusion equations are stated in section 7.

2. Assumptions and main results.

2.1. Assumptions. Throughout the paper, Ω is a bounded open and connected subset in $\mathbb{R}^N (N \geq 2)$ of class $C^{2+\bar{\gamma}}$ for some $0 < \bar{\gamma} \leq 1$, and q is a positive constant satisfying $q > \frac{N}{2} + 1$. The operator A is of the form $Ay(x) = -\sum_{i,j=1}^{N} D_i(a_{ij}(x) D_j y(x))$, where D_i denotes the partial derivative with respect to x_i. The coefficients a_{ij} of A belong to $\mathcal{C}^{1+\bar{\gamma}}(\overline{\Omega})$ and satisfy the condition

$$a_{ij}(x) = a_{ji}(x) \ \text{ for all } i, j \in \{1, \ldots, N\}, \quad m_o |\xi|^2 \leq \sum_{i,j=1}^{N} a_{ij}(x) \xi_i \xi_j$$

for every $\xi \in \mathbb{R}^N$ and every $x \in \overline{\Omega}$, with $m_o > 0$.

Let us define the notation.

- The conormal derivative of y with respect to A is denoted by $\dfrac{\partial y}{\partial n_A}$; that is

$$\frac{\partial y}{\partial n_A}(s,t) = \sum_{i,j} a_{ij}(s) D_j y(s,t) n_i(s),$$

  where $n = (n_1, \ldots, n_N)$ is the unit normal to Γ outward Ω.
- $\overline{\Omega}_o = \overline{\Omega} \times \{0\}$, $\overline{\Omega}_T = \overline{\Omega} \times \{T\}$. For any $\tau > 0$, we set $Q_{\tau T} = \Omega \times]\tau, T[$, $\Omega_\tau = \{ x \in \Omega \mid d(x, \Gamma) > \tau \}$, $Q^\tau = \Omega_\tau \times]\tau, T[$ (d is the Euclidean distance).
- For every $1 \le d \le \infty$, the usual norms in the spaces $L^d(\Omega)$, $L^d(Q)$, $L^d(\Sigma)$ will be denoted by $\|\cdot\|_{d,\Omega}$, $\|\cdot\|_{d,Q}$, $\|\cdot\|_{d,\Sigma}$.
- If $\mathcal{O}$ is a locally compact subset of $\mathbb{R}^{N+1}$, $\mathcal{C}_b(\mathcal{O})$ denotes the space of bounded continuous functions on $\mathcal{O}$, and $\mathcal{C}_o(\mathcal{O})$ the space of all continuous functions from $\mathcal{O}$ into $\mathbb{R}$ vanishing at infinity. The dual space of $\mathcal{C}_o(\mathcal{O})$ is denoted by $\mathcal{M}_b(\mathcal{O})$ (it is the space of bounded Radon measures on $\mathcal{O}$).

Now let us set the assumptions.

(A1) The control set is defined as

$$U_{ad} = \{ u \in L^q(Q) \mid u(x,t) \in K_U(x,t) \text{ for almost all } (x,t) \in Q \},$$

where $K_U(\cdot)$ is a measurable multivalued mapping with nonempty, convex, and closed values in $\mathcal{P}(\mathbb{R})$. The set of constraints on g is defined by

$$G_{ad} = \{ g \in L^\infty(\Omega) \mid g(x) \in K_G(x) \subset G \text{ for almost all } x \in \Omega \},$$

where $K_G(\cdot)$ is a measurable multivalued mapping with nonempty, convex, and closed values in $\mathcal{P}(\mathbb{R})$, and G is a compact subset of $\mathbb{R}$. We suppose that U_{ad} and G_{ad} are nonempty.

REMARK 2.1. *Observe that U_{ad} is a closed, convex subset of $L^q(Q)$. Similarly, G_{ad} is convex and bounded in $L^\infty(\Omega)$ and closed in $L^s(\Omega)$ for all $1 \le s \le +\infty$.*

(A2) Φ is a Carathéodory function from $Q \times \mathbb{R}$ into $\mathbb{R}$. For almost every $(x,t) \in Q$, $\Phi(x,t,\cdot)$ is of class $\mathcal{C}^1$. Moreover, the following estimates hold.

$$|\Phi(x,t,0)| \le \Phi_1(x,t), \ 0 \le \Phi'_y(x,t,y) \le \Phi_1(x,t)\, \eta(|y|),$$

where $\Phi_1 \in L^q(Q)$, and η is a nondecreasing function from $\mathbb{R}^+$ to $\mathbb{R}^+$.

(A3) F and H are Carathéodory functions from $Q \times \mathbb{R}$ to $\mathbb{R}$. For almost all $(x,t) \in Q$, $F(x,t,\cdot)$ is of class $\mathcal{C}^1$ and $H(x,t,\cdot)$ is convex. Moreover, the following estimates hold.

$$-C_1|y|^\sigma \le F(x,t,y) \le F_1(x,t)\, \eta(|y|), \quad |F'_y(x,t,y)| \le F_2(x,t)\, \eta(|y|),$$

$$C_1|u|^q \le H(x,t,u) \le H_1(x,t) + C_2\, |u|^q,$$

where $C_1 > 0$, $C_2 > 0$, $1 \le \sigma < q$, $F_1 \in L^1(Q)$, $H_1 \in L^1(Q)$, $F_2 \in L^m(Q)$ with $m > 1$, and η is defined as in (A2) .

(A4) ℓ and L are Carathéodory functions from $\Omega \times \mathbb{R}$ into $\mathbb{R}$. For almost all $x \in \Omega$, $\ell(x,\cdot)$ is $\mathcal{C}^1$, $L(x,\cdot)$ is concave, and the following estimates hold.

$$-C_1|y|^\sigma \le \ell(x,y) \le \ell_1(x)\, \eta(|y|), \ |\ell'_y(x,y)| \le \ell_2(x)\, \eta(|y|), \ |L(x,g)| \le L_1(x)\, \eta(|g|),$$

where $\ell_1 \in L^1(\Omega)$, $L_1 \in L^1(\Omega)$, $\ell_2 \in L^m(\Omega)$; $m > 1$ and σ are the same exponents as in (A3), and η is as in (A2).

2.2. Statement of the main results. Let us define the Hamiltonian functions:

$$\mathcal{H}_Q(x,t,u,p) = H(x,t,u) - pu \qquad \text{for all } (x,t,u,p) \in Q \times \mathbb{R}^2,$$

$$\mathcal{H}_\Omega(x,y,w,p) = L(x,w) - pw \qquad \text{for all } (x,w,p) \in \Omega \times \mathbb{R}^2.$$

The following result provides necessary optimality conditions (as a Pontryagin's principle) for solutions to $(\mathcal{P}_g)$, where $g \in G_{ad}$ is fixed.

THEOREM 2.1. *Suppose that* (A1)–(A4) *are fulfilled. For any $g \in G_{ad}$, let u_g be a solution of $(\mathcal{P}_g)$. Then there exists $p_g \in L^1(0,T;W_o^{1,1}(\Omega))$, such that*

$$\begin{cases} -\dfrac{\partial p_g}{\partial t} + A\,p_g + \Phi'_y(x,t,y(u_g,g))\,p_g + F'_y(x,t,y(u_g,g)) = 0 & \text{in } Q, \\ \\ p_g(x,T) + \ell'_y(x,y(u_g,g)(T)) = 0 & \text{in } \Omega, \end{cases} \tag{2.1}$$

and

$$\mathcal{H}_Q(x,t,u_g(x,t),p_g(x,t)) = \min_{u \in K_U(x,t)} \mathcal{H}_Q(x,t,u,p_g(x,t)) \quad \text{for almost every } (x,t) \in Q. \tag{2.2}$$

Next we are concerned with necessary optimality conditions for solutions to problem $(\mathcal{P})$.

THEOREM 2.2. *Assume* (A1)–(A4). *Then $(\mathcal{P})$ admits at least a solution $\bar g$. In addition, there exists a solution $\bar u$ to $(\mathcal{P}_{\bar g})$, and $\bar p \in L^1(0,T;W_o^{1,1}(\Omega))$, such that*

$$\begin{cases} -\dfrac{\partial \bar p}{\partial t} + A\bar p + \Phi'_y(x,t,y(\bar u,\bar g))\,\bar p + F'_y(x,t,y(\bar u,\bar g)) = 0 & \text{in } Q, \\ \\ \bar p(x,T) + \ell'_y(x,y(\bar u,\bar g)(T)) = 0 & \text{in } \Omega, \end{cases} \tag{2.3}$$

and

$$\mathcal{H}_Q(x,t,\bar u(x,t),\bar p(x,t)) = \min_{u \in K_U(x,t)} \mathcal{H}_Q(x,t,u,\bar p(x,t)) \quad \text{for a.e. } (x,t) \in Q, \tag{2.4}$$

$$\mathcal{H}_\Omega(x,\bar g(x),\bar p(x,0)) = \max_{g \in K_G(x)} \mathcal{H}_\Omega(x,g,\bar p(x,0)) \quad \text{for a.e. } x \in \Omega. \tag{2.5}$$

In the case of linear equations, and when the cost functional is convex with respect to the state variable, a more accurate statement is given below.

THEOREM 2.3. *Suppose that* (A1)–(A4) *are fulfilled. Suppose in addition that Φ is of the form $\Phi(\cdot,y) = a(\cdot)y + b(\cdot)$ (with $a \in L^q(Q)$, $a \ge 0$, $b \in L^q(Q)$), and that $F(x,t,\cdot)$ and $\ell(x,\cdot)$ are convex. Let $\bar g$ be a solution of (P) and let $u_{\bar g}$ be in $Arg\inf(P_{\bar g})$. Then there exists $\bar p \in L^1(0,T;W_0^{1,1}(\Omega))$ satisfying the equation*

$$\begin{cases} -\dfrac{\partial \bar p}{\partial t} + A\bar p + \Phi'_y(x,t,y(u_{\bar g},\bar g))\,\bar p + F'_y(x,t,y(u_{\bar g},\bar g)) = 0 & \text{in } Q, \\ \\ \bar p(x,T) + \ell'_y(x,y(u_{\bar g},\bar g)(T)) = 0 & \text{in } \Omega, \end{cases} \tag{2.6}$$

and such that

$$\mathcal{H}_Q(x,t,u_{\bar g}(x,t),\bar p(x,t)) = \min_{u \in K_U(x,t)} \mathcal{H}_Q(x,t,u,\bar p(x,t)) \text{ for a.e. } (x,t) \in Q, \tag{2.7}$$

$$\mathcal{H}_\Omega(x,\bar g(x),\bar p(x,0)) = \max_{g \in K_G(x)} \mathcal{H}_\Omega(x,g,\bar p(x,0)) \text{ for a.e. } x \in \Omega. \tag{2.8}$$

3. State equation. Adjoint equation.

3.1. Existence and regularity for the solution of state equation. Let a be a nonnegative function in $L^q(Q)$, let ϕ be in $L^q(Q)$, let f be in $L^\infty(\Sigma)$, and let w be in $L^\infty(\Omega)$. Consider the following equation.

$$\frac{\partial y}{\partial t} + Ay + ay = \phi \ \text{ in } Q, \qquad y = f \ \text{ on } \Sigma, \qquad y(0) = w \ \text{ in } \Omega. \tag{3.1}$$

DEFINITION 3.1. *A function $y \in L^1(Q)$ is a weak solution of* (3.1) *if and only if $a\,y \in L^1(Q)$ and*

$$\int_Q y \left(-\frac{\partial z}{\partial t} + A\,z + a\,z\right) dx\,dt = \int_Q \phi\,z\,dx\,dt + \int_\Omega w\,z(0)\,dx - \int_\Sigma f\,\frac{\partial z}{\partial n_A} ds\,dt$$

for all $z \in \mathcal{C}^2(\overline{Q})$ such that $z(T) = 0$ and $z_{|\Sigma} = 0$.

PROPOSITION 3.1 (see [2, Proposition 3.6]). *Equation* (3.1) *admits a unique weak solution $y \in L^1(Q)$. This solution belongs to $\mathcal{C}_b(Q \cup \Omega_T)$ and satisfies*

$$\|y\|_{\infty,Q} \le C(\|\phi\|_{q,Q} + \|f\|_{\infty,\Sigma} + \|w\|_{\infty,\Omega}),$$

where $C \equiv C(T, \Omega, N, q)$ does not depend on a.

PROPOSITION 3.2 (see [2, Proposition 3.7]). *Let a be a nonnegative function in $L^q(Q)$ such that $\|a\|_{q,Q} \le M$. For every $\tau > 0$, the weak solution y of* (3.1) *is Hölder continuous on $\overline{Q^\tau}$ and satisfies*

$$\|y\|_{\mathcal{C}^{\nu,\nu/2}(\overline{Q^\tau})} \le C(\tau)(\|\phi\|_{q,Q} + \|f\|_{\infty,\Sigma} + \|w\|_{\infty,\Omega}) \quad \text{for some } 0 < \nu < 1,$$

where $C(\tau) \equiv C(T, \Omega, N, M, q, \tau)$.

Now we recall some results for the (nonlinear) state equation.

DEFINITION 3.2. *A function $y \in L^1(Q)$ is a weak solution of* (1.1) *if and only if $\Phi(\cdot, y(\cdot)) \in L^1(Q)$ and*

$$\int_Q y \left(-\frac{\partial z}{\partial t} + A\,z\right) dx\,dt + \int_Q (\Phi(x,t,y) - u)\,z\,dx\,dt = -\int_\Sigma \psi \frac{\partial z}{\partial n_A} ds\,dt + \int_\Omega (y_o + g)\,z(0)\,dx$$

for all $z \in \mathcal{C}^2(\overline{Q})$ satisfying $z(T) = 0$ and $z_{|\Sigma} = 0$.

THEOREM 3.1 (see [2, Theorem 3.9]). *Let u be in $L^q(Q)$, let g be in $L^\infty(\Omega)$, and let y_o be in $L^\infty(\Omega)$. Equation* (1.1) *admits a unique weak solution $y(u,g)$. This solution belongs to $\mathcal{C}_b(Q \cup \Omega_T)$ and satisfies*

$$\|y(u,g)\|_{\infty,Q} \le C(\|u\|_{q,Q} + \|\psi\|_{\infty,\Sigma} + \|g\|_{\infty,\Omega} + \|y_o\|_{\infty,\Omega} + 1), \tag{3.2}$$

where $C = C(T, \Omega, N, q)$.

THEOREM 3.2 (see [2, Theorem 3.10]). *For every $M > 0$ and every $\tau > 0$, there exists a positive constant $C(\tau) \equiv C(T, \Omega, N, q, \tau, M)$ and $\nu > 0$ such that, for every $(u,g) \in L^q(Q) \times L^\infty(\Omega)$ satisfying $\|u\|_{q,Q} + \|g\|_{\infty,\Omega} \le M$, the weak solution $y(u,g)$ of* (1.1) *corresponding to (u,g) is Hölder continuous on $\overline{Q^\tau}$ and satisfies*

$$\|y(u,g)\|_{C^{\nu,\nu/2}(\overline{Q^\tau})} \le C(\tau).$$

3.2. Adjoint equation. Consider the following terminal boundary value problem.

$$-\frac{\partial p}{\partial t} + A\,p + a\,p = \phi \ \text{ in } Q, \qquad p = 0 \text{ on } \Sigma, \qquad p(T) = w \ \text{ in } \Omega_T, \tag{3.3}$$

where a is a nonnegative function in $L^q(Q)$, $\phi \in L^1(Q)$, and $w \in L^1(\Omega)$.

DEFINITION 3.3. *A function $p \in L^1(0,T;W_o^{1,1}(\Omega))$ is a weak solution of* (3.3) *if and only if $a\,p \in L^1(Q)$ and*

$$\int_Q \left(p\,\frac{\partial z}{\partial t} + \sum_{i,j=1}^{N} a_{ij} D_j p\, D_i z + a\,z\,p \right) dx\,dt = \int_Q \phi z\,dx\,dt + \int_\Omega w z(T)\,dx$$

for all $z \in \mathcal{C}^1(\overline{Q}) \cap \mathcal{C}_o(Q \cup \Omega_T)$.

THEOREM 3.3. (see [2, Theorem 4.2]) *Let $\phi \in L^1(Q)$, $w \in L^1(\Omega)$, and let a be a nonnegative function in $L^q(Q)$ satisfying $\|a\|_{q,Q} \le M$. Then* (3.3) *admits a unique weak solution p in $L^1(0,T;W_o^{1,1}(\Omega))$. This solution belongs to $L^\delta(0,T;W_o^{1,d}(\Omega))$ for all (δ,d) satisfying $\delta > 1$, $d > 1$, $\frac{N}{2d} + \frac{1}{\delta} > \frac{N+1}{2}$. Moreover, there exists a function in $L^1(\Sigma)$, denoted by $\frac{\partial p}{\partial n_A}$, and a function in $L^1(\Omega)$, denoted by $p(0)$, such that*

$$\int_Q \left(\frac{\partial z}{\partial t} + A\,z + a\,z\right) p\,dx\,dt - \int_\Sigma z\,\frac{\partial p}{\partial n_A} ds\,dt + \int_\Omega z(0)\,p(0)\,dx = \int_Q \phi z\,dx\,dt + \int_\Omega w z(T)\,dx$$

for all $z \in \left\{ \ y \in \mathcal{C}_b(Q \cup \Omega_T) | \frac{\partial y}{\partial t} + A\,y \in L^q(Q),\ y_{|\Sigma} \in L^\infty(\Sigma),\ y(0) \in L^\infty(\Omega) \ \right\}$.

4. Existence results for min and max problems. Before dealing with existence results for the different problems, we have to establish some lower semicontinuity properties for the cost functional. As the state variable is implicitly involved in the cost functional, the following result is crucial for what follows.

THEOREM 4.1. *For every $\tau \in]0,T[$, the mapping $(u,g) \mapsto y(u,g)$ is sequentially continuous from $U_{ad} \times G_{ad}$, endowed with its weak-$L^q(Q)\times$ weak-star-$L^\infty(\Omega)$ topology, into $\mathcal{C}(\overline{Q^\tau})$.*

Proof. Let $(u_n, g_n)_n$ be a sequence converging to $(\tilde{u}, \tilde{g})$ for the weak-$L^q(Q)\times$ weak-star-$L^\infty(\Omega)$ topology. Let y_n be the solution of (1.1), corresponding to (u_n, g_n). Due to (3.2), the sequence $(y_n)_n$ is bounded in $L^\infty(Q)$. Then there exists a subsequence, still indexed by n, and $\tilde{y} \in L^\infty(Q)$ such that $(y_n)_n$ converges to $\tilde{y}$ for the weak-star topology of $L^\infty(Q)$. Moreover, from Theorem 3.2, the sequence $(y_n)_n$ is bounded in $\mathcal{C}^{\nu,\nu/2}(\overline{Q^\tau})$ for some $\nu > 0$ and for all $\tau \in]0,T[$. Since the imbedding from $\mathcal{C}^{\nu,\nu/2}(\overline{Q^\tau})$ into $\mathcal{C}(\overline{Q^\tau})$ is compact, $(y_n)_n$ converges to $\tilde{y}$ uniformly on $\overline{Q^\tau}$ for all $\tau > 0$. On the other hand, observe that y_n satisfies

$$\begin{aligned}\int_Q y_n \left(-\frac{\partial z}{\partial t} + A\,z\right) dx\,dt + \int_Q (\Phi(x,t,y_n) - u_n)\ z\,dx\,dt \\ = -\int_\Sigma \psi \frac{\partial z}{\partial n_A}\,ds\,dt + \int_\Omega (y_o + g_n)\,z(0)\,dx\end{aligned}$$

for all $z \in \mathcal{C}^2(\overline{Q})$ satisfying $z(T) = 0$ and $z_{|\Sigma} = 0$. With (A2) on Φ and Lebesgue's theorem, we can pass to the limit when n tends to infinity, and we obtain

$$\int_Q \tilde{y} \left(-\frac{\partial z}{\partial t} + A\,z\right) dx\,dt + \int_Q (\Phi(x,t,\tilde{y}) - \tilde{u})\ z\,dx\,dt$$

$$= - \int_\Sigma \psi \frac{\partial z}{\partial n_A} ds\, dt + \int_\Omega (y_o + \tilde{g})\, z(0)\, dx$$

for all $z \in \mathcal{C}^2(\overline{Q})$ satisfying $z(T) = 0$ and $z_{|\Sigma} = 0$. Therefore, $\tilde{y}$ is the solution of (1.1) corresponding to $(\tilde{u}, \tilde{g})$. ∎

COROLLARY 4.1. *For all $g \in G_{ad}$, the mapping $u \mapsto J(u, g)$ is sequentially lower semicontinuous on U_{ad}, endowed with the weak-topology of $L^q(Q)$.*

Proof. Let g be in G_{ad}, and let $(u_n)_n$ be a sequence converging to some $\tilde{u}$ for the weak topology of $L^q(Q)$. Let $y_n = y(u_n, g)$ and $\tilde{y} = y(\tilde{u}, g)$ be the associated solutions of (1.1). We observe that

$$\begin{aligned} &J(\tilde{u}, g) - J(u_n, g) \\ &= \int_Q F_n(x,t)(\tilde{y} - y_n)\, dx\, dt + \int_\Omega \ell_n\, (\tilde{y} - y_n)\, (T)\, dx + \int_Q (H(x,t,\tilde{u}) - H(x,t,u_n))\, dx\, dt, \end{aligned}$$

where

$$\begin{aligned} F_n(x,t) &= \int_0^1 F'_y\, (x, t, (1-\theta)\, y_n + \theta \tilde{y})\ d\theta \\ \text{and } \ell_n(x) &= \int_0^1 \ell'_y\, (x, (1-\theta) y_n(T) + \theta\, \tilde{y}(T))\ d\theta. \end{aligned}$$

Due to Theorem 4.1, the sequence $(y_n)_n$ converges to $\tilde{y}$ uniformly on $\overline{Q^\tau}$ for all $\tau > 0$. With assumptions on F and ℓ and Lebesgue's theorem, we obtain

$$\lim_{n \to +\infty} \int_Q F_n(x,t)\, (\tilde{y} - y_n)\, dx\, dt = \lim_{n \to +\infty} \int_\Omega \ell_n(x)\, (\tilde{y} - y_n)(T)\, dx = 0\,. \tag{4.1}$$

On the other hand, from [9, Theorem 2.1, Chapter 8], we deduce that

$$\int_Q H(x,t,\tilde{u})\, dx\, dt \leq \liminf_{n \to +\infty} \int_Q H(x,t,u_n)\, dx\, dt\,. \tag{4.2}$$

The sequential lower semicontinuity of $J(\cdot, g)$ follows from (4.1) and (4.2). ∎

Now we are able to give an existence result of solutions to problem $(\mathcal{P}_g)$.

THEOREM 4.2. *Let g be in G_{ad}. If* (A1)–(A4) *are fulfilled, then problem $(\mathcal{P}_g)$ admits at least one solution.*

Proof. Let g be in G_{ad}, and let u be in U_{ad}. From (A3)–(A4) and Theorem 3.1, it follows that

$$\begin{aligned} J(u,g) \geq\ & C_1 ||u||^q_{q,Q} - C(||u||^\sigma_{q,Q} + ||\Psi||^\sigma_{\infty,\Sigma} + ||y_0||^\sigma_{\infty,\Omega} + ||g||^\sigma_{\infty,\Omega} + 1) \\ & - ||L_1||_{1,\Omega}\ ||\eta(g + y_0)||_{\infty,\Omega}. \end{aligned} \tag{4.3}$$

With Young's inequality we can prove that the infimum of $(\mathcal{P}_g)$ belongs to $\mathbb{R}$. Let $(u_n)_n$ be a minimizing sequence for $(\mathcal{P}_g)$. Due to (4.3), the sequence $(u_n)_n$ is bounded in $L^q(Q)$. Then there exist a subsequence, still indexed by n, and $u \in U_{ad}$, such that $(u_n)_n$ converges to u for the weak topology of $L^q(Q)$. Due to Corollary 4.1, we have

$$J(u,g) \leq \liminf_n J(u_n, g) \leq \inf(\mathcal{P}_g),$$

and the proof is complete. ∎

To study the existence of solutions to problem $(\mathcal{P})$, we have to set some continuity results.

PROPOSITION 4.1. *Let $(g_n)_n \subset G_{ad}$ be a sequence converging to some $g \in G_{ad}$ for the weak-star topology of $L^\infty(\Omega)$. Let u_{g_n} be an element in Argmin$(\mathcal{P}_{g_n})$. Then there exists $u_g \in$ Argmin$(\mathcal{P}_g)$ such that the following conditions hold.*

- *$(u_{g_n})_n$ converges (up to a subsequence) to u_g for the weak topology of $L^q(Q)$.*
- *$(y(u_{g_n}, g_n))_n$ converges (up to a subsequence) to $y(u_g, g)$ in $\mathcal{C}(\overline{Q^\tau})$ for all $\tau > 0$.*

Proof. Due to (4.3), since $(g_n)_n$ is bounded in $L^\infty(\Omega)$, and since $J(u_{g_n}, g_n) \leq J(u_o, g_n) \leq M$ (for some u_o fixed in U_{ad}), the sequence $(u_{g_n})_n$ is bounded in U_{ad}. Then there exists a subsequence, still indexed by n, and $\tilde{u}$, such that $(u_{g_n})_n$ converges to $\tilde{u}$ for the weak topology of $L^q(Q)$. Since U_{ad} is closed and convex in $L^q(Q)$, it is also weakly closed and $\tilde{u} \in U_{ad}$. Due to Theorem 4.1, the sequence $(y(u_{g_n}, g_n))_n$ converges to $y(\tilde{u}, g)$ in $\mathcal{C}(\overline{Q^\tau})$ for all $\tau > 0$. Let us prove that $\tilde{u}$ belongs to Argmin $(\mathcal{P}_g)$. By definition, u_{g_n} satisfies

$$J(u_{g_n}, g_n) \leq J(u, g_n) \qquad \text{for all } u \in U_{ad}. \tag{4.4}$$

With arguments similar to those used in Corollary 4.1, we can prove that

$$\lim_{n\to+\infty} \int_Q F(x,t,y(u_{g_n}, g_n))\,dx\,dt = \int_Q F(x,t,y(\tilde{u}, g))\,dx\,dt, \tag{4.5}$$

$$\lim_{n\to+\infty} \int_Q F(x,t,y(u, g_n))\,dx\,dt = \int_Q F(x,t,y(u, g))\,dx\,dt\,, \tag{4.6}$$

$$\lim_{n\to+\infty} \int_\Omega \ell(x,y(u_{g_n}, g_n)(T))\,dx = \int_\Omega \ell(x,y(\tilde{u}, g)(T))\,dx\,, \tag{4.7}$$

$$\lim_{n\to+\infty} \int_\Omega \ell(x,t,y(u, g_n)(T))\,dx = \int_\Omega \ell(x,y(u, g)(T))\,dx\,, \tag{4.8}$$

$$\int_Q H(x,t,\tilde{u})\,dx\,dt \leq \liminf_{n\to+\infty} \int_Q H(x,t,u_{g_n})\,dx\,dt\,. \tag{4.9}$$

From (4.5)–(4.9) and passing to the limit when n tends to infinity in (4.4), we obtain

$$J(\tilde{u}, g) \leq J(u, g) \qquad \text{for all } u \in U_{ad}.$$

Therefore, $\tilde{u}$ belongs to Argmin $(\mathcal{P}_g)$, and we set $\tilde{u} \equiv u_g$. The second assertion follows from Corollary 4.1. ∎

THEOREM 4.3. *Assume* (A1)–(A4) *are fulfilled; then the minimax problem $(\mathcal{P})$ admits at least one solution $\bar{g}$.*

Proof. First we observe that problem $(\mathcal{P})$ is equivalent to

$$\max\ \{J(u_g, g) \mid u_g \in \text{Argmin}\,(\mathcal{P}_g)\ g \in G_{ad}\}.$$

Let $(g_n)_n$ be a maximizing sequence for $(\mathcal{P})$. Since G_{ad} is bounded in $L^\infty(\Omega)$, there exist a subsequence, still indexed by n, and $\tilde{g}$, such that $(g_n)_n$ converges to $\tilde{g}$ for the

weak-star topology of $L^\infty(\Omega)$. In addition, $\tilde{g}$ is also the weak limit of g_n in $L^s(\Omega)$ for all $s \geq 1$. Since G_{ad} is convex and closed in $L^s(\Omega)$, it is also weakly closed and $\tilde{g} \in G_{ad}$. From assumption (A4) on L, we deduce that

$$\limsup_{n\to+\infty} \int_\Omega (L(x, g_n) - L(x, \tilde{g}))\, dx \leq 0. \tag{4.10}$$

On the other hand, with Proposition 4.1, we know that $(u_{g_n})_n$ (where $u_{g_n} \in \text{Argmin } (\mathcal{P}_{g_n})$) converges (up to a subsequence) to some $u_{\tilde{g}} \in \text{Argmin } (\mathcal{P}_{\tilde{g}})$ for the weak topology of $L^q(Q)$, and $(y(u_{g_n}, g_n))_n$ converges (up to a subsequence) to $y(u_{\tilde{g}}, \tilde{g})$ in $\mathcal{C}(\overline{Q^\tau})$ for all $\tau > 0$. Therefore,

$$\begin{aligned} J(u_{g_n}, g_n) - J(u_{\tilde{g}}, \tilde{g}) &= (J(u_{g_n}, g_n) - J(u_{\tilde{g}}, g_n)) + (J(u_{\tilde{g}}, g_n) - J(u_{\tilde{g}}, \tilde{g})) \\ &\leq J(u_{\tilde{g}}, g_n) - J(u_{\tilde{g}}, \tilde{g}) \\ &= \int_Q (F(x, t, y(u_{\tilde{g}}, g_n)) - F(x, t, y(u_{\tilde{g}}, \tilde{g})))\, dx\, dt \\ &\quad + \int_\Omega (\ell(x, y(u_{\tilde{g}}, g_n)(T)) - \ell(x, y(u_{\tilde{g}}, \tilde{g})(T)))\, dx \\ &\quad + \int_\Omega (L(x, g_n) - L(x, \tilde{g}))\, dx\ . \end{aligned} \tag{4.11}$$

As in the proof of Proposition 4.1, we have

$$\lim_{n\to+\infty} \int_Q (F(x, t, y(u_{\tilde{g}}, g_n)) - F(x, t, y(u_{\tilde{g}}, \tilde{g})))\, dx\, dt = 0, \tag{4.12}$$

$$\lim_{n\to+\infty} \int_\Omega (\ell(x, y(u_{\tilde{g}}, g_n)(T)) - \ell(x, y(u_{\tilde{g}}, \tilde{g})(T)))\, dx = 0. \tag{4.13}$$

Consequently, with (4.11), (4.10), (4.12), and (4.13), it follows that

$$\begin{aligned} \sup_{g\in G_{ad}} \{J(u_g, g) \mid u_g \in \text{Arg inf}(P_g)\} &= \lim_{n\to+\infty} J(u_{g_n}, g_n) \\ &= \limsup_{n\to+\infty} J(u_{g_n}, g_n) \leq J(u_{\tilde{g}}, \tilde{g}). \end{aligned}$$

Since $\tilde{g}$ belongs to G_{ad} and $u_{\tilde{g}}$ belongs to $\text{Arg inf}(P_{\tilde{g}})$, $\tilde{g}$ is a solution to $(\mathcal{P})$. ∎

5. Taylor expansions. In order to give optimality conditions we need some Taylor expansions. In what follows, $\mathcal{L}^{N+1}$ denotes the $(N+1)$-dimensional Lebesgue measure.

THEOREM 5.1 (Taylor expansion of y with respect to u). *Let ρ be in $]0, 1[$. For every u_1, $u_2 \in U_{ad}$, and every $g \in G_{ad}$, there exist measurable subsets $Q_\rho \subset Q$ such that*

$$\mathcal{L}^{N+1}(Q_\rho) = \rho \mathcal{L}^{N+1}(Q),$$

$$\int_{Q_\rho} (H(x, t, u_2) - H(x, t, u_1))\, dx\, dt = \rho \int_Q (H(x, t, u_2) - H(x, t, u_1))\, dx\, dt,$$

$$y_\rho = y_1 + \rho z + r_\rho, \quad \textit{with} \quad \lim_{\rho\to 0} \frac{1}{\rho} ||r_\rho||_{\mathcal{C}(\overline{Q})} = 0,$$

$$J(u_\rho, g) = J(u_1, g) + \rho \left(J'_y(u_1, g_1)z + \int_Q (H(x,t,u_2) - H(x,t,u_1))\, dx\, dt \right) + o(\rho),$$

where

$$u_\rho(x,t) = \begin{cases} u_1(x,t) & \text{in } Q \setminus Q_\rho, \\ \\ u_2(x,t) & \text{in } Q_\rho, \end{cases}$$

y_ρ and y_1 are the solutions of (1.1) *corresponding to (u_ρ, g) and (u_1, g), and z is the solution of*

$$\frac{\partial z}{\partial t} + Az + \Phi'_y(\cdot, y_1)z = u_2 - u_1 \text{ in } Q, \qquad z = 0 \text{ on } \Sigma, \qquad z(0) = 0 \text{ in } \Omega.$$

Proof. See [16, Theorem 4.1] or [15]. □

THEOREM 5.2 (Taylor expansion of y with respect to g). *Let ρ be in $]0,1[$, and set $\tau_\rho = \rho^{m' + \frac{q\bar{q}}{q-\bar{q}}}$, where $q > \bar{q} > \frac{N}{2} + 1$ ($m > 1$ is the exponent in* (A4)*, and m' is the conjugate exponent to m). For every g_1, $g_2 \in G_{ad}$, there exist measurable subsets $\Omega_\rho \subset \Omega$, and there exists $u_{g_1} \in Arg\inf\ (P_{g_1})$, such that the following hold.*

$$\mathcal{L}^N(\Omega_\rho) = \rho \mathcal{L}^N(\Omega),$$

$$\int_{\Omega_\rho} (L(x,g_2) - L(x,g_1))\, dx = \rho \int_\Omega (L(x,g_2) - L(x,g_1))\, dx, \tag{5.1}$$

$$y(u_{g_\rho}, g_\rho) = y(u_{g_\rho}, g_1) + \rho z_1 + r_\rho, \text{ with } \lim_{\rho \to 0} \frac{1}{\rho} ||r_\rho||_{\mathcal{C}(\overline{Q}_{\tau_\rho T})} = 0, \tag{5.2}$$

$$J(u_{g_\rho}, g_\rho) = J(u_{g_\rho}, g_1) + \rho \left(J'_y(u_{g_1}, g_1)z_1 + \int_\Omega (L(x,g_2) - L(x,g_1))\, dx \right) + o(\rho), \tag{5.3}$$

where

$$g_\rho(x) = \begin{cases} g_1(x) & \text{on } \Omega \setminus \Omega_\rho, \\ \\ g_2(x) & \text{on } \Omega_\rho, \end{cases} \qquad u_{g_\rho} \in Arg\inf\ (P_{g_\rho}),$$

$y(u_{g_\rho}, g_\rho)$ and $y(u_{g_\rho}, g_1)$ are the solutions of (1.1) *corresponding to (u_ρ, g_ρ) and (u_ρ, g_1), and z_1 is the solution of*

$$\frac{\partial z}{\partial t} + Az + \Phi'_y(x, t, y(u_{g_1}, g_1))\, z = 0 \text{ in } Q, \qquad z = 0 \text{ on } \Sigma, \qquad z(0) = g_2 - g_1 \text{ in } \Omega.$$

The proof is based on the following lemmas.

LEMMA 5.1. (see [16, Lemma 4.1]) *Let g_1, g_2 be in $L^\infty(\Omega)$. For every $\rho \in]0,1[$, there exists a sequence of measurable subsets $(\Omega_\rho^n)_n$ in Ω, such that*

$$\mathcal{L}^N(\Omega_\rho^n) = \rho \mathcal{L}^N(\Omega),$$

$$\int_{\Omega_\rho^n} (L(x,g_2) - L(x,g_1))dx = \rho \int_\Omega (L(x,g_2) - L(x,g_1))dx,$$

$$\left(\frac{\chi_{\Omega_\rho^n}}{\rho}\right)_n \text{ converges to } 1 \text{ for the weak-star topology of } L^\infty(\Omega),$$

where $\chi_{\Omega_\rho^n}$ *denotes the characteristic function of* Ω_ρ^n.

LEMMA 5.2. *Let be* $\rho \in]0,1[$, *and let* $(\Omega_\rho^n)_n$ *be the sequence of measurable subsets defined in Lemma* 5.1. *Set*

$$g_\rho^n(x) = \begin{cases} g_1(x) & \text{on } \Omega \setminus \Omega_\rho^n, \\ g_2(x) & \text{on } \Omega_\rho^n. \end{cases}$$

Then the sequence $(g_\rho^n)_n$ *converges to* $\rho g_2 + (1-\rho)g_1$ *for the weak-star topology of* $L^\infty(\Omega)$.

Proof. Let be $\varphi \in L^1(\Omega)$. We have

$$\left| \int_\Omega g_\rho^n \, \varphi \, dx - \int_\Omega (\rho g_2 + (1-\rho)g_1) \, \varphi \, dx \right| = \left| \int_\Omega (g_\rho^n - \rho g_2 - (1-\rho)g_1) \, \varphi \, dx \right|$$

$$= \rho \left| \int_\Omega \left(\frac{\chi_{\Omega_\rho^n}}{\rho} - 1 \right) (g_2 - g_1) \, \varphi \, dx \right| \to 0 \quad \text{as } n \to \infty. \qquad \square$$

Proof of Theorem 5.2. Let $u_{g_\rho^n}$ be in Argmin $(\mathcal{P}_{g_\rho^n})$. With Lemma 5.2 and Proposition 4.1, we can prove that $(u_{g_\rho^n})_n$ (or at least a subsequence) weakly converges to some $u_\rho \in L^q(Q)$, $u_\rho \in$ Argmin $\left(\mathcal{P}_{(\rho g_2 + (1-\rho)g_1)}\right)$, and

$$\lim_{n\to\infty} \left\| y(u_{g_\rho^n}, g_\rho^n) - y\left(u_\rho, \rho g_2 + (1-\rho)g_1\right) \right\|_{\mathcal{C}(\bar{Q}^\tau)} = 0 \qquad \text{for all } \tau > 0. \tag{5.4}$$

With similar arguments we prove that $(u_\rho)_\rho$ weakly converges to some $u_{g_1} \in$Argmin$(\mathcal{P}_{g_1})$, and

$$\lim_{\rho\to 0} \left\| y(u_{g_1}, g_1) - y\left(u_\rho, \rho g_2 + (1-\rho)g_1\right) \right\|_{\mathcal{C}(\bar{Q}^\tau)} = 0 \qquad \text{for all } \tau > 0. \tag{5.5}$$

Step 1. We first establish (5.2). The function $\zeta_\rho^n = \frac{1}{\rho}\big(y(u_{g_\rho^n}, g_\rho^n) - y(u_{g_\rho^n}, g_1)\big) - z_1$ belongs to$C_b(Q \cup \Omega_T)$, and it is the solution of

$$\frac{\partial \zeta}{\partial t} + A\zeta + \beta_\rho^n \zeta = h_\rho^n \text{ in } Q, \quad \zeta = 0 \text{ on } \Sigma, \quad \zeta(0) = f_\rho^n \text{ in } \Omega,$$

where

$$\beta_\rho^n = \int_0^1 \Phi_y' \left(\cdot, \theta \; y(u_{g_\rho^n}, g_\rho^n) + (1-\theta) \; y(u_{g_\rho^n}, g_1) \right) d\,\theta,$$

$$h_\rho^n = \left(\Phi_y'(\cdot, y(u_{g_1}, g_1)) - \beta_\rho^n\right) z_1, \; f_\rho^n = \left(1 - \frac{1}{\rho}\chi_{\Omega_\rho^n}\right)(g_1 - g_2).$$

We look for $n \in \mathbb{N}^*$ as a function of ρ, say $n(\rho)$, such that

$$\lim_{\rho \searrow 0} \left\| \zeta_\rho^{n(\rho)} \right\|_{\mathcal{C}(\overline{Q}_{\tau_\rho T})} = 0. \tag{5.6}$$

Set $\zeta_\rho^n = \zeta_\rho^{n,1} + \zeta_\rho^{n,2}$, where $\zeta_\rho^{n,1} \in \mathcal{C}(\overline{Q})$ is the solution of

$$\frac{\partial \zeta}{\partial t} + A\zeta + \beta_\rho^n \zeta = h_\rho^n \text{ in } Q, \quad \zeta = 0 \text{ on } \Sigma, \quad \zeta(0) = 0 \text{ in } \Omega,$$

and $\zeta_\rho^{n,2} \in C_b(Q \cup \Omega_T)$ is the solution of

$$\frac{\partial \zeta}{\partial t} + A\zeta + \beta_\rho^n \zeta = 0 \text{ in } Q, \quad \zeta = 0 \text{ on } \Sigma, \quad \zeta(0) = f_\rho^n \text{ in } \Omega.$$

Let $\eta_\rho^n \in C_b(Q \cup \Omega_T)$ be the solution of

$$\frac{\partial \eta}{\partial t} + A\eta + \beta\eta = 0 \text{ in } Q, \quad \eta = 0 \text{ on } \Sigma, \quad \eta(0) = f_\rho^n \text{ in } \Omega,$$

where $\beta(\cdot) = \Phi'_y(\cdot, y(u_{g_1}, g_1))$. The operator $\mathcal{T}$ which associates ζ, the solution of

$$\frac{\partial \zeta}{\partial t} + A\zeta + \beta\zeta = 0 \text{ in } Q, \quad \zeta = 0 \text{ on } \Sigma, \quad \zeta(0) = w \text{ in } \Omega,$$

with $w \in L^\infty(\Omega)$, is continuous from $L^\infty(\Omega)$ into $C^{\nu,\nu/2}(\overline{Q}_{\tau_\rho T})$ for some $0 < \nu < 1$ (see Proposition 3.2). Since the imbedding from $C^{\nu,\nu/2}(\overline{Q}_{\tau_\rho T})$ into $\mathcal{C}(\overline{Q}_{\tau_\rho T})$ is compact, $\mathcal{T}$ can be considered as a compact operator from $L^\infty(\Omega)$ into $\mathcal{C}(\overline{Q}_{\tau_\rho T})$. Since the sequence $(f_\rho^n)_n$ converges to 0 for the weak-star topology of $L^\infty(\Omega)$, we obtain

$$\lim_{n \to \infty} ||\eta_\rho^n||_{\mathcal{C}(\overline{Q}_{\tau_\rho T})} = 0. \tag{5.7}$$

From (5.4) and (5.7), we deduce the existence of some integer $n(\rho)$ such that

$$\left\| y\left(u_{g_\rho^{n(\rho)}}, g_\rho^{n(\rho)}\right) - y\left(u_\rho, \rho g_2 + (1-\rho) g_1\right) \right\|_{\mathcal{C}(\bar{Q}^{\tau_\rho})} + \|\eta_\rho^{n(\rho)}\|_{\mathcal{C}(\overline{Q}_{\tau_\rho T})} \leq \rho. \tag{5.8}$$

On the other hand, the function $\zeta_\rho^{n(\rho),2} - \eta_\rho^{n(\rho)}$ belongs to $\mathcal{C}(\overline{Q})$ and satisfies

$$\frac{\partial \zeta}{\partial t} + A\zeta + \beta_\rho^{n(\rho)} \zeta = (\beta - \beta_\rho^{n(\rho)}) \eta_\rho^{n(\rho)} \text{ in } Q, \quad \zeta = 0 \text{ on } \Sigma, \quad \zeta(0) = 0 \text{ in } \Omega.$$

There exists $C \equiv C(T, \Omega, q, N) > 0$, independent of ρ, such that

$$\|\zeta_\rho^{n(\rho),1}\|_{\mathcal{C}(\overline{Q})} \leq C \|h_\rho^{n(\rho)}\|_{q,Q},$$

$$\|\zeta_\rho^{n(\rho),2} - \eta_\rho^{n(\rho)}\|_{\mathcal{C}(\overline{Q})} \leq C \|(\beta - \beta_\rho^{n(\rho)}) \eta_\rho^{n(\rho)}\|_{\bar{q},Q} \leq C \|\beta - \beta_\rho^{n(\rho)}\|_{q,Q} \, \|\eta_\rho^{n(\rho)}\|_{r,Q},$$

where $\frac{1}{r} = \frac{1}{\bar{q}} - \frac{1}{q}$ ($\bar{q}$ obeys $q > \bar{q} > \frac{N}{2} + 1$; see assumptions). It follows that

$$\begin{aligned} \|\zeta_\rho^{n(\rho),2} - \eta_\rho^{n(\rho)}\|_{\mathcal{C}(\overline{Q})} \\ \leq C \|\beta - \beta_\rho^{n(\rho)}\|_{q,Q} \left(\|\eta_\rho^{n(\rho)}\|_{\mathcal{C}(\overline{Q}_{\tau_\rho T})} + (\mathcal{L}^{N+1}(Q \setminus Q_{\tau_\rho T}))^{\frac{1}{r}} \|\eta_\rho^{n(\rho)}\|_{\infty,Q} \right) \end{aligned}$$

$$\begin{aligned}
&\leq C\|\beta-\beta_\rho^{n(\rho)}\|_{q,Q}\left(\|\eta_\rho^{n(\rho)}\|_{\mathcal{C}(\overline{Q}_{\tau_\rho T})}+\tau_\rho^{\frac{1}{r}}\|f_\rho^{n(\rho)}\|_{\infty,\Omega}\right)\\
&\leq C\|\beta-\beta_\rho^{n(\rho)}\|_{q,Q}\left(\|\eta_\rho^{n(\rho)}\|_{\mathcal{C}(\overline{Q}_{\tau_\rho T})}+\tau_\rho^{\frac{1}{r}}\,\rho^{-1}\|g_1-g_2\|_{\infty,\Omega}\right)\\
&\leq C\|\beta-\beta_\rho^{n(\rho)}\|_{q,Q}\left(\|\eta_\rho^{n(\rho)}\|_{\mathcal{C}(\overline{Q}_{\tau_\rho T})}+\rho^{m'\frac{q-\bar q}{q\bar q}}\right)\\
&\leq C\|\beta-\beta_\rho^{n(\rho)}\|_{q,Q}\left(\ \rho+\rho^{m'\frac{q-\bar q}{q\bar q}}\right).
\end{aligned}$$

Therefore,

$$\begin{aligned}
\|\zeta_\rho^{n(\rho)}\|_{\mathcal{C}(\overline{Q}_{\tau_\rho T})} \leq\ & \|\zeta_\rho^{n(\rho),1}\|_{\mathcal{C}(\overline{Q})}+\|\zeta_\rho^{n(\rho),2}-\eta_\rho^{n(\rho,\tau)}\|_{\mathcal{C}(\overline{Q})}+\|\eta_\rho^{n(\rho)}\|_{\mathcal{C}(\overline{Q}_{\tau_\rho T})}\\
\leq\ & C\left(\|h_\rho^{n(\rho)}\|_{q,Q}+\|\beta-\beta_\rho^{n(\rho)}\|_{q,Q}(\rho+\rho^{m'\frac{q-\bar q}{q\bar q}})+\rho\right).
\end{aligned}$$

With assumption (A2) and with Lebesgue's theorem of dominated convergence, we can prove that $\left(h_\rho^{n(\rho)}\right)_\rho$ converges to 0 in $L^q(Q)$. Thus (5.6) follows from the last inequality. Let us set $\Omega_\rho=\Omega_\rho^{n(\rho)}$ and $g_\rho=g_\rho^{n(\rho)}$. We have $y(u_{g_\rho},g_\rho)=y\big(u_{g_\rho^{n(\rho)}},g_\rho^{n(\rho)}\big)$, $y(u_{g_\rho},g_1)=y\big(u_{g_\rho^{n(\rho)}},g_1\big)$, $\frac{r_\rho}{\rho}=\zeta_\rho^{n(\rho)}$, and (5.2) is proved.

Step 2. Now we establish (5.3). First observe that, with (5.8), we have

$$\begin{aligned}
c\|y(u_{g_\rho,g_\rho})-\ y\ (u_{g_1},g_1)\|_{\mathcal{C}(\overline{Q^\tau})} \leq\ & \|y(u_{g_\rho},g_\rho)-y\left(u_\rho,\rho g_2+(1-\rho)g_1\right)\|_{\mathcal{C}(\overline{Q^\tau})}\\
&+\|y(u_{g_1},g_1)-y\left(u_\rho,\rho g_2+(1-\rho)g_1\right)\|_{\mathcal{C}(\overline{Q^\tau})}\\
\leq\ & \rho+\|y(u_{g_1},g_1)-y\left(u_\rho,\rho g_2+(1-\rho)g_1\right)\|_{\mathcal{C}(\overline{Q^\tau})}\ \text{for all } \tau>0.
\end{aligned}$$

Therefore, from (5.5) it follows that

$$\lim_{\rho\to 0}\|y(u_{g_\rho},g_\rho)-y(u_{g_1},g_1)\|_{\mathcal{C}(\overline{Q^\tau})}=0 \text{ for all } \tau>0. \tag{5.9}$$

On the other hand,

$$\left|\frac{J(u_{g_\rho},g_\rho)-J(u_{g_\rho},g_1)}{\rho}-\Delta J\right|$$

$$\leq\left|\int_Q\left(\frac{F(x,t,y(u_{g_\rho},g_\rho))-F(x,t,y(u_{g_1},g_1))}{\rho}-F'_y(x,t,y(u_{g_1},g_1))z_1(x,t)\right)dx\,dt\right|$$

$$+\left|\int_\Omega\left(\frac{\ell(x,y(u_{g_\rho},g_\rho)(T))-\ell(x,y(u_{g_1},g_1)(T))}{\rho}-\ell'_y(x,y(u_{g_1},g_1)(T))\,z_1(x,T)\right)dx\right|$$

$$=I_\rho^1+I_\rho^2.$$

(Due to (5.1), the integrand L does not appear in the above estimate.) From the equation satisfied by $\dfrac{r_\rho}{\rho}=\zeta_\rho^{n(\rho)}$, we deduce that

$$\left\|\frac{r_\rho}{\rho}\right\|_{\infty,Q}\leq C\left(\|h_\rho^{n(\rho)}\|_{q,Q}+\|f_\rho^{n(\rho)}\|_{\infty,\Omega}\right)\leq C\left(\|h_\rho^{n(\rho)}\|_{q,Q}+\frac{1}{\rho}\right).$$

With (A3), we obtain

$$\begin{aligned}
I_\rho^1 \leq\ & \left\| F_\rho \frac{r_\rho}{\rho} \right\|_{1,Q} + \| \left(F_\rho - F'_y(\cdot, y(u_{g_1}, g_1)) \right) z_1 \|_{1,Q} \\
\leq\ & C \left(\|F_\rho\|_{m,Q} \left\| \frac{r_\rho}{\rho} \right\|_{m',Q} + \|F_\rho - F'_y(\cdot, y(u_{g_1}, g_1))\|_{1,Q} \right) \\
\leq\ & C \left(\left\| \frac{r_\rho}{\rho} \right\|_{\mathcal{C}(\overline{Q}_{\tau_\rho T})} + \left[\mathcal{L}^{N+1}(Q \setminus Q_{\tau_\rho T}) \right]^{\frac{1}{m'}} \left\| \frac{r_\rho}{\rho} \right\|_{\infty,Q} + \|F_\rho - F'_y(\cdot, y(u_{g_1}, g_1))\|_{1,Q} \right) \\
\leq\ & C \left(\left\| \frac{r_\rho}{\rho} \right\|_{\mathcal{C}(\overline{Q}_{\tau_\rho T})} + (\tau_\rho)^{\frac{1}{m'}} \left[\|h_\rho^{n(\rho)}\|_{q,Q} + \frac{1}{\rho} \right] + \|F_\rho - F'_y(\cdot, y(u_{g_1}, g_1))\|_{1,Q} \right) \\
\leq\ & C \left(\rho + (\tau_\rho)^{\frac{1}{m'}} \|h_\rho^{n(\rho)}\|_{q,Q} + \rho^{\frac{q\bar{q}}{m'(q-\bar{q})}} + \|F_\rho - F'_y(\cdot, y(u_{g_1}, g_1))\|_{1,Q} \right),
\end{aligned}$$

where

$$F_\rho = \int_0^1 F'_y(\cdot, (1-\theta) y(u_{g_1}, g_1) + \theta y(u_{g_\rho}, g_\rho))\, d\theta,$$

and C is a positive constant independent of ρ. Similarly to the calculus of I_ρ^1, and due to (A4), we prove that

$$I_\rho^2 \leq C \left(\rho + (\tau_\rho)^{\frac{1}{m'}} \|h_\rho^{n(\rho)}\|_{q,Q} + \rho^{\frac{q\bar{q}}{m'(q-\bar{q})}} + \|\ell_\rho(\cdot) - \ell'_y(\cdot, y(u_{g_1}, g_1)(T))\|_{1,\Omega} \right),$$

where

$$\ell_\rho(\cdot) = \int_0^1 \ell'_y(\cdot, (1-\theta) y(u_{g_1}, g_1)(T) + \theta y(u_{g_\rho}, g_\rho)(T))\, d\theta,$$

and where C is a positive constant independent of ρ. Due to (5.9), by assumptions on F and ℓ, and Lebesgue's theorem of dominated convergence, we have

$$\lim_{\rho \to 0} \|F_\rho - F'_y(\cdot, y(u_{g_1}, g_1))\|_{1,Q} = 0 \quad \text{and} \quad \lim_{\rho \to 0} \|L_\rho(T) - L'_y(\cdot, y(u_{g_1}, g_1)(T))\|_{1,\Omega} = 0.$$

Thus $\lim_{\rho \to 0} I_\rho^1 = \lim_{\rho \to 0} I_\rho^2 = 0$, and the proof is complete. ∎

6. Proof of the optimality conditions.

6.1. Proof of Theorem 2.1: Optimality conditions for $(\mathcal{P}_g)$. The proof is similar to the one given in [2, Theorem 2.1]. We rewrite it for the convenience of the reader. Let ρ be in $]0,1[$. Let g be in G_{ad}, u_g in Argmin $(\mathcal{P}_g)$, and u in U_{ad}. Due to Theorem 5.1, there exists a measurable subset Q_ρ such that $\mathcal{L}^{N+1}(Q_\rho) = \rho \mathcal{L}^{N+1}(Q)$, and

$$y(u_\rho(g), g) = y(u_g, g) + \rho z_g + r_\rho, \quad \text{with } \lim_{\rho \to 0} \frac{1}{\rho} \|r_\rho\|_{\mathcal{C}(\overline{Q})} = 0,$$

$$J(u_\rho(g), g) = J(u_g, g) + \rho \Delta J + o(\rho),$$

where $u_\rho(g)$ is defined by

$$u_\rho(g)(x,t) = \begin{cases} u_g(x,t) & \text{in } Q \setminus Q_\rho, \\ u(x,t) & \text{in } Q_\rho, \end{cases}$$

$y(u_\rho(g),g)$ is the solution to (1.1) corresponding to $(u_\rho(g),g)$, z_g is the weak solution of

$$\frac{\partial z}{\partial t} + A\,z + \Phi'_y(x,t,y(u_g,g))\,z = u - u_g \text{ in } Q, \qquad z = 0 \text{ on } \Sigma, \qquad z(0) = 0 \text{ in } \Omega,$$

and

$$\Delta J = \int_Q \left(F'_y(x,t,y(u_g,g))z_g + (H(x,t,u) - H(x,t,u_g))\right)\,dx\,dt + \int_\Omega \ell'_y(x,y(u_g,g))\,z_g(T)\,dx.$$

Since $(u_\rho(g),g)$ is admissible for $(\mathcal{P}_g)$, it follows that $J(u_g,g) \le J(u_\rho(g),g)$ and

$$-\Delta J \le \lim_{\rho\to 0} \frac{J(u_g,g) - J(u_\rho(g),g)}{\rho} \le 0\ . \tag{6.1}$$

Let p_g be the weak solution of (2.1). Using the Green formula of Theorem 3.3, we obtain

$$-\int_Q F'_y(x,t,y(u_g,g))z_g\,dx\,dt - \int_\Omega \ell'_y(x,y(u_g,g)(T))\,z_g(T)\,dx$$

$$= \int_Q p_g\left(\frac{\partial z_g}{\partial t} + A\,z_g + \Phi'_y(x,t,y(u_g,g))\,z_g\right)\,dx\,dt = \int_Q p_g\,(u-u_g)\,dx\,dt.$$

Taking the definition of ΔJ into account, we have

$$\Delta J = \int_Q H(x,t,u) - H(x,t,u_g)\,dx\,dt - \int_Q p_g(x,t)(u(x,t) - u_g(x,t)\,dx\,dt. \tag{6.2}$$

From (6.1) and (6.2), we finally obtain

$$\int_Q \mathcal{H}_Q(x,t,\bar u_g,p_g)\,dx\,dt \le \int_Q \mathcal{H}_Q(x,t,u,p_g)\,dx\,dt \quad \text{for all } u \in U_{ad}.$$

The pointwise Pontryagin's principle (2.2) is next obtained by the method developed in [16, section 5.2]. □

6.2. Proof of Theorem 2.2: Optimality conditions for $(\mathcal{P})$. Let ρ be in $]0,1[$, let τ_ρ be as in Theorem 5.2, and let $g \in G_{ad}$. We recall that $\bar g$ is the optimal solution to $(\mathcal{P})$ that we want to characterize. Due to Theorem 5.2, there exist $\bar u \in$ Argmin $(\mathcal{P}_{\bar g})$ and measurable subsets Ω_ρ such that

$$\mathcal{L}^N(\Omega_\rho) = \rho\mathcal{L}^N(\Omega),$$

$$y(u_{g_\rho}, g_\rho) = y(u_{g_\rho}, \bar{g}) + \rho \bar{z} + \tilde{r}_\rho, \quad \text{with } \lim_{\rho \to 0} \frac{1}{\rho} \|\tilde{r}_\rho\|_{\mathcal{C}(\overline{Q}_{\tau_\rho T})} = 0,$$

$$J(u_{g_\rho}, g_\rho) = J(u_{g_\rho}, \bar{g}) + \rho \left(J'_y(\bar{u}, \bar{g})\bar{z} + \int_\Omega (L(x, g) - L(x, \bar{g}))dx \right) + o(\rho), \tag{6.3}$$

where

$$g_\rho(x) = \begin{cases} \bar{g}(x) & \text{in } \Omega \setminus \Omega_\rho, \\ g(x) & \text{in } \Omega_\rho, \end{cases} \qquad u_{g_\rho} \in \text{ Argmin } (\mathcal{P}_{g_\rho}),$$

and $\bar{z}$ is the solution of the equation

$$\frac{\partial z}{\partial t} + A\, z + \Phi'_y(\cdot, y(\bar{u}, \bar{g}))\, z = 0 \ \text{ in } Q, \ z = 0 \text{ on } \Sigma, \ z(0) = g - \bar{g} \ \text{ in } \Omega\ .$$

Since $\bar{g}$ is a solution to $(\mathcal{P})$ and $\bar{u} \in \text{Argmin}(\mathcal{P}_{\bar{g}})$, we see that

$$0 \le \frac{J(\bar{u}, \bar{g}) - J(u_{g_\rho}, g_\rho)}{\rho} \le \frac{J(u_{g_\rho}, \bar{g}) - J(u_{g_\rho}, g_\rho)}{\rho}.$$

With (6.3), we obtain

$$0 \le \lim_{\rho \to 0} \frac{J(u_{g_\rho}, \bar{g}) - J(u_{g_\rho}, g_\rho)}{\rho}$$

$$= -\int_Q F'_y(x, t, y(\bar{u}, \bar{g}))\, \bar{z}\, dx\, dt - \int_\Omega \ell'_y(x, y(\bar{u}, \bar{g})(T))\, \bar{z}(T)\, dx - \int_\Omega (L(x, g) - L(x, \bar{g}))\, dx\ .$$

Using the Green formula of Theorem 3.3, we have

$$-\int_Q F'_y(x, t, y(\bar{u}, \bar{g}))\, \bar{z}\, dx\, dt - \int_\Omega \ell'_y(x, y(\bar{u}, \bar{g})(T))\, \bar{z}(T)\, dx = \int_\Omega \bar{p}(x, 0)\, (g - \bar{g})(x)\, dx\ .$$

Therefore,

$$\int_\Omega \mathcal{H}_\Omega(x, g, \bar{p}(x, 0))\, dx \le \int_\Omega \mathcal{H}_\Omega(x, \bar{g}, \bar{p}(x, 0))\, dx \text{ for all } g \in G_{ad}.$$

The pointwise Pontryagin's principle (2.5) is obtained by the method developed in [16, section 5.2]. Finally, since $\bar{u}$ belongs to Argmin $(\mathcal{P}_{\bar{g}})$, (2.4) follows from Theorem 2.1. □

6.3. Proof of Theorem 2.3. The optimality condition for $u_{\bar{g}}$ is a direct consequence of Theorem 2.1. Let us prove optimality conditions for $\bar{g}$. Let g be in G_{ad} and u_g be an element of Arg inf(P_g). For ρ in $]0, 1[$, set $y_\rho = (1 - \rho)\ y(u_{\bar{g}}, \bar{g}) + \rho\ y(u_g, g)$. Since Φ is affine with respect to y, y_ρ is the solution of (1.1) corresponding to $((1 - \rho)\ u_{\bar{g}} + \rho\ u_g, (1 - \rho)\ \bar{g} + \rho\ g)$. Due to the convexity of F and ℓ, we have

$$\begin{aligned} \int_Q F(x, t, y_\rho)\, dx\, dt + \int_\Omega \ell(x, y_\rho(T))\, dx & \\ \le\ (1 - \rho) \left(\int_Q F(x, t, y(u_{\bar{g}}, \bar{g}))\, dx\, dt + \int_\Omega \ell(x, y(u_{\bar{g}}, \bar{g})(T))\, dx \right) & \\ +\ \rho \left(\int_Q F(x, t, y(u_g, g))\, dx\, dt + \int_\Omega \ell(x, y(u_g, g)(T))\, dx \right). & \end{aligned}$$

Set

$$J_\rho = \int_Q F(x,t,y_\rho)\,dx\,dt + \int_\Omega \ell(x,y_\rho(T))\,dx + (1-\rho)\left(\int_Q H(x,t,u_{\bar g})\,dx\,dt + \int_\Omega L(x,\bar g)dx\right) + \rho\left(\int_Q H(x,t,u_g)\,dx\,dt + \int_\Omega L(x,g)\,dx\right).$$

With the previous inequality, we obtain

$$J_\rho \le (1-\rho)J(u_{\bar g},\bar g) + \rho J(u_g,g).$$

From the optimality of $\bar g$, we have $J(u_g,g) \le J(u_{\bar g},\bar g)$, and thus

$$J_\rho \le (1-\rho)J(u_{\bar g},\bar g) + \rho J(u_g,g) \le J(u_{\bar g},\bar g).$$

It follows that

$$\begin{aligned} 0 \le \lim_{\rho\to 0}\frac{J(u_{\bar g},\bar g)-J_\rho}{\rho} = {} & \int_Q F'_y(x,t,y(u_{\bar g},\bar g))\ (y(u_{\bar g},\bar g)-y(u_g,g))\ dx\,dt \\ & + \int_\Omega \ell'_y(x,y(u_{\bar g},\bar g)(T)),\ (y(u_{\bar g},\bar g)-y(u_g,g)(T))\ dx \\ & + \int_Q (H(x,t,u_{\bar g})-H(x,t,u_g))\,dx\,dt \\ & + \textstyle\int_\Omega (L(x,\bar g)-L(x,g))\,dx. \end{aligned} \tag{6.4}$$

From (6.4), by using the Green formula of Theorem 3.3 (with $z = y(u_{\bar g},\bar g) - y(u_g,g)$), we obtain

$$\int_Q \bar p(u_{\bar g}-u_g)\,dx\,dt + \int_\Omega \bar p(0)(\bar g - g)\,dx$$

$$\le \int_Q (H(x,t,u_{\bar g})-H(x,t,u_g))\,dx\,dt + \int_\Omega (L(x,\bar g)-L(x,g))\,dx.$$

Recalling (2.7), we deduce that

$$\int_\Omega (L(x,g)-\bar p(0)g)\,dx - \int_\Omega (L(x,\bar g)-\bar p(0)\bar g)\,dx$$

$$\le \int_Q (H(x,t,u_{\bar g})-\bar p u_{\bar g})dx\,dt - \int_Q (H(x,t,u_g)-\bar p u_g)\,dx\,dt \le 0.$$

The proof is complete. □

7. Some extensions. We can also consider models which take into account the first order derivatives of the state variable both in the equation and the cost functional. For example, consider the parabolic equation

$$\begin{aligned} \frac{\partial y}{\partial t} + Ay + \overrightarrow{V}\cdot\nabla y + \Phi(x,t,y) = u \text{ in } Q,\ y = 0 \ \text{ on } \Sigma, \\ y(x,0) = y_o(x) + g(x) \ \text{ in } \Omega, \end{aligned} \tag{7.1}$$

where A, Φ, and u satisfy the assumptions of section 2, $\overrightarrow{V}$ belongs to $L^\infty(0,T;(L^\infty(\Omega))^N)$, and the cost functional is defined by

$$\mathcal{I}(y,u,g) = \int_Q \left(F(\cdot,y) + G(\cdot,\nabla y) + H(\cdot,u)\right)\, dx\, dt + \int_\Omega \left(\ell(\cdot,y(T)) + L(\cdot,g)\right)\, dx. \tag{7.2}$$

The regularity results used in the proof of Theorem 5.2 are still true for the above equation [10], [8]. The adjoint equation corresponding to a solution $\bar{y}$ of (7.1) is

$$\begin{cases} -\dfrac{\partial p}{\partial t} + Ap - \overrightarrow{V}\cdot\nabla p + \Phi'_y(\cdot,\bar{y})\, p + F'_y(\cdot,\bar{y}) - \operatorname{div}\left(G'_z(\cdot,\nabla\bar{y})\right) = 0 & \text{in } Q, \\ \\ p(\cdot,T) + \ell'_y(\cdot,\bar{y}(T)) = 0 & \text{in } \Omega, \end{cases} \tag{7.3}$$

where G'_z denotes the derivative of G with respect to ∇y. If $G'_z(\cdot,\nabla\bar{y})$ belongs to $L^r(Q)$ with $r > 1$, then (7.3) admits a unique solution in $L^1(0,T;W_0^{1,1}(\Omega))$. The condition $G'_z(\cdot,\nabla\bar{y}) \in L^r(Q)$ can be easily checked if G'_z satisfies a suitable growth condition. For $\overrightarrow{V} \equiv 0$ and $\Phi'_y \equiv 0$, the existence to (7.3) can be deduced from [17]. For the general case, that is if $\overrightarrow{V} \not\equiv 0$ and $\Phi'_y \not\equiv 0$, the existence can be shown by using a fixed point technique as in [13]. Therefore, Theorem 2.2 may be extended (with obvious modifications) to problem (P) corresponding to (7.1) and the functional (7.2). Notice that we have considered homogeneous boundary conditions, as in the example of convection-diffusion systems studied in [1]. The analysis corresponding to nonhomogeneous boundary conditions of the form $y = \psi \in L^\infty(\Sigma)$ is more delicate since we only know that ∇y belongs to $L^1_{loc}(Q)$.

For a state equation with a nonlinear term depending on the gradient ∇y, the analysis is more complicated and some additional material is needed (see for example [7, Chapter 4]). The results presented in our paper do not recover this case, and the extension to such models requires another lengthy study of the state equation which cannot be included here.

REFERENCES

[1] N. U. Ahmed and X. Xiang, *Nonlinear uncertain systems and necessary conditions of optimality,* SIAM J. Control Optim., 35 (1997), pp. 1755–1772.

[2] N. Arada and J. P. Raymond, *Dirichlet Boundary Control of Semilinear Parabolic Equations. Part* 1: *Problems with No State Constraints*, UMR CNRS 5640, Report n^o 98-05, Université Paul Sabatier, Touluse, France, 1998.

[3] N. Arada and J. P. Raymond, *Dirichlet Boundary Control of Semilinear Parabolic Equations. Part* 2: *Problems with Pointwise State Constraints,* UMR CNRS 5640, Report n^o 98-06, Université Paul Sabatier, Touluse, France, 1998.

[4] N. Arada, *Minimax Dirichlet boundary control problem with state constraints,* Nonlinear Anal., to appear.

[5] H. T. Banks, R. C. Smith, and Y. Wang, *Smart Material Structures: Modelling, Estimation and Control*, Masson, Paris, 1996.

[6] V. Barbu and Th. Precupanu, *Convexity and Optimization in Banach Spaces*, D. Reidel, Boston, Dordrecht, The Netherlands, 1986.

[7] D. Daners and P. K. Medina, *Abstract Evolution Equations, Periodic Problems and Applications*, Longman, London, 1992.

[8] E. DiBenedetto, *Degenerate Parabolic Equations*, Springer-Verlag, New York, 1993.

[9] I. Ekeland and R. Temam, *Analyse convexe et problèmes variationnels*, Dunod, Paris, 1974.

[10] O. A. Ladyzenskaja, V. A. Solonnikov, and N. N. Ural'ceva, *Linear and quasilinear equations of parabolic type,* Transl. Math. Monogr. 23, AMS, Providence, RI, 1968.

[11] J. L. Lions, *Contrôle à moindres regrets des systèmes distribués,* C. R. Acad. Sci. Paris Sér. I Math., 315 (1992), pp. 1253–1257.

[12] B. Mordukhovich and K. Zhang, *Minimax control of parabolic systems with Dirichlet boundary conditions and state constraints,* Appl. Math. Optim., 36 (1997), pp. 323–360.

[13] P. A. Nguyen and J. P. Raymond, *Control problems for convection-diffusion equations with control localized on manifolds,* in Proceedings of the 19th International Federation for Information Processing Conference on System Modelling and Optimization, Cambridge, UK, 1999.

[14] N. Papageorgiou, *Minimax control of nonlinear evolution equations,* Comment. Math. Univ. Carolin., 36 (1995), pp. 39–56.

[15] J. P. Raymond and H. Zidani, *Hamiltonian Pontryagin's principles for control problems governed by semilinear parabolic equations,* Appl. Math. Optim., 39 (1999), pp. 143–177.

[16] J. P. Raymond and H. Zidani, *Pontryagin's principles for state-constrained control problems governed by parabolic equations with unbounded controls,* SIAM J. Control Optim., 36 (1998), pp. 1853–1879.

[17] V. Vespri, *Analytic semigroups in $H^{-m,p}$ by elliptic variational operators and applications to linear Cauchy problems,* in Semigroup Theory and Applications, P. Clement et al., eds, Marcel Dekker, New York, 1989, pp. 419–431.

SIAM J. CONTROL OPTIM.
Vol. 38, No. 5, pp. 1501–1520

INVARIANT SOLUTIONS OF DIFFERENTIAL GAMES AND HAMILTON–JACOBI–ISAACS EQUATIONS FOR TIME-MEASURABLE HAMILTONIANS*

PIERRE CARDALIAGUET† AND SŁAWOMIR PLASKACZ‡

Abstract. We characterize invariance of time-varying domains with respect to differential games with time-measurable dynamics. We deduce from this result a new definition of viscosity solutions to some first order Hamilton–Jacobi equations with time-measurable Hamiltonians.

Key words. differential games, Hamilton–Jacobi–Isaacs equation, viscosity solutions, time-measurable Hamiltonians

AMS subject classifications. 49J52, 49L25, 90D25

PII. S0363012998296219

1. Introduction. We study, on one hand, invariance of time-varying domains with respect to differential games with time-measurable dynamics (sections 2 and 3) and, on the other hand, first order Hamilton–Jacobi equations with time-measurable Hamiltonians (section 4).

We consider a differential game

$$\begin{cases} x'(t) = f(t, x(t), y(t), z(t)), \\ x(0) = x_0, \end{cases}$$

where f is only measurable with respect to the time. A time-varying domain $t \rightsquigarrow P(t) \subset R^N$ (a tube) is a set-valued map with nonempty closed values. The paper is mainly concerned with the following problem: Under which condition can one of the players prevent the state $x(t)$ of the system from leaving $P(t)$ under any action of the other player? The problem was considered by Cardaliaguet in [7] and by Cardaliaguet–Plaskacz in [8] for differential games with a more regular right-hand side. The time-measurable case was studied by Frankowska–Plaskacz–Rzeżuchowski in [15] and by Frankowska–Plaskacz in [14] for differential inclusions (see also [1], [2]).

In the second part we consider Hamilton–Jacobi–Isaacs equations

$$\frac{\partial V}{\partial t} + H\left(t, x, \frac{\partial V}{\partial x}\right) = 0$$

with a time-measurable Hamiltonian H given by

$$H(t, x, p) = \inf_z \sup_y \langle f(t, x, y, z), p \rangle.$$

We generalize a definition of viscosity solutions and obtain the uniqueness and the existence of such solutions. Our approach to the definition of a solution and our

*Received by the editors March 11, 1998; accepted for publication (in revised form) February 8, 1999; published electronically May 23, 2000.
http://www.siam.org/journals/sicon/38-5/29621.html

†CENTRE V.J.C., E.R.S. 2064, Université Paris IX Dauphine, Place du Maréchal de Lattre de Tassigny, 75775 Paris cedex, France (pierre@cardaliaguet@dauphine.fr).

‡Department of Mathematics and Informatics, Nicholas Copernicus University, Chopina 12/18, 87–100 Toruń, Poland (plaskacz@mat.uni.torun.pl). The research of this author was supported in part by KBN grant PB 801/P03/95/09.

methods of the proof are similar to the ones used by Frankowska for the first time in [12], and later by Frankowska–Plaskacz–Rzezuchowski in [15].

Hamilton–Jacobi equations with time-measurable Hamiltonians were considered also by Ishii in [17], by Lions–Perthame in [18] and by Barron–Jensen in [5]. Our definition of a generalized viscosity solution seems to be more elementary than others collected in [18]. It is difficult to compare them directly. The equivalence of different definitions follows from the uniqueness results which hold true in all cases.

Let us consider a tube $P : [0,T] \rightsquigarrow R^N$ and a dynamic f as previously. As usual in two players differential games, we are interested in the following problems:

The first problem. For any initial position $x_0 \in P(t_0)$, does there exist a nonanticipative strategy[1] α of the first player such that, for any time-measurable control $z(\cdot)$ chosen by the second player, the solution to

$$\begin{cases} x'(t) = f(t, x(t), \alpha(z(\cdot))(t), z(t)), \\ x(t_0) = x_0 \end{cases} \tag{1.1}$$

satisfies

$$x(t) \in P(t) \qquad \forall t \in [t_0, T]?$$

In Theorem 2.2 we show that a left absolutely continuous tube P enjoys the above stated property if and only if it is a discriminating tube, i.e., there exists a set $C \subset [0,T]$ of full measure such that, for any $t \in C$ and any $x \in P(t)$,

$$\forall (n_t, n_x) \in N^0_{\mathrm{Graph}(P)}(t,x), \sup_z \inf_y \langle (n_t, n_x), (1, f(t,x,y,z)) \rangle \leq 0.$$

The second problem. For any initial position $x_0 \in P(t_0)$, for any $\epsilon > 0$ and any nonanticipative strategy α of the first player, does there exist a time-measurable control $z(\cdot)$ of the second player, such that the solution $x(\cdot)$ to (1.1) satisfies

$$x(t) \in P(t) + B(0,\epsilon) \qquad \forall t \in [t_0, T]?$$

In Theorem 3.2 we show that a left absolutely continuous tube P enjoys this property if and only if it is a leadership tube, i.e., there exists a set $C \subset [0,T]$ of full measure such that, for any $t \in C$ and for any $x \in P(x)$,

$$\forall (n_t, n_x) \in N^0_{\mathrm{Graph}(P)}(t,x), \inf_z \sup_y \langle (n_t, n_x), (1, f(t,x,y,z)) \rangle \leq 0.$$

In the second part of the paper we use the above results to prove that the value function of the differential game is the unique solution of the appropriate Hamilton–Jacobi–Isaacs equations. We use and generalize some results of Evans–Souganidis [11]. In [11] it was proved that the value function is a viscosity solution of a Hamilton–Jacobi–Isaacs equation. We prove that the value function is a generalized viscosity solution and also we give an elementary proof (using Theorems 2.2 and 3.2) that a generalized solution of a Hamilton–Jacobi–Isaacs equation is equal to the value function. In this way we prove uniqueness of solution for some Hamilton–Jacobi equations with time-measurable Hamiltonians.

[1] These strategies are also often called Elliot–Kalton's strategies. The definition is recalled below.

2. Discriminating domains. We consider a differential game with dynamics given by $x'(t) = f(t, x(t), y, z)$. By $x(\cdot; t_0, x_0, y(\cdot), z(\cdot))$ we denote the solution of the Cauchy problem

$$\begin{cases} x'(t) = f(t, x(t), y(t), z(t)) \ \text{ for a.e. } \ t \in [0, T], \\ x(t_0) \ = \ x_0, \end{cases}$$

where $y : [0, T] \to Y$, $z : [0, T] \to Z$ are measurable controls (open loops) of player I and II, respectively, and Y, Z are compact metric spaces. We further introduce a set-valued map $P : [0, T] \rightsquigarrow R^n$, i.e., $P(t) \subset R^n$ and $P(t) \neq \emptyset$ for every $t \in [0, T]$, regarded as a time-dependent constraints set or a tube of constraints. A tube P is called left absolutely continuous if there exists an integrable function $\mu : [0, T] \to [0, +\infty)$ such that for every $t_1 < t_2$ we have

$$P(t_1) \subset P(t_2) + B\left(0, \int_{t_1}^{t_2} \mu(s)\, \mathrm{ds}\right),$$

where $B(x_0, r)$ denotes the ball in R^n centered at x_0 with radius r and $A + D = \{a + d \ : \ a \in A \& d \in D\}$ for A, $D \subset R^n$. Given a closed subset K of Euclidean space E, the Bouligand contingent cone $T_K(x)$ to K at $x \in K$ is defined by

$$T_K(x) = \left\{ e \in E \ : \ \liminf_{h \to 0^+} \frac{\text{dist } (x + he, K)}{h} = 0 \right\}.$$

For $T \subset E$ we let $T^\perp$ for the polar cone to T,

$$T^\perp = \{w \in E \ \ \langle w, v \rangle \leq 0, \ \text{for every } v \in T\}.$$

We set

$$N_K^0(x) = T_K^\perp(x)$$

and say that $N_K^0(x)$ is the normal cone to K at $x \in K$.

Let $M_t = \{y : [t, T] \to Y \ \ : \ \ y \text{ is measurable}\}$ and $N_t = \{z : [t, T] \to Z \ \ : \ z \text{ is measurable}\}$. We say that a map $\alpha : N_t \to M_t$ is a nonanticipative strategy if for every controls z_1, $z_2 \in N_t$ such that

$$z_1(s) = z_2(s) \text{ for almost all } s \in [t, \tau]$$

we have

$$\alpha(z_1)(s) = \alpha(z_2)(s) \text{ for almost all } s \in [t, \tau].$$

Let Γ_t denote the set of all nonanticipative strategies $\alpha : N_t \to M_t$.

Definition 2.1 (discriminating tube). *A tube $P : [0, T] \rightsquigarrow R^n$ is a discriminating tube for $f : [0, T] \times R^n \times Y \times Z \to R^n$ if there exists a full measure set $C \subset [0, T]$ such that for every $t \in C$ and every $x \in P(t)$ we have*

$$\begin{gathered} \forall (n_t, n_x) \in N^0_{Graph(P)}(t, x), \ \forall z \in Z, \ \exists y \in Y, \\ \langle (n_t, n_x), (1, f(t, x, y, z)) \rangle \leq 0. \end{gathered} \tag{2.1}$$

THEOREM 2.2. *We assume that a tube $P : [0,T] \rightsquigarrow R^n$ is left absolutely continuous and that a right-hand side $f : [0,T] \times R^n \times Y \times Z \to R^n$ satisfies the following conditions:*

$$f(\cdot, x, y, z) \text{ is measurable for every } x, y, z; \tag{2.2}$$

$$\begin{gathered}\exists l \in L^1(0,T),\ \forall x_1,\ x_2,\ \forall y \in Y,\ \forall z \in Z,\\ \|f(t,x_1,y,z) - f(t,x_2,y,z)\| \le l(t)\|x_1 - x_2\| \quad \text{for a.a. } t \in [0,T];\end{gathered} \tag{2.3}$$

$$f(t,x,\cdot,\cdot) \text{ is continuous for every } t, x; \tag{2.4}$$

$$\exists\, \mu \in L^1(0,T),\ \forall t,x,y,z,\ \ \|f(t,x,y,z)\| \le \mu(t); \tag{2.5}$$

$$\forall (t,x,z) \in [0,T] \times R^n \times Z,\ \ \{f(t,x,y,z) \,:\, y \in Y\} \text{ is convex}. \tag{2.6}$$

If P is a discriminating tube for f, then for each $t_0 \in [0,T]$ and $x_0 \in P(t_0)$

$$\exists \alpha \in \Gamma_{t_0},\ \forall z(\cdot) \in N_{t_0},\ \forall t \in [t_0,T],\ x(t;t_0,x_0,\alpha(z),z) \in P(t). \tag{2.7}$$

Conversely, if for each $t_0 \in [0,T]$ and $x_0 \in P(t_0)$

$$\begin{gathered}\forall \varepsilon > 0,\ \exists \alpha \in \Gamma_{t_0},\ \forall z(\cdot) \in N_{t_0},\ \forall t \in [t_0,T],\\ x(t;t_0,x_0,\alpha(z),z) \in P(t) + B(0,\varepsilon),\end{gathered} \tag{2.8}$$

then P is a discriminating tube for f.

The proof of Theorem 2.2 makes use of a viability result for differential inclusions and a nonexpansive selection theorem in ultrametric spaces. First, we recall a viability result in an appropriate version.

DEFINITION 2.3 (viability tube). *A tube $P : [0,T] \rightsquigarrow R^n$ is a viability tube for $F : [0,T] \times R^n \rightsquigarrow R^n$ if there exists a full measure set $C \subset [0,T]$ such that for every $t \in C$ and every $x \in P(t)$ we have*

$$(\{1\} \times F(t,x)) \cap \overline{co}\left(T_{Graph(P)}(t,x)\right) \neq \emptyset. \tag{2.9}$$

THEOREM 2.4. *Assume that a nonempty closed valued tube $P : [0,T] \rightsquigarrow R^n$ is left absolutely continuous and that a set valued-map $F : [0,T] \times R^n \rightsquigarrow R^n$ satisfies the following conditions:*

$$F(t,x) \text{ is nonempty closed convex;} \tag{2.10}$$

$$d_H(F(t,x_1),F(t,x_2)) \le l(t)|x_1 - x_2| \text{ a.e. in } [0,T],\ l(\cdot) \in L^1[0,T]; \tag{2.11}$$

$$F(\cdot,x) \text{ is measurable;} \tag{2.12}$$

$$\|F(t,x)\| \le \mu(t) \text{ a.e. in } [0,T],\ \ \mu \in L^1(0,T). \tag{2.13}$$

Then the tube P is a viability tube for F if and only if $\forall\ t_0 \in [0,T)$, $x_0 \in P(t_0)$ there exists an absolutely continuous solution $x : [t_0,T] \to R^n$ of

$$\begin{cases} x'(t) \in F(t,x(t)) \ \ \textit{a.e. in } [t_0,T], \\ x(t_0) = x_0, \\ x(t) \in P(t) \ \ \forall t \in [t_0,T]. \end{cases} \tag{2.14}$$

Proof. Fix $t_0 \in [0,T)$, $x_0 \in P(t_0)$. Define

$$g(t) = \inf\{\operatorname{dist}(x(t),P(t)) \ : \ x \in \operatorname{Sol}_F(t_0,x_0)\},$$

where $\operatorname{Sol}_F(t_0,x_0)$ denotes the set of solutions of Cauchy problem

$$\begin{cases} x' \in F(t,x), \\ x(t_0) = x_0. \end{cases}$$

By Lemma 4.8 in [16], if the tube P is absolutely continuous, then the function g is absolutely continuous. Assuming that P is only left absolutely continuous, we obtain that g has bounded variation and Gronwall inequality holds true for g. Namely, we have the following lemma.

LEMMA 2.5. *If P is left absolutely continuous, then*

(a) $g(t_2) \le g(t_1) + 2\int_{t_1}^{t_2} \mu(s)\,ds$, *for* $t_0 \le t_1 < t_2 \le T$;

(b) *g has bounded variation, in particular g is differentiable a.e. in $[t_0,T]$;*

(c) *if there exists $c \in L^1(t_0,T)$ such that $g'(t) \le c(t)g(t)$, a.e. in $[t_0,T]$, then $g \equiv 0$.*

Proof. Fix $t_1 < t_2$. Since $\operatorname{Sol}_F(t_0,x_0)$ is compact, there exists $x \in \operatorname{Sol}_F(t_0,x_0)$ such that $g(t_1) = \operatorname{dist}(x(t_1),P(t_1))$. We have

$$\begin{gathered} g(t_2) \le \operatorname{dist}(x(t_2),P(t_2)) \\ \le \|x(t_2) - x(t_1)\| + \operatorname{dist}(x(t_1),P(t_1)) + \sup\{\operatorname{dist}(y,P(t_2)) \ : \ y \in P(t_1)\} \\ \le \int_{t_1}^{t_2} \mu(s)\,ds + g(t_1) + \int_{t_1}^{t_2} \mu(s), \end{gathered}$$

which is our assertion (a).

To estimate the variation of g on $[t_0,T]$, we take a partition $t_0 < t_1 < \cdots < t_k = T$. Let $S = \{i \in \{1,2,\dots,k\} \ : \ g(t_i) - g(t_{i-1}) \ge 0\}$ and $S' = \{1,2,\dots,k\} \setminus S$. We have

$$g(t_k) - g(t_0) = \sum_{i=1}^{k} g(t_i) - g(t_{i-1}) = \sum_{i\in S} |g(t_i) - g(t_{i-1})| - \sum_{i\in S'} |g(t_i) - g(t_{i-1})|.$$

Thus

$$\sum_{i=1}^{k} |g(t_i) - g(t_{i-1})| = 2\sum_{i\in S} |g(t_i) - g(t_{i-1})| + g(t_k) - g(t_0) \le 6\int_{t_0}^{T} \mu(s)\,ds,$$

which gives us (b).

We set $h(t) = \sup\{g(s) \ : \ s \in [t_0,t]\}$. The function h is nonnegative, nondecreasing, $g(t) \le h(t)$ for $t \in [t_0,T]$, and

$$h(t_2) - h(t_1) \le 2\int_{t_1}^{t_2} \mu(s)\,ds$$

for $t_1 < t_2$. Hence h is absolutely continuous. Moreover, we have

$$\limsup_{\tau\to 0^+} \frac{h(t+\tau)-h(t)}{\tau} \le \max\left(\limsup_{\tau\to 0^+} \frac{g(t+\tau)-g(t)}{\tau}, 0\right).$$

Thus, for almost all $t \in [t_0, T]$, we have $h'(t) \le c(t)\, h(t)$. Gronwall inequality now yields $h \equiv 0$, which completes the proof of Lemma 2.5. □

The proof of Theorem 2.4 runs as the proof of Theorem 4.7 in [16] with the difference that we use our Lemma 2.5 instead of Lemma 4.8 in [16]. □

A metric ρ in a space M is an ultrametric if it satisfies strong triangle inequality

$$\rho(x,z) \le \max(\rho(x,y), \rho(y,z)).$$

We say that a subset K of an ultrametric space M is (*)-closed if for every sequence $\{y_n\} \subset K$ and every sequence $\{c_n\}$ ($c_n \ge c_{n+1} \ge 0$) such that $\rho(y_n, y_{n+1}) \le c_n$, there is $\bar{y} \in K$ such that $\rho(\bar{y}, y_n) \le c_n$, for every n.

REMARK 2.6. *If D_1, D_2 are nonempty (*)-closed subsets of an ultrametric space M, then the Hausdorff distance $d_H(D_1, D_2) \le r$ if and only if for every $d_1 \in D_1$ there is $d_2 \in D_2$ such that $\rho(d_1, d_2) \le r$ and for every $d_2 \in D_2$ there is $d_1 \in D_1$ such that $\rho(d_1, d_2) \le r$.*

We say that a set-valued map $A : N \rightsquigarrow M$ is nonexpansive set-valued map from an ultrametric space (N, ρ_N) into another ultrametric space (M, ρ_M) if A satisfies

$$\begin{aligned}
&\forall (n_1, n_2) \in N \times N,\\
&\text{(a)}\ \ \forall m_1 \in A(n_1),\ \exists m_2 \in A(n_2),\ \rho_M(m_1, m_2) \le \rho_N(n_1, n_2),\\
&\text{(b)}\ \ \forall m_2 \in A(n_2),\ \exists m_1 \in A(n_1),\ \rho_M(m_1, m_2) \le \rho_N(n_1, n_2).
\end{aligned}$$

LEMMA 2.7 (nonexpansive selection). *If $A : N \rightsquigarrow M$ is a nonexpansive set-valued map from an ultrametric space (N, ρ_N) into an ultrametric space (M, ρ_M) with nonempty (*)-closed values, then there exists a nonexpansive selection $\alpha : N \mapsto M$ of A.*

The proof of Lemma 2.7 is given in the appendix.

REMARK 2.8. *Given $y_1, y_2 \in M_{t_0}$ we define*

$$\rho(y_1, y_2) = T - \sup\{t \in [t_0, T] \,:\, y_1(s) = y_2(s) \text{ for a.a. } t \in [t_0, t]\}.$$

It is easy to see that (M_{t_0}, ρ) is an ultrametric space. Moreover, a strategy $\alpha : N_{t_0} \to M_{t_0}$ is nothing but a nonexpansive map in the meaning of the ultrametric ρ.

Proof of Theorem 2.2. Fix $t_0 \in [0, T]$, $x_0 \in P(t_0)$, and $\tilde{z}(\cdot) \in N_{t_0}$.

We define a set-valued map $F_{\tilde{z}(\,\cdot\,)}(t,x) = \{f(t,x,y,\tilde{z}(t)) \,:\, y \in Y\}$. By the regularity of f: (2.2), (2.3), (2.4), (2.5), and (2.6), the set-valued map $F_{\tilde{z}(\,\cdot\,)}$ satisfies (2.10), (2.11), (2.12), (2.13). By the separation theorem and (2.2), we have for every $t \in C$ and $x \in P(t)$

$$\forall z \in Z,\ \exists y \in Y,\ (1, f(t,x,y,z)) \in \overline{co}\left(T_{\mathrm{Graph}(P)}(t,x)\right). \tag{2.15}$$

Thus $F_{\tilde{z}(\,\cdot\,)}$ satisfies condition (2.9). Therefore there exists an absolutely continuous solution $\tilde{x} : [t_0, T] \to R^n$ of the differential inclusion $\tilde{x}'(t) \in F_{\tilde{z}(\,\cdot\,)}(t, \tilde{x}(t))$ such that $\tilde{x}(t_0) = x_0$ and $\tilde{x}(t) \in P(t)$ for every $t \in [t_0, T]$. By measurable selection Theorem 8.2.10 in [3], there exists a measurable map $\tilde{y} : [t_0, T] \to Y$ such that $x(t; t_0, x_0, \tilde{y}(\,\cdot\,), \tilde{z}(\,\cdot\,)) = \tilde{x}(t)$ for $t \in [t_0, T]$.

We define a set-valued map $A : N_{t_0} \rightsquigarrow M_{t_0}$ by

$$A(z(\cdot)) = \{y(\cdot) \in M_{t_0} \,:\, x(t; t_0, x_0, y(\cdot), z(\cdot)) \in P(t) \text{ for } t \in [t_0, T]\}.$$

We have shown that the values of the map A are nonempty. Now we verify that the map A satisfies the remaining assumptions of Lemma 2.7.

Let $z_1, z_2 \in N_{t_0}$ and $y_1 \in A(z_1)$. We set $t_1 = T - \rho(z_1, z_2)$ and $x_1 = x(t_1\,;, t_0, x_0, y_1, z_1)$. We have $x_1 \in P(t_1)$. By (2.15) and Theorem 2.4, there exists a solution $\hat{x} : [t_1, T] \to R^n$ of a differential inclusion $\hat{x}'(t) \in F_{z_2}(t, \hat{x}(t))$ such that $\hat{x}(t_1) = x_1$ and $\hat{x}(t) \in P(t)$ for $t \in [t_1, T]$, where $F_{z_2}(t, x) = \{f(t, x, y, z_2(t)) \,:\, y \in Y\}$. By Theorem 8.2.10 in [3], there exists a measurable map $y_3 : [t_1, T] \to Y$ such that $x(t; t_1, x_1, y_3, z_2) = \hat{x}(t)$ for $t \in [t_1, T]$. Setting

$$y_2(t) = \begin{cases} y_1(t) & \text{for} \quad t \in [t_0, t_1), \\ y_3(t) & \text{for} \quad t \in [t_1, T], \end{cases}$$

we get $y_2 \in A(z_2)$ such that $\rho(y_1, y_2) \le \rho(z_1, z_2)$, which means that the map A is nonexpansive.

Now, we show that the set $A(z)$ is $(*)$-closed for every $z \in N_{t_0}$. Let $0 \le \cdots \le c_{k+1} \le c_k \le \cdots \le c_1 \le T - t_0$, $c = \lim_{k\to\infty} c_k$ and $y_k \in A(z)$ satisfy $\rho(y_k, y_{k+1}) \le c_k$. We set $t_k = T - c_k$. Obviously, we have $x(t; t_0, x_0, y_k, z) = x(t; t_0, x_0, y_{k+1}, z)$ for $t \in [t_0, t_k]$. We define a map $y_\infty : [t_0, T - c[\to Y$ by

$$y_\infty(t) = \begin{cases} y_1(t) & \text{for} \quad t \in [t_0, t_1), \\ y_k(t) & \text{for} \quad t \in [t_{k-1}, t_k) \text{ and } k = 2, 3, \ldots . \end{cases}$$

We set $x_\infty = \lim_{t\to(T-c)^-} x(t, t_0, x_0, y_\infty, z)$. It is easy to check that $x_\infty \in P(T - c)$. By (2.15) and Theorem 2.4, there exists a solution $\bar{x} : [T - c, T] \to R^n$ of a differential inclusion $\bar{x}'(t) \in F_z(t, \bar{x}(t))$ such that $\bar{x}(T - c) = x_\infty$ and $\bar{x}(t) \in P(t)$ for $t \in [T - c, T]$. By Theorem 8.2.10 in [3], there exists a measurable map $\bar{y} : [T - c, T] \to Y$ such that $x(t; T - c, x_\infty, \bar{y}, z) = \bar{x}(t)$ for $t \in [T - c, T]$. Setting

$$y(t) = \begin{cases} y_\infty(t) & \text{for} \quad t \in [t_0, T - c), \\ \bar{y}(t) & \text{for} \quad t \in [T - c, T], \end{cases}$$

we get $y \in A(z)$ such that $\rho(y_k, y) \le c_k$, which means that the set $A(z)$ is $(*)$-closed.

Finally, by Lemma 2.7, there exists a nonexpansive selection $\alpha : N_{t_0} \to M_{t_0}$ of A, which is the desired strategy.

For the converse, we set $F_z(t, x) = \{f(t, x, y, z) \,:\, y \in Y\}$. By Lemma 2.6 in [16], there is a full measure set $C \in [0, T]$ such that

$$\forall\, (t_0, x_0, z) \in C \times R^d \times Z,\ \forall\, \varepsilon > 0,\ \exists\, \delta > 0,\ \forall\, x(\cdot) \in \mathrm{Sol}_{F_z}(t_0, x_0),$$

$$\forall\, 0 < |h| < \delta,\ \frac{1}{h}(x(t_0 + h) - x_0) \in F_z(t_0, x_0) + B(0, \varepsilon).$$

Fix $t_0 \in C$, $x_0 \in P(t_0)$, $z_0 \in Z$. Applying (2.7) we obtain an $\alpha_n \in \Gamma_{t_0}$ such that $x_n(t) := x(t, t_0, x_0, \alpha_n(z), z) \in P(t) + B(0, 1/n)$, for $t \in [t_0, T]$, where $z(\cdot)$ is a constant control on $[t_0, T]$ equal to z_0. For fixed $h > 0$ let $x(t_0 + h)$ be a condensing point of the

sequence $(x_n(t_0+h))$. Obviously, we have $x(t_0+h) \in P(t_0+h)$ and for sufficiently small h

$$\frac{x(t_0+h)-x(t_0)}{h} \in F_{z_0}(t_0,x_0) + B(0,\varepsilon).$$

There is a sequence $h_n > 0$ tending to zero such that $v := \lim_{n\to\infty} \frac{x(t_0+h_n)-x(t_0)}{h_n} \in F_{z_0}(t_0,x_0)$. We find $y_0 \in Y$ such that $v = f(t_0,x_0,y_0,z_0)$. We have $(1,v) \in T_{\mathrm{Graph}(P)}(t_0,x_0)$, which yields

$$\langle (n_t,n_x), (1, f(t_0,x_0,y_0,z_0))\rangle \le 0$$

for every $(n_t,n_x) \in N^0_{\mathrm{Graph}(P)}(t_0,x_0)$. ∎

3. Leadership domains.

DEFINITION 3.1 (leadership tube). *The tube $P(\cdot)$ is a leadership tube for f if there exists a set C of full measure in $[0,T]$ such that for every $t \in C$ and $x \in P(t)$*

$$\begin{gathered} \forall (n_t,n_x) \in N^0_{Graph(P)}(t,x),\ \exists z \in Z,\ \forall y \in Y, \\ \langle (n_t,n_x), (1, f(t,x,y,z))\rangle \le 0. \end{gathered} \tag{3.1}$$

THEOREM 3.2. *We assume that a tube $P : [0,T] \rightsquigarrow R^n$ is left absolutely continuous and that the right-hand side $f : [0,T] \times R^n \times Y \times Z \to R^n$ satisfies* (2.2), (2.3), (2.4), and (2.5). *Then $P(\cdot)$ is a leadership tube for f if and only if for any $t_0 \in [0,T]$ and $x_0 \in P(t_0)$*

$$\begin{gathered} \forall \varepsilon > 0,\ \forall \alpha \in \Gamma_{t_0},\ \exists z(\cdot) \in N_{t_0},\ \forall t \in [t_0,T], \\ x(t;t_0,x_0,\alpha(z),z) \in P(t) + B(0,\varepsilon). \end{gathered} \tag{3.2}$$

The proof is based on the following lemma.

LEMMA 3.3. *Let f and $P(\cdot)$ be as in Theorem* 3.2. *The following assertions are equivalent:*

(i) *$P(\cdot)$ is a leadership tube for f.*

(ii) *From any initial condition (t_0,z_0) belonging to Graph(P), and for any measurable map $a : [0,T] \times Z \to Y$, there is at least one solution to the differential inclusion:*

$$\begin{cases} v'(t) \in \overline{co} \bigcup_z f(t,v(t),a(t,z),z) \ \ a.e.\ in\ [t_0,T], \\ v(t_0) = v_0 \end{cases} \tag{3.3}$$

with $v(t) \in P(t)$ for every $t \in [t_0,T]$.

Proof. Assume that $P(\cdot)$ is a leadership tube. There exists a set C of full measure in $[t_0,T]$ such that $\forall t \in C$, $x \in P(t)$, $(n_t,n_x) \in N^0_{\mathrm{Graph}(P)}(t,x)$,

$$\inf_z \sup_y \langle f(t,x,y,z), n_x\rangle + n_t \le 0.$$

For any measurable map $a : [t_0,T] \times Z \to Y$, the set-valued map F_a, defined by

$$F_a(t,x) := \overline{co} \bigcup_z f(t,x,a(t,z),z),$$

is measurable, bounded by $\mu(\cdot)$, has convex compact values, and, for almost every $t \in [t_0,T]$, $x \rightsquigarrow F_a(t,x)$, is $l(t)-$Lipschitz.

Let us now prove that the tube $P(\cdot)$ is viable for F_a. Let $t \in C$, $x \in P(t)$, $(n_t, n_x) \in N^0_{\mathrm{Graph}(P)}(t,x)$. Then

$$\begin{array}{rcl} \inf_{w \in F_a(t,x)} \langle w, n_x \rangle + n_t & = & \inf_z \langle f(t,x,a(t,z),z), n_x \rangle + n_t \\ & \leq & \inf_z \sup_y \langle f(t,x,y,z), n_x \rangle + n_t \ \leq \ 0. \end{array}$$

So $P(\cdot)$ is a viability tube for F_a and Theorem 2.4 states that (ii) holds true.

Conversely, assume that the tube $P(\cdot)$ enjoys property (ii). Fix $n_x \in R^n$, $x_0 \in R^n$ and define

$$Y_{n_x,x_0}(t,z) := \left\{ \bar{y} \in Y \mid \langle f(t,x_0,\bar{y},z), n_x \rangle = \sup_y \langle f(t,x_0,y,z), n_x \rangle \right\}$$

and

$$G_{n_x,x_0}(t,x) := \overline{co} \left\{ f(t,x,y,z) \mid y \in Y_{n_x,x_0}(t,z) \ \& \ z \in Z \right\}.$$

First we prove that the tube $P(\cdot)$ is viable for the set-valued map G_{n_x,x_0} for any n_x and x_0. The set-valued map $Y_{n_x,x_0}(\cdot,\cdot)$ is measurable, so it enjoys a measurable selection $a_{n_x,x_0}(\cdot,\cdot)$. Note that

$$F_{a_{n_x,x_0}}(t,x) = \overline{co} \bigcup_z f(t,x,a_{n_x,x_0}(t,z),z) \ \subset G_{n_x,x_0}(t,x)$$

for almost all $t \in [t_0,T]$ and for all x. Thus, from (ii), there is a solution to the differential inclusion for G_{n_x,x_0} which remains in the tube $P(\cdot)$.

Let us point out that G_{n_x,x_0} is measurable and bounded. Moreover, G_{n_x,x_0} is upper semicontinuous with respect to (n_x,x_0,x) and has convex compact values for almost every $t \in [t_0,T]$. Thus Lemma 2.6 of [16] yields the existence of a set C of full measure in $[t_0,T]$ such that $\forall (\tau, x_\tau, n_x, x_0) \in C \times R^n \times R^n \times R^n$, $\forall \epsilon > 0$, $\exists \delta > 0$ such that, for any solution $x(\cdot)$ to the differential inclusion for G_{n_x,x_0} starting at x_τ at time τ, one has

$$\forall 0 < |h| < \delta, \ \ \frac{1}{h}(x(\tau+h) - x_\tau) \in G_{n_x,x_0}(\tau,x_\tau) + \epsilon B. \tag{3.4}$$

Let $\tau \in C$, $x_\tau \in P(\tau)$, and $(n_t,n_x) \in N^0_{Graph(P(\cdot))}(\tau,x_\tau)$. We have already proved that there is a solution $x(\cdot)$ of the differential inclusion for G_{n_x,x_τ} starting from x_τ at time τ which remains in the tube $P(\cdot)$ on $[\tau,T]$. From (3.4), for any $h \in]0,\delta[$, there is some $w_h \in G_{n_x,x_\tau}(\tau,x_\tau)$ such that

$$\frac{1}{h}(x(\tau+h) - x_\tau) \in w_h + \epsilon B.$$

Since $G_{n_x,x_\tau}(\tau,x_\tau)$ is compact, w_h converges, up to a subsequence, to some $w \in G_{n_x,x_\tau}(\tau,x_\tau)$. Thus, $(1,w)$ belongs to $T_{Graph(P(\cdot))}(\tau,x_\tau)$, and $\langle n_x, w \rangle + n_t \leq 0$. From the very definition of $G_{n_x,x_\tau}(\tau,x_\tau)$, one has

$$\begin{array}{rcl} 0 \ \geq \ \langle n_x, w \rangle + n_t & \geq & \inf_{v \in G_{n_x,x_\tau}(\tau,x_\tau)} \langle v, n_x \rangle + n_t \\ & = & \inf_z \inf_{y \in Y_{n_x,x_\tau}(\tau,z)} \langle f(\tau,x_\tau,y,z), n_x \rangle + n_t \\ & = & \inf_z \sup_y \langle f(\tau,x_\tau,y,z), n_x \rangle + n_t. \end{array}$$

So we have finally proved that, for any $\tau \in C$, for any $x_\tau \in P(\tau)$, and for any $(n_t,n_x) \in N^0_{Graph(P(\cdot))}(\tau,x_\tau)$,

$$\inf_y \sup_z \langle f(\tau,x_\tau,y,z), n_x \rangle + n_t \leq 0,$$

i.e., $P(\cdot)$ is a leadership tube. □

Proof of Theorem 3.2. Assume that $P(\cdot)$ enjoys the property described in Theorem 3.2, and let us prove that $P(\cdot)$ is a leadership tube. Let $a(\cdot,\cdot) : [0,T] \times Z \to Y$, and define the nonanticipative strategy α in the following way:

$$\forall z(\cdot) \in N, \ \ \alpha(z(\cdot))(t) := a(t, z(t)).$$

For any initial position (t_0, z_0) belonging to the graph of $P(\cdot)$, for any $\epsilon > 0$, there is a control $z_\epsilon(\cdot)$ such that the solution $x_\epsilon(\cdot) := x(t_0, x_0, \alpha(z_\epsilon(\cdot)), z_\epsilon(\cdot))$ satisfies

$$\forall t \in [t_0, T], \ \ d_{P(t)}(x_\epsilon(t)) \leq \epsilon.$$

Note that the $x_\epsilon(\cdot)$ are solutions of the differential inclusion (3.3). Since the set of solutions of this differential inclusion is compact, there exists a subsequence of the $x_\epsilon(\cdot)$ convergent to a solution $x(\cdot)$ of (3.3) satisfying $x(t) \in P(t)$ for any $t \in [t_0, T]$. Then Lemma 3.3 states that the tube $P(\cdot)$ is a leadership tube.

Conversely, assume now that $P(\cdot)$ is a leadership tube and fix any $\epsilon > 0$. The idea of the proof consists in constructing the desired control $z(\cdot)$ step by step, on intervals $[n\tau, (n+1)\tau)$, where $\tau > 0$ is fixed and shall be chosen later in function of ϵ.

For that purpose, we need the following estimation.

LEMMA 3.4. *Let f and $P(\cdot)$ be as in Theorem* 3.2, *$t_0 \in [0,T)$, and $x_0 \notin P(t_0)$. Assume that $P(\cdot)$ is a leadership tube. For any nonanticipative strategy α, there is a control $z(\cdot)$ such that, if we set $x(\cdot) := x(t_0, x_0, \alpha(z(\cdot)), z(\cdot))$, then for every $t \in [t_0, T]$,*

$$\begin{array}{rl} d^2_{P(t)}(x(t)) \leq & \left(1 + 2\int_{t_0}^t l(s)ds\right) d^2_{P(t_0)}(x_0) + 4\left(\int_{t_0}^t \mu(s)ds\right)^2 \\ & +2d_{P(t_0)}(x_0) \int_{t_0}^t l(s) \int_{t_0}^s \mu(\sigma)\, d\sigma ds. \end{array}$$

Proof of Lemma 3.4. The proof is based on Lemma 3.3. Let v_0 belong to the proximal projection of x_0 onto $P(t_0)$. Set $\nu := x_0 - v_0$. Consider the following set-valued map:

$$(s,z) \rightsquigarrow \left\{ \bar{y} \in Y \mid \langle f(s, v_0, \bar{y}, z), \nu \rangle = \max_y \langle f(s, v_0, y, z), \nu \rangle \right\}.$$

This set-valued map is measurable and, so, enjoys a measurable selection $a(\cdot,\cdot)$. In the same way, the set-valued map

$$s \rightsquigarrow \left\{ \bar{z} \in Z \mid \max_y \langle f(s, v_0, y, \bar{z}), \nu \rangle = \min_z \max_y \langle f(s, v_0, y, z), \nu \rangle \right\}$$

is measurable and enjoys a measurable selection $z(\cdot) \in N_{t_0}$.

Let us denote now $x(\cdot) := x(t_0, x_0, \alpha(z(\cdot)), z(\cdot))$ and let $v(\cdot)$ be a solution to

$$\begin{cases} v'(t) \in \overline{co} \bigcup_z f(t, v(t), a(t,z), z) \ \text{ f.a.e. } t \in [t_0, T], \\ v(t_0) = v_0, \end{cases}$$

which remains in the tube on $[t_0, T]$ (Lemma 3.3). Then

$$\begin{array}{rcl} d^2_{P(t)}(x(t)) & \leq & \|x(t) - v(t)\|^2 \\ & = & \|(x(t) - x_0) + (\nu) + (v_0 - v(t))\|^2 \\ & = & \|x(t) - x_0\|^2 + \|\nu\|^2 + \|v_0 - v(t)\|^2 + 2\langle x(t) - x_0, \nu \rangle \\ & + & 2\langle x(t) - x_0, v_0 - v(t) \rangle + 2\langle \nu, v_0 - v(t) \rangle. \end{array}$$

Note that $\|x(t)-x_0\|^2$, $\|v_0-v(t)\|^2$, and $\langle x(t)-x_0, v_0-v(t)\rangle$ are bounded by $(\int_{t_0}^t \mu(s)ds)^2$. Note also that $\|\nu\|^2 = d^2_{P(t_0)}(x(t_0))$.

Let us now estimate $\langle x(t)-x_0,\nu\rangle$:

$$\begin{array}{rcl}\langle x(t)-x_0,\nu\rangle &=& \int_{t_0}^t \langle f(s,x(s),\alpha(z(\cdot))(s),z(s)),\nu\rangle ds\\ &\le& \int_{t_0}^t \langle f(s,v_0,\alpha(z(\cdot))(s),z(s)),\nu\rangle ds\\ &+& \|\nu\|\int_{t_0}^t l(s)\|x(s)-v_0\|ds.\end{array}$$

For almost every s,

$$\begin{array}{l}\langle f(s,v_0,\alpha(z(\cdot))(s),z(s)),\nu\rangle\\ \quad\le \langle f(s,v_0,a(s,z(s)),z(s)),\nu\rangle\\ \quad= \min_z\langle f(s,v_0,a(s,z),z),\nu\rangle\\ \quad= \min_{w\in\overline{co}\bigcup_z f(s,v_0,a(s,z),z)}\langle w,\nu\rangle\\ \quad\le \langle v'(s),\nu\rangle + l(s)\|\nu\|\|v(s)-v_0\|\end{array}$$

from the very definition of $a(\cdot,\cdot)$ and of $z(\cdot)$ and because $x\rightsquigarrow \overline{co}\bigcup_z f(s,v,a(s,z),z)$ is $l(s)-$Lipschitz for almost all s. So, we have finally

$$\begin{array}{rl}\langle x(t)-x_0,\nu\rangle &\le \langle v(t)-v(0),\nu\rangle\\ &\quad+\|\nu\|\int_{t_0}^t l(s)(\|x(s)-v_0\|+\|v(s)-v_0\|)ds.\end{array}$$

Since f is bounded by $\mu(\cdot)$,

$$\|x(s)-v_0\|\le\int_{t_0}^s \mu(\sigma)\,d\sigma+\|\nu\| \text{ and } \|v(s)-v_0\|\le\int_{t_0}^s\mu(\sigma)\,d\sigma,$$

so that

$$\langle x(t)-x_0,\nu\rangle+\langle\nu,v_0-v(t)\rangle\le\|\nu\|\int_{t_0}^t l(s)\left(\|\nu\|+2\int_{t_0}^s\mu(\sigma)\,d\sigma\right)ds.$$

In conclusion,

$$\begin{array}{rl}d^2_{P(t)}(x(t))\le & \|\nu\|^2+4\left(\int_{t_0}^t\mu(s)ds\right)^2\\ &+2\|\nu\|\int_{t_0}^t l(s)\left(\|\nu\|+2\int_{t_0}^s\mu(\sigma)\,d\sigma\right)ds.\end{array}$$

□

Construction of $z(\cdot)$. We construct $z(\cdot)$ step by step, on intervals of the form $[n\tau,(n+1)\tau)$, where $\tau>0$ is fixed and shall be chosen below (τ depends mainly on ϵ).

Assume that we have already defined $z(\cdot)$ on $[0,n\tau]$. Then set $x_n := x(n\tau;t_0,x_0,\alpha(z(\cdot)),z(\cdot))$ (Note that x_n is well defined because α is nonanticipative.)

• If x_n belongs to $P(n\tau)$, then choose any $z\in Z$ and set $z(\cdot):=z$ on $[n\tau,(n+1)\tau)$.

• Otherwise, let $z_1(\cdot)$ be the control defined in Lemma 3.4 for $(t_0,x_0):=(n\tau,x_n)$. Then we set $z(\cdot):=z_1(\cdot)$ on $[n\tau,(n+1)\tau)$.

Note that the distance between $x(t):=x(t;t_0,x_0,\alpha(z(\cdot)),z(\cdot))$ and $P(t)$ ($t\in[t_0,T]$) is maximal if $x_n\notin P(n\tau)$ for any $n>0$. In that case, this distance satisfies $\forall t\in[n\tau,(n+1)\tau)$,

$$\begin{array}{rl}d^2_{P(t)}(x(t))\le & \left(1+2\int_{n\tau}^t l(s)ds\right)d^2_{P(n\tau)}(x_n)+4\left(\int_{n\tau}^t\mu(s)ds\right)^2\\ &+2d_{P(n\tau)}(x_n)\int_{n\tau}^t l(s)\int_{n\tau}^s\mu(\sigma)\,d\sigma\,ds\end{array}$$

from Lemma 3.4. In particular,

$$\forall t \in [n\tau, (n+1)\tau),\ d^2_{P(t)}(x(t)) \le d_{n+1}(\tau),$$

where $d_n(\tau)$ is the sequence defined by $d_0(\tau) = 0$ and

$$d_{n+1}(\tau) = (1 + \alpha_n(\tau))\, d_n(\tau) + \beta_n(\tau),$$

where $\alpha_n(\tau) := 2\int_{n\tau}^{(n+1)\tau} l(s)ds$,

$$\beta := \max\left\{4; 2 \sup_{z(\cdot)\in N_{t_0}} \sup_{t\in[t_0,T]} d_{P(t)}(x(t; t_0, x_0, \alpha(z(\cdot)), z(\cdot)))\right\}$$

(note that $\beta < +\infty$ because f is bounded and $P(\cdot)$ is absolutely continuous), and

$$\beta_n(\tau) := \beta\left[\left(\int_{n\tau}^{(n+1)\tau} \mu(s)ds\right)^2 + \int_{n\tau}^{(n+1)\tau} l(s)\int_{n\tau}^{s} \mu(\sigma)\, d\sigma ds\right].$$

To prove that the sequence constructed step by step satisfies the conclusion of Theorem 3.2, it is sufficient to apply the following lemma.

LEMMA 3.5. *Let d_n be the sequence defined previously. For any $\epsilon > 0$, there is $\tau_0 > 0$ such that, if $0 < \tau < \tau_0$, then*

$$\forall n \le \frac{T+\tau}{\tau},\ \ d_n(\tau) \le \epsilon.$$

Proof. Fix $\epsilon > 0$. To simplify the notations, we shall write d_i instead of $d_i(\tau)$, α_i instead of $\alpha_i(\tau)$, and so on.

It is easy to prove by induction that

$$d_{n+1} = \sum_{i=0}^{n}\left(\prod_{j=i}^{n-1}(1+\alpha_j)\right)\beta_i$$

(where, by convention, $\prod_{j=n}^{n-1}(1+\alpha_j) = 1$). Note that

$$\left(\prod_{j=0}^{n}(1+\alpha_j)\right) = \exp\left[\sum_{j=0}^{n}\ln(1+\alpha_j)\right] \le \exp\left[\sum_{j=0}^{n}\alpha_j\right] \le \exp\left[2\|l(\cdot)\|_1\right]$$

from the very definition of α_i. So,

$$d_{n+1} \le \exp\left[2\|l(\cdot)\|_1\right]\sum_{i=0}^{n}\beta_i.$$

Set $\epsilon_0 := \epsilon/(\beta e^{2\|l\|_1}\|l+\mu\|_1)$. Now choose τ small enough (say, $\tau < \tau_0$) in such a way that $\int_{i\tau}^{(i+1)\tau} l(s)ds \le \epsilon_0$ and $\int_{i\tau}^{(i+1)\tau}\mu(s)ds \le \epsilon_0$ for any i such that $i\tau \le T$.

Then, for any $n \le \frac{T}{\tau}$,

$$\sum_{i=0}^{n}\beta_i \le \beta\epsilon_0\sum_{i=0}^{n}\int_{i\tau}^{(i+1)\tau}(\mu(s)+l(s))ds \le \beta\epsilon_0\|\mu+l\|_1$$

so that

$$d_{n+1} \le \beta\epsilon_0\|\mu + l\|_1 e^{2\|l\|_1} \le \epsilon. \qquad \square$$

REMARK 3.6. *If $Z = \{z_0\}$, then the differential game reduces to the control system with dynamics given by $\hat{f}(t,x,y) = f(t,x,y,z_0)$. Assume, moreover, that $\{f(t,x,y,z_0) : y \in Y\}$ is convex for every t and x. Then leadership tube condition* (3.1) *implies that*

$$\forall y \in Y,\ (1, f(t,x,y,z_0)) \in \overline{co}\left(T_{Graph(P)}(t,x)\right)$$

and discriminating tube condition (2.1) *implies that*

$$\exists y \in Y,\ (1, f(t,x,y,z_0)) \in \overline{co}\left(T_{Graph(P)}(t,x)\right).$$

4. Value function and Hamilton–Jacobi–Isaacs equations. We mostly adopt here the notation of Evans–Souganidis [11].

Dynamic of a differential game is given by $f : [0,T] \times R^n \times Y \times Z \to R^n$. We assume that f satisfies (2.2), (2.3), (2.4), and (2.5). We shall consider a terminal time payoff functional

$$Q(y,z) = Q_{t_0x_0}(y,z) = g(x(T,\, t_0, x_0, y, z)),$$

where $g : R^n \to R$ is a continuous function, $y \in M_{t_0}$, $z \in N_{t_0}$.

We define

$$U(t_0,x_0) = \sup_{\alpha\in\Gamma_{t_0}} \inf_{z\in N_{t_0}} Q_{t_0x_0}(\alpha(z),z),$$
$$H(t,x,p) = \min_{z\in Z} \max_{y\in Y} \langle f(t,x,y,z), p\rangle.$$

We call $U : [0,T] \times R^n \to R$ the value function of the differential game and $H : [0,T] \times R^n \times R^n \to R$ the Hamiltonian. In [11], U and H are called the upper value function and the upper Hamiltonian. Besides these, a lower value function and a lower Hamiltonian are considered there. We formulate all results only for the upper value function. Following [11], we can reformulate them for the lower value function. If the value function U is differentiable in its domain, then it satisfies the Isaacs (Hamilton–Jacobi–Isaacs) equation:

$$U_t + H(t,x,U_x) = 0.$$

Evans and Souganidis in [11] proved that if g is Lipschitz continuous and f is continuous and Lipschitz continuous with respect to x, then U is a viscosity solution of Isaacs equation, i.e., for every $t \in (0,T)$ and $x \in R^n$

$$(h_1) \quad \forall(-p_t,-p_x,1) \in N^0_{\mathrm{Hyp}(U)}(t,x,U(t,x)),\ \ p_t + H(t,x,p_x) \ge 0$$

and

$$(h_2) \quad \forall(p_t,p_x,-1) \in N^0_{\mathrm{Epi}(U)}(t,x,U(t,x)),\ \ p_t + H(t,x,p_x) \le 0,$$

where $\mathrm{Hyp}(U) = \{(t,x,u) \in [0,T] \times R^n \times R : u \le U(t,x)\}$ and $\mathrm{Epi}(U) = \{(t,x,u) \in [0,T] \times R^n \times R : u \ge U(t,x)\}$.

We start with some properties of value function.

THEOREM 4.1 (Theorem 3.1 in [11]). *For each $0 \le t < t+h \le T$ and $x \in R^n$,*

$$U(t,x) = \sup_{\alpha\in\Gamma_t} \inf_{z\in N_t} U(t+h, x(t+h,\, t, x, \alpha(z), z)).$$

This is the dynamic programming optimality condition.

Next we examine regularity of the value function. We recall that the modulus of continuity $m_{f,A}(\delta)$ of a function $f : X \to Y$ (X, Y are metric spaces) on a subset $A \subset X$ is given by

$$m_{f,A}(\delta) = \sup\{d(f(x_1), f(x_2)) \,:\, x_1, x_2 \in A,\ d(x_1,x_2) \le \delta\}$$

for $\delta > 0$. It is easy to check that f is uniformly continuous on A if and only if $\lim_{\delta\to 0^+} m_{f,A}(\delta) = 0$. Moreover, $m_{f,A}(\cdot)$ is nondecreasing and if $A \subset B \subset X$, then $m_{f,A}(\delta) \le m_{f,B}(\delta)$.

PROPOSITION 4.2. *If f satisfies* (2.2), (2.3), (2.4), (2.5) *and $g : R^n \to R$ is continuous, then we have*

$$m_{U(t_0,\,\cdot\,),\, B(0,R)}(\delta) \le m_{g,\, B(0,R+\int_{t_0}^T \mu(s)\,ds)}\left(\delta \exp\left(\int_{t_0}^T l(s)\,ds\right)\right).$$

Proof. Fix $t_0 \in [0,T]$, α and z. By (2.3) and the Gronwall inequality, we have

$$\|x(T,\, t_0, x_1, \alpha(z), z) - x(T,\, t_0, x_2, \alpha(z), z)\| \le \exp\left(\int_{t_0}^T l(s)\,ds\right)\|x_1 - x_2\|.$$

By (2.5), we obtain

$$\|x(T,\, t_0, x_0, \alpha(z), z) - x_0\| \le \int_{t_0}^T \mu(s)\,ds.$$

If we use the above estimations, the proof is straightforward. □

COROLLARY 4.3. *The function $U(t_0,\,\cdot\,)$ is continuous for every $t_0 \in [0,T]$.*

We define the tubes $E,\ H : [0,T] \rightsquigarrow R^n$ by $E(t) = \{(x,u) \,:\, u \ge U(t,x)\}$, $H(t) = \{(x,u) \,:\, u \le U(t,x)\}$. We call E the epitube and H the hypotube generated by value function U. Obviously $\mathrm{Graph}(H) = \mathrm{Hyp}(U)$ and $\mathrm{Graph}(E) = \mathrm{Epi}(U)$.

PROPOSITION 4.4. *If f satisfies* (2.5)*, then the epitube E and the hypotube H generated by the value function U are left absolutely continuous, namely,*

$$E(t_1) \subset E(t_2) + \left(\int_{t_1}^{t_2} \mu(s)\right) B\,ds,$$

$$H(t_1) \subset H(t_2) + \left(\int_{t_1}^{t_2} \mu(s)\right) B\,ds$$

for $t_1 < t_2$.

Proof. Fix $x \in R^n$, $0 \le t_1 < t_2 \le T$. By Theorem 4.1,

$$U(t_1,x) = \sup_{\alpha\in\Gamma_{t_1}} \inf_{z\in N_{t_1}} U(t_2, x(t_2,\, t_1, x, \alpha(z), z)).$$

Take $\varepsilon > 0$. There is $\alpha_0 \in \Gamma_{t_1}$ such that

$$U(t_1, x) - \varepsilon < \inf_{z \in N_{t_1}} U(t_2, x(t_2,\, t_1, x, \alpha_0(z), z)) \le U(t_1, x).$$

Next, there is $z_0 \in N_{t_1}$ such that

$$U(t_1, x) - \varepsilon < U(t_2, x(t_2,\, t_1, x, \alpha_0(z_0), z_0)) < U(t_1, x) + \varepsilon. \tag{4.1}$$

We set $x(\cdot) = x(\cdot\,,\, t_1, x, \alpha_0(z_0), z_0)$. If $u \ge U(t_1, x)$, then $U(t_2, x(t_2)) < u + \varepsilon$. Thus $(x(t_2), u + \varepsilon) \in E(t_2)$. Therefore,

$$\operatorname{dist}((x, u), E(t_2)) \le (\|x(t_2) - x\|^2 + \varepsilon^2)^{1/2}.$$

Hence,

$$E(t_1) \subset E(t_2) + B\bigg(\int_{t_1}^{t_2} \mu(s)\, ds\bigg).$$

Now, let $(x, u) \in H(t_1)$. By (4.1), $u - \varepsilon < U(t_2, x(t_2))$. Thus $(x(t_2), u - \varepsilon) \in H(t_2)$. Therefore,

$$\operatorname{dist}((x, u), H(t_2)) \le (\|x(t_2) - x\|^2 + \varepsilon^2)^{1/2},$$

which completes the proof. □

PROPOSITION 4.5. *If $U : [0, T] \times R^n \to R$ is a value function, then for each $t_0 \in [0, T]$ and $x_0 \in R^n$*

$$\begin{gathered} \forall \varepsilon > 0,\ \exists \alpha \in \Gamma_{t_0},\ \forall z \in N_{t_0},\ \forall t \in [t_0, T], \\ U(t_0, x_0) \le U(t, x(t; t_0, x_0, \alpha(z), z)) + \varepsilon, \end{gathered} \tag{4.2}$$

$$\begin{gathered} \forall \varepsilon > 0,\ \forall \alpha \in \Gamma_{t_0},\ \exists z \in N_{t_0},\ \forall t \in [t_0, T], \\ U(t_0, x_0) \ge U(t, x(t; t_0, x_0, \alpha(z), z)) - \varepsilon. \end{gathered} \tag{4.3}$$

Proof. Fix $t_0 \in [0, T]$, $x_0 \in R^n$, and $\varepsilon > 0$.

First we prove that if U is the value function, then (4.2) holds true. By the definition of the value function, there exists an $\alpha_\varepsilon \in \Gamma_{t_0}$ such that

$$U(t_0, x_0) \le \inf_{z \in N_{t_0}} g(x(T; t_0, x_0, \alpha(z), z)) + \varepsilon/2.$$

We show that (4.2) holds true for α_ε. To the contrary, assume that there are $z_0 \in N_{t_0}$ and $t_1 \in [0, T]$ such that

$$U(t_0, x_0) > U(t_1, x_1) + \varepsilon,$$

where $x_1 = x(t_1; t_0, x_0, \alpha_\varepsilon(z_0), z_0)$. Given $z_1 \in N_{t_1}$, we set

$$(z_0, z_1)(s) = \begin{cases} z_0(s) & \text{for} \quad s \in [t_0, t_1), \\ z_1(s) & \text{for} \quad s \in [t_1, T]. \end{cases}$$

Let $\alpha_1 \in \Gamma_{t_1}$ be given by

$$\alpha_1(z_1)(s) = \alpha_\varepsilon(z_0, z_1)(s) \tag{4.4}$$

for $s \in [t_1, T]$. Obviously, $U(t_1, x_1) \geq \inf_{z_1 \in N_{t_1}} g(x(T; t_1, x_1, \alpha_1(z_1), z_1)$ and hence there is $z_1 \in N_{t_1}$ such that

$$\inf_{z \in N_{t_1}} g(x(T; t_1, x_1, \alpha_1(z), z)) > g(x(T; t_1, x_1, \alpha_1(z_1), z_1)) - \varepsilon/2.$$

Setting $\tilde{z} = (z_0, z_1)$ we have

$$x(T; t_0, x_0, \alpha_\varepsilon(\tilde{z}), \tilde{z}) = x(T; t_1, x_1, \alpha_1(z_1), z_1).$$

Thus

$$U(t_0, x_0) > U(t_1, x_1) + \varepsilon > g(x(T; t_0, x_0, \alpha_\varepsilon(\tilde{z}), \tilde{z})) - \frac{\varepsilon}{2} + \varepsilon,$$

which is the desired contradiction.

Fix $\alpha \in \Gamma_{t_0}$. We divide the proof of (4.3) into two steps.

Step 1. We fix a division $t_0 < t_1 < \cdots < t_k = T$ of the interval $[t_0, T]$. By the dynamic programming optimality condition (Theorem 4.1), there is $z_0 \in N_{t_0}$ such that

$$U(t_0, x_0) > U(t_1, x_1) - \frac{\varepsilon}{2k},$$

where $x_1 = x(t_1; t_0, x_0, \alpha(z_0), z_0)$. Taking $\alpha_0 = \alpha$ in (4.4) we obtain an $\alpha_1 \in \Gamma_{t_1}$. By the dynamic programming optimality condition again, we obtain $z_1 \in N_{t_1}$ such that

$$U(t_1, x_1) > U(t_2, x_2) - \frac{\varepsilon}{2k},$$

where $x_1 = x(t_2; t_1, x_1, \alpha_1(z_1), z_1)$.

We proceed by induction getting a sequence $z_2 \in N_{t_2}, \ldots, z_{k-1} \in N_{t_{k-1}}$. Setting $\tilde{z}(s) = z_i(s)$, for $s \in [t_{i-1}, t_i)$, we obtain

$$U(t_0, x_0) \geq U(t_i, x(t_i; t_0, x_0, \alpha(\tilde{z}), \tilde{z})) - \frac{i\varepsilon}{2k}.$$

Step 2. We set $R = \|x_0\| + 1$ and find $\delta > 0$ such that

$$m_{g, B(R + \int_{t_0}^T \mu(s)ds)} \left(\delta \int_{t_0}^T l(s)ds \right) < \frac{\varepsilon}{2}.$$

Next we choose a division $t_0 < t_1 < \cdots < t_k = T$ of the interval $[t_0, T]$ such that $\int_{t_{i-1}}^{t_i} \mu(s)ds < \frac{\delta}{2}$, for $i = 1, 2, \ldots, k$. By Step 1, we find $\tilde{z} \in N_{t_0}$ such that $U(t_0, x_0) > U(t_i, \tilde{x}(t_i)) - \frac{\varepsilon}{2}$, where $\tilde{x}(t) = x(t; t_0, x_0, \alpha(\tilde{z}), \tilde{z})$. Fix $t \in [t_{i-1}, t_i]$. By the dynamic programming optimality condition $U(t, \tilde{x}(t)) \in [\inf\{U(t_i, y) \ : \ \|y - \tilde{x}(t)\| \leq \int_t^{t_i} \mu(s)ds\}, \sup\{U(t_i, y) \ : \ \|y - \tilde{x}(t)\| \leq \int_t^{t_i} \mu(s)ds\}] (:= J)$. Since $\|\tilde{x}(t_i - \tilde{x}(t)\| \leq \int_t^{t_i} \mu(s)ds$ then also $U(t_i, \tilde{x}(t_i)) \in J$. Thus $\|U(t, \tilde{x}(t)) - U(t_i, \tilde{x}(t_i)\| \leq \sup\{\|U(t_1, y_1) - U(t_1, y_2)\| \ : \ y_1, y_2 \in B(\tilde{x}(t_0), \int_t^{t_i} \mu(s)ds)\} \leq \varepsilon/2$, which completes the proof. □

THEOREM 4.6. *Suppose that $g : R^n \to R$ is continuous and $f : [0, T] \times R^n \times Y \times Z \to R^n$ satisfies* (2.2), (2.3), (2.4), (2.5), *and* (2.6). *Let the Hamiltonian $H : [0, T] \times R^n \times R^n \to R$ and the value function $U : [0, T] \times R^n \to R$ be generated by f, g. Then a function $W : [0, T] \times R^n \to R$ is the value function, i.e., $W = U$, if and only if W satisfies the following conditions:*

(a) $W(T, \cdot) = g(\cdot)$;

(b) $W(t,\cdot)$ *is continuous function, for every* $t \in [0,T]$*;*

(c) *the epitube* E_W *and the hypotube* H_W *are left absolutely continuous, where* $E_W(t) = \{(x,w) \in R^n \times R : w \geq W(t,x)\}$ *and* $H_W(t) = \{(x,w) \in R^n \times R : w \leq W(t,x)\}$*;*

(d) *there exists a full measure set* $C \subset [0,T]$ *such that for every* $t \in C$ *and* $x \in R^n$

$$\forall (n_t, n_x, n_u) \in N^0_{Graph(H_W)}(t,x,W(t,x)), \quad -n_t + H(t,x,-n_x) \geq 0 \tag{4.5}$$

$$\forall (n_t, n_x, n_u) \in N^0_{Graph(E_W)}(t,x,W(t,x)), \quad n_t + H(t,x,n_x) \leq 0. \tag{4.6}$$

REMARK 4.7. *Note that, if* $W = U$ *is smooth, then* (4.5) *and* (4.6) *mean nothing but that* W *satisfies the Hamilton–Jacobi–Isaacs equation:*

$$\frac{\partial W}{\partial t}(t,x) + H\left(t,x,\frac{\partial W}{\partial x}(t,x)\right) = 0.$$

Proof. Suppose that $W = U$. Corollary 4.3 and Proposition 4.4 yield (b) and (c). Let $\tilde{f}(t,x,u,y,z) = (f(t,x,y,z),0)$, $u \in R$. The function $(x(t),u(t)) = (x(t;t_0,x_0,\alpha(z),z),U(t_0,x_0))$ is the solution of the Cauchy problem

$$\begin{cases} (x'(t),u'(t)) = \tilde{f}(t,x(t),u(t),\alpha(z),z), \\ (x(t_0),u(t_0)) = (x_0,U(t_0,x_0)). \end{cases}$$

From (4.2) it follows that (2.8) holds true for $P = H_W$ and $f = \tilde{f}$. By Theorem 2.2, the hypotube H_W is a discriminating tube for $\tilde{f}$. From this we conclude (4.5). From (4.3) it follows that (3.2) holds true for $P = E_W$ and $f = \tilde{f}$. By Theorem 3.2, the epitube E_W is a leadership tube for $\tilde{f}$. From this we conclude (4.6).

Now, suppose that a function W satisfies (a), (b), (c), and (d). From (4.5) it follows that the hypotube H_W is a discriminating tube for $\tilde{f}$. Fix t_0 and x_0. By Theorem 2.2, there is $\alpha \in \Gamma_{t_0}$ such that for every $z \in N_{t_0}$

$$(x(T;t_0,x_0,\alpha(z),z),W(t_0,x_0)) \in H_W(T).$$

Hence

$$\forall \alpha, \ \exists z, \quad W(t_0,x_0) \leq W(T,x(T;t_0,x_0,\alpha(z),z)).$$

Thus

$$W(t_0,x_0) \leq \sup_\alpha \inf_z g(x(T;t_0,x_0,\alpha(z),z)).$$

From (4.6) it follows that the epitube E_W is a leadership tube for $\tilde{f}$. Fix t_0, x_0. By Theorem 3.2, for every $\varepsilon > 0$ and every $\alpha \in \Gamma_{t_0}$, there is $z \in N_{t_0}$

$$(x(T;t_0,x_0,\alpha(z),z),W(t_0,x_0)) \in E_W(T) + B(0,\varepsilon).$$

Since g is uniformly continuous on $B(x_0, \int_{t_0}^T \mu(s)ds)$, we have

$$W(t_0,x_0) \geq \sup_\alpha \inf_z g(x(T;t_0,x_0,\alpha(z),z)),$$

which completes the proof. □

5. Appendix. The aim of the appendix is to prove Lemma 2.7. Lemma 2.7 has been announced in [8]. We proceed the proof with some elementary properties of ultrametric spaces.

PROPOSITION 5.1. *If $y_1, y_2, y_3 \in M$ and $\rho(y_1, y_2) < \rho(y_2, y_3)$, then $\rho(y_1, y_3) = \rho(y_2, y_3)$.*

Proof. We have $\rho(y_2, y_3) \leq \max(\rho(y_1, y_2), \rho(y_1, y_3))$. Thus $\rho(y_2, y_3) \leq \rho(y_1, y_3)$. Moreover, $\rho(y_1, y_3) \leq \max(\rho(y_1, y_3), \rho(y_2, y_3)) = \rho(y_2, y_3)$. □

Let K be a nonempty subset of M. We denote by Q_K the family of nonempty subsets of K of the form

$$\{y \in K \,:\, \rho(y, y_0) \leq c\},$$

where $y_0 \in K$ and $c \geq 0$.

PROPOSITION 5.2. *If $D \in Q_K$ and $\bar{y} \in D$, then $D = \{y \in K \,:\, \rho(y, \bar{y}) \leq diam\, D\}$.*

Proof. Fix $y_0 \in K$, $c \geq 0$ and define $D = \{y \in K \,:\, \rho(y, y_0) \leq c\}$. Let $\bar{y} \in D$. Obviously, we have $D \subset \{y \in K \,:\, \rho(y, \bar{y}) \leq \mathrm{diam} D\}$. If $y \in K$ and $\rho(y, \bar{y}) \leq \mathrm{diam} D$, then $\rho(y, y_0) \leq \max(\rho(y, \bar{y}), \rho(\bar{y}, y_0)) \leq \max(\mathrm{diam} D, c) = c$. □

PROPOSITION 5.3. *If $D \in Q_K$, $y_1 \in M$, $c \geq 0$ are such that $D_1 = \{y \in D \,:\, \rho(y, y_1) \leq c\}$ is a nonempty set, then $D_1 \in Q_K$.*

Proof. Fix $\bar{y} \in D_1$. By Proposition 5.2, we have $D = \{y \in K \,:\, \rho(y, \bar{y}) \leq \mathrm{diam}\, D\}$.

Case 1. If $\mathrm{diam}\, D \leq c$, then $D_1 = D$.

Indeed, for any $y \in D$, we have $\rho(y, y_1) \leq \max(\rho(y, \bar{y}), \rho(\bar{y}, y_1)) \leq \max(\mathrm{diam}\, D, c) = c$.

Case 2. If $\mathrm{diam}\, D > c$, then $D_1 = \{y \in K \,:\, \rho(y, \bar{y}) \leq c\}$.

If $y \in D_1$, then $\rho(y, \bar{y}) \leq \max(\rho(y, y_1), \rho(y_1, \bar{y})) \leq c$. Thus $D_1 \subset \{y \in K \,:\, \rho(y, \bar{y}) \leq c\}$. If $y \in K$ and $\rho(y, \bar{y}) \leq c$, then $y \in D$ and $\rho(y, y_1) \leq \max(\rho(y, \bar{y}), \rho(\bar{y}, y_1)) \leq c$. □

PROPOSITION 5.4. *If $D_1, D_2 \in Q_K$, $D_1 \subset D_2$, and $D_1 \neq D_2$, then $diam D_1 < diam D_2$.*

Proof. Suppose that $\mathrm{diam}\, D_1 = \mathrm{diam}\, D_2 = d$ and $D_1 \subset D_2$. If $y_1 \in D_1$ and $y_2 \in D_2$, then $D_1 = \{y \in K \,:\, \rho(y, y_1) \leq d\}$ and $D_2 = \{y \in K \,:\, \rho(y, y_2) \leq d\}$. We have $\rho(y_1, y_2) \leq d$. If $y \in D_2$, then $\rho(y, y_1) \leq \max(\rho(y, y_2), \rho(y_2, y_1)) \leq d$, which involves that $y \in D_1$. □

PROPOSITION 5.5. *Suppose that K is a nonempty $(^*)$-closed subset of M, and a family $\{D_\omega \in Q_K \,:\, \omega \in \Omega\}$ satisfies the following condition:*

$$\forall \omega_1, \omega_2 \in \Omega,\ D_{\omega_1} \subset D\omega_2 \text{ or } D_{\omega_2} \subset D\omega_1.$$

Then

(1) $\forall \omega_1, \omega_2 \in \Omega,\ (diam\, D_{\omega_1} \leq diam\, D_{\omega_2} \Longrightarrow D_{\omega_1} \subset D_{\omega_2})$.

(2) *If a sequence $\{\omega_n\} \subset \Omega$ satisfies the conditions*
- *$diam\, D_{\omega_{n+1}} \leq diam\, D_{\omega_n}$ $(:= d_n)$,*
- *$\lim_{n\to\infty} d_n = \inf_{\omega\in\Omega} diam\, D_\omega$ $(:= d)$,*

then for every sequence $y_n \in D_{\omega_n}$ there is $\bar{y} \in K$ such that $\rho(\bar{y}, y_n) \leq d_n$ and

$$\bigcap_{\omega\in\Omega} D_\omega = \{y \in K \,:\, \rho(y, \bar{y}) \leq d\}. \tag{5.1}$$

(3) *The set $D = \bigcap_{\omega\in\Omega} D_\omega$ belongs to Q_K.*

Proof. Assertion (1) is an immediate consequence of Proposition 5.4.
By Proposition 5.2,

$$D_{\omega_n} = \{y \in K \,:\, \rho(y, y_n) \le d_n\}.$$

According to assertion (1), $D_{\omega_{n+1}} \subset D_{\omega_n}$ for every n. Therefore, $\rho(y_{n+1}, y_n) \le d_n$. Since K is (*)-closed, there is $\bar{y} \in K$ such that $\rho(\bar{y}, y_n) \le d_n$.

If $y \in D_{\omega_n}$, then $\rho(y, \bar{y}) \le \max(\rho(y, y_n), \rho(y_n, \bar{y})) \le d_n$.

Let us choose $y \in K$ such that $\rho(y, \bar{y}) \le d$ and pick some $\omega \in \Omega$. There is ω_n such that $d_n \le \operatorname{diam} D_\omega$. By the assertion (a), we have $D_{\omega_n} \subset D_\omega$. Moreover, $\rho(y, y_n) \le \max(\rho(y, \bar{y}), \rho(\bar{y}, y_n)) \le \max(d, d_n) = d_n$. Hence $y \in D_{\omega_n}$.

By (5.1), we obtain statement (3). □

Proof of Lemma 2.7. First we define a partial order $(\mathcal{P}, \le)$. The family $\mathcal{P}$ consists of all nonempty valued nonexpansive maps $C : N \rightsquigarrow M$ such that $C(z) \in Q_{A(z)}$ for every $z \in N$. Since $A \in \mathcal{P}$, then the family $\mathcal{P}$ is nonempty. We say that $C_1 \le C_2$ if $C_1(z) \subset C_2(z)$ for every $z \in N$.

Step 1. Let $\{C_\omega\}_{\omega\in\Omega} \subset \mathcal{P}$ be a chain. Define a set-valued map $C : N \rightsquigarrow M$ by $C(z) = \bigcap_{\omega\in\Omega} C_\omega(z)$. By Proposition 5.5(3), we have $C(z) \in Q_{A(z)}$ for every $z \in N$. Now, we show that C is a nonexpansive map. Let us take $z_1, z_2 \in N$ and $\bar{y_1} \in C(z_1)$.

Case 1. $\rho(z_1, z_2) \ge \inf_{\omega\in\Omega} \operatorname{diam} C_\omega(z_2)$. We choose a sequence $\{\omega_n\} \subset \Omega$ such that $d_{n+1} \le d_n \,(:= \operatorname{diam} C_{\omega_n}(z_2))$ and $\lim_{n\to\infty} d_n = \inf_{\omega\in\Omega} \operatorname{diam} C_\omega(z_2)$. Since C_{ω_n} is a nonexpansive map, then there is $y_n \in C_{\omega_n}(z_2)$ such that $\rho(\bar{y_1}, y_n) \le \rho(z_1, z_2)$. By Proposition 5.5(2), there is $\bar{y} \in C(z_2)$ such that $\rho(y_n, \bar{y}) \le d_n$. Therefore $\rho(\bar{y_1}, \bar{y}) \le \max(\rho(\bar{y_1}, y_n), \rho(y_n, \bar{y})) \le \max(\rho(z_1, z_2), d_n)$.

Case 2. $\rho(z_1, z_2) < \inf_{\omega\in\Omega} \operatorname{diam} C_\omega(z_2)$. Let us fix $\omega_0 \in \Omega$ and choose $y_0 \in C_{\omega_0}(z_2)$ such that $\rho(\bar{y_1}, y_0) \le \rho(z_1, z_2)$. We claim that $y_0 \in C(z_2)$. We pick some $\omega \in \Omega$ and choose $y_\omega \in C_\omega(z_2)$ such that $\rho(\bar{y_1}, y_\omega) \le \rho(z_1, z_2)$. Thus $\rho(y_0, y_\omega) \le \max(\rho(y_0, y_1), \rho(y_1, y_\omega)) \le \rho(z_1, z_2)$. By Proposition 5.2, we have $C_\omega(z_2) = \{y \in A(z_2) \,:\, \rho(y, y_\omega) \le \operatorname{diam} C_\omega(z_2)\}$. Therefore $y_0 \in C_\omega(z_2)$.

Step 2. Suppose that $C \in \mathcal{P}$ and there is $z_0 \in N$ such that $\operatorname{diam} C(z_0) > 0$, i.e., C is not a single-valued map. We define a map $\tilde{C} : N \rightsquigarrow M$ by

$$\tilde{C}(z) = \begin{cases} C(z) & \text{if} \quad \rho(z, z_0) \ge d, \\ \{y \in C(z) \,:\, \rho(y, y_0) \le \rho(z, z_0)\} & \text{if} \quad \rho(z, z_0) < d, \end{cases}$$

where $d = \operatorname{diam} C(z_0)$ and y_0 is a fixed element of $C(z_0)$. Obviously, $\tilde{C}(z_0) = \{y_0\} \ne C(z_0)$. Since C is a nonexpansive map, then $\tilde{C}(z) \ne \emptyset$ for every $z \in N$. By Proposition 5.3, we have $\tilde{C}(z) \in Q_{A(z)}$ for every $z \in N$. Now, we show that $\tilde{C}$ is a nonexpansive map. Let us take $z_1, z_2 \in N$ and $y_1 \in \tilde{C}(z_1)$.

Case 1. $\rho(z_1, z_0) \ge d$ and $\rho(z_2, z_0) < d$. Let us take an arbitrary $y_2 \in \tilde{C}(z_2)$. By Proposition 5.1, we have $\rho(z_1, z_2) = \rho(z_1, z_0)$. Since C is a nonexpansive map, then there is $\bar{y_0} \in C(z_0)$ such that $\rho(y_1, \bar{y_0}) \le \rho(z_1, z_0)$. Therefore $\rho(y_1, y_2) \le \max(\rho(y_1, \bar{y_0}), \rho(\bar{y_0}, y_0), \rho(y_0, y_2)) \le \max(\rho(z_0, z_2), d) \le \rho(z_1, z_2)$.

Case 2. $\rho(z_1, z_0) < d$ and $\rho(z_2, z_0) < d$.

- $\rho(z_1, z_0) \le \rho(z_2, z_0)$. By Proposition 5.1, we have $\rho(z_1, z_2) = \rho(z_2, z_0)$. For any $y_2 \in \tilde{C}(z_2)$ we have $\rho(y_1, y_2) \le \max(\rho(y_1, y_0), \rho(y_0, y_2)) \le \max(\rho(z_1, z_0), \rho(z_0, z_2)) = \rho(z_1, z_2)$.
- $\rho(z_1, z_0) > \rho(z_2, z_0)$. By Proposition 5.1, we have $\rho(z_1, z_2) = \rho(z_1, z_0)$. Let y_2 be an arbitrary element of $\tilde{C}(z_2)$. Therefore, $\rho(y_1, y_2) \le \max(\rho(y_1, y_0), \rho(y_0, y_2)) \le \max(\rho(z_1, z_0), \rho(z_0, z_2)) = \rho(z_1, z_2)$.

- $\rho(z_1, z_0) = \rho(z_2, z_0)$. Since C is a nonexpansive map, then there is $y_2 \in C(z_2)$ such that $\rho(y_1, y_2) \leq \rho(z_1, z_2)$. Observe that $\rho(z_1, z_2) \leq \max(\rho(z_1, z_0), \rho(z_0, z_2)) = \rho(z_1, z_0)$. So $\rho(y_2, y_0) \leq \max(\rho(y_2, y_1), \rho(y_1, y_0)) \leq \max(\rho(z_2, z_1, \rho(z_1, z_0)) = \rho(z_1, z_0) = \rho(z_2, z_0)$. Therefore, $y_2 \in \tilde{C}(z_2)$.

By Step 1 and Step 2 together and Kuratowski–Zorn's lemma, we obtain the existence of a nonexpansive (single-valued) selection $\alpha : N \mapsto M$ of the set-valued map $A : N \rightsquigarrow M$. □

REFERENCES

[1] J.-P. Aubin, *Viability Theory*, Birkhäuser, Boston, Basel, Berlin, 1991.

[2] J.-P. Aubin and A. Cellina, *Differential Inclusions*, Springer-Verlag, Berlin, 1984.

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

[4] G. Barles, *Discontinuous viscosity solutions of first-order Hamilton-Jacobi equations: A guided visit*, Nonlinear Anal., 9 (1993), pp. 1123–1134.

[5] E. N. Barron and R. Jensen, *Generalized viscosity solutions for Hamilton-Jacobi equations with time-measurable hamiltonians*, J. Differential Equations, 68 (1987), pp. 10–21.

[6] E. N. Barron and R. Jensen, *Optimal control and semicontinuous viscosity solutions*, Proc. Amer. Math. Soc. 113, 1992, pp. 397–402.

[7] P. Cardaliaguet, *A differential game with two players and one target*, SIAM J. Control Optim., 34 (1996), pp. 1441–1460.

[8] P. Cardaliaguet and S. Plaskacz, *Viability and invariance for differential games with applications to Hamilton-Jacobi-Isaacs equations*, in Topology in Nonlinear Analysis, Banach Center Publ. 35, Polish Academy of Sciences, 1996, pp. 149–158.

[9] M. G. Crandall and P.-L. Lions, *Viscosity solutions of Hamilton-Jacobi equations*, Trans. Amer. Math. Soc., 277 (1983), pp. 1–42.

[10] R. J. Elliott and N. J. Kalton, *The existence of value in differential games*, Mem. Amer. Math. Soc. 126, (1972).

[11] L. C. Evans and P. E. Souganidis, *Differential games and representation formulas for solutions of Hamilton-Jacobi-Isaacs equations,* Indiana Univ. Math. J. 33, (1984), pp. 773–797.

[12] H. Frankowska, *Optimal trajectories associated to a solution of contingent Hamilton-Jacobi equations*, Appl. Math. Optim., 19 (1989), pp. 291–311.

[13] H. Frankowska, *Lower semicontinuous solutions of Hamilton-Jacobi-Bellman equations*, SIAM J. Control Optim., 31 (1993), pp. 257–272.

[14] H. Frankowska and S. Plaskacz, *A measurable upper semicontinuous viability theorem for tubes*, Nonlinear Anal., 26 (1996), pp. 565–582.

[15] H. Frankowska, S. Plaskacz, and T. Rzeżuchowski, *Théorèmes de viabilité mesurables et l'équation d'Hamilton-Jacobi-Bellman*, C. R. Acad. Sci. Paris Sér. 1 Math. 315, 1992, pp. 131–134.

[16] H. Frankowska, S. Plaskacz, and T. Rzeżuchowski, *Measurable viability theorems and Hamilton-Jacobi-Bellman equation*, J. Differential Equations, 116 (1995), pp. 265–305.

[17] H. Ishii, *Hamilton-Jacobi equations with discontinuous Hamiltonians on arbitrary open sets*, Bull. Fac. Sci. Engrg Chuo Univ., 28 (1985), pp. 33–77.

[18] P.-L. Lions and B. Perthame, *Remarks on Hamilton-Jacobi equations with measurable time-dependent Hamiltonians*, J. Nonlinear Anal., 11 (1987), pp. 613–621.

SIAM J. CONTROL OPTIM.
Vol. 38, No. 5, pp. 1521–1533

ON SOME NECESSARY CONDITIONS OF OPTIMALITY FOR A NONLOCAL VARIATIONAL PRINCIPLE*

JULIO MUÑOZ†

Abstract. The aim of this work is to deduce optimality conditions for a nonlocal variational principle in the one-dimensional scalar case. We consider the relaxation of the problem in terms of Young measures. This relaxation is a new problem where we perform variations to derive optimality conditions, and those conditions yield explicit information about minimizers in the homogeneous case. It also provides a method capable of finding minimizers for some specific problems.

Key words. optimization, nonlocality, optimality conditions, relaxation

AMS subject classifications. 49K27, 49J27

PII. S0363012998342829

1. Introduction. This paper is devoted to the obtainment of generalized equilibrium conditions for the one-dimensional scalar variational principle

$$\inf \{J(u) : u \in \mathcal{A}\}, \tag{1.1}$$

where

$$J(u) = \int_{I\times I} W\left(u'(x_1), u'(x_2)\right) dx_1 dx_2, \tag{1.2}$$

$$\mathcal{A} = \left\{ u \in W^{1,p}(I) : u - u_0 \in W_0^{1,p}(I) \right\}, \tag{1.3}$$

I is an open interval in $\mathbf{R}$, and $u_0 \in W^{1,p}(I)$ such that $J(u_0) < \infty$.

It is well known that the existence of solutions is strongly related to the weak lower semicontinuity of the functional (1.2). For the problem above this property is equivalent to a sort of inequality of nonlocal nature: the energy density W must satisfy

$$\sum_{i,j=1}^{2n} W(\lambda_i, \lambda_j) \geq 4 \sum_{i,j=1}^{n} W\left(\frac{\lambda_{2i-1}+\lambda_{2i}}{2}, \frac{\lambda_{2j-1}+\lambda_{2j}}{2}\right) \tag{1.4}$$

for any $n \in \mathbf{N}$ and any choice $\lambda_1, \lambda_2, \ldots, \lambda_{2n} \in \mathbf{R}$ (see [22]).

Unfortunately inequalities like (1.4) are complicated to check and that makes the analysis of existence too difficult. Moreover, this kind of problem is different if we compare it to the relaxation of nonconvex classical variational problems where we can substitute the original problem with its convexified version. The nonlocal nature of problems like (1.1)–(1.3) seems to block this approach because an equivalent convexification does not make sense, or at least it is not so clear in this situation. We shall deal with this question later.

*Received by the editors August 3, 1999; accepted for publication (in revised form) November 3, 1999; published electronically May 26, 2000. This research was supported by DGES (Spain) through grant PB96-0534.

http://www.siam.org/journals/sicon/38-5/34282.html

†ETSI Industriales, UCLM, 13071 C. Real, Spain (jmunoz@ind-cr.uclm.es).

Notice we restrict our attention to the homogenous problem, $W = W(u'(x_1), u'(x_2))$. The general case $W = W(x_1, x_2, u'(x_1), u'(x_2))$ is briefly discussed in section 6.

Our approach to the problem is not new (cf. Young [31], [32], McShane [18], Warga [28]). In order to study the original problem (1.1)–(1.3) we shall consider its relaxation in terms of the Young measures, generated by sequences of gradients of admissible functions. We shall call it the *generalized problem* (see (2.6)–(2.8)). Its analysis on problems without the weak lower semicontinuity property can be used to anticipate the oscillatory behavior of minimizing sequences. Also, the existence of solutions to the original problem (1.1)–(1.3) can be dealt with once the generalized problem is solved.

The contribution of this work is to provide a tool to study the generalized problem. We mainly concentrate on the obtainment of necessary conditions of optimality for this principle. We get generalized equilibrium conditions (Theorems 3.1 and 3.2). They are established by making variations on the Young measures and are only useful when we are dealing with the homogeneous problem (i.e., when W depends only on the gradients). In that case they enable us to solve and describe the optimal structure in some examples. It also helps us to understand the appearance of microstructure and its dependence on the imposed boundary conditions.

Some basic references on optimization and relaxation are [2], [7], [8], [10], [13], [15], [17], [23], [26], [28], and [33]. About variational calculus using Young measures, [9], [16], [19], [20], [21], [22], [23], [24], [29], and [33] can be looked at. The analysis of principles similar to (1.1)–(1.3) can be found in [1], [11], [12], [22], and [25].

The paper is organized as follows. In section 2 we revise some preliminaries and tools. Section 3 is devoted to the variational analysis and the equilibrium conditions. In section 4 we apply those conditions exploring one example, already proposed in [22]. In section 5 we propose a relaxation through what we shall call the nonlocal convex envelope. In the final section we talk about the limitations of our method for the nonhomogeneous problem.

2. Some preliminaries and tools. Consider the optimization problem (1.1)–(1.3). We assume the density energy

$$W : \mathbf{R} \times \mathbf{R} \to \mathbf{R} \tag{2.1}$$

is smooth and satisfies the bounds

$$c\left(|A_1|^p + |A_2|^p - 1\right) \le W(A_1, A_2) \le C\left(|A_1|^p + |A_2|^p + 1\right), \tag{2.2}$$

$$\left|\frac{\partial W}{\partial A_i}(A_1, A_2)\right| \le C\left(|A_1|^{p-1} + |A_2|^{p-1} + 1\right), \tag{2.3}$$

$$\left|\frac{\partial^2 W}{\partial A_i^2}(A_1, A_2)\right| \le C\left(|A_1|^{p-2} + |A_2|^{p-2} + 1\right), \tag{2.4}$$

$i = 1, 2$, $2 < p < \infty$, $0 < c < C$. For simplicity we assume $I = (0, 1)$ and $u_0(x) = \gamma x$, $\gamma \in \mathbf{R}$, so that (1.1)–(1.3) can be written in a simpler way:

$$\inf\left\{\int_0^1 \int_0^1 W(u'(x_1), u'(x_2))\, dx_1 dx_2 : u \in W^{1,p}(I), \right. \\ \left. \text{with } u(0) = 0 \text{ and } u(1) = \gamma\right\}. \tag{2.5}$$

As we have mentioned, the lack of the weak lower semicontinuity property or the difficulties to check it induces us to consider the following problem:

$$\inf\left\{\overline{J}(\nu) : \nu \in \overline{\mathcal{A}}\right\}, \tag{2.6}$$

where

$$\overline{J}(\nu) = \int_{I\times I}\int_{\mathbf{R}\times\mathbf{R}} W(\lambda_1,\lambda_2)\, d\nu_{x_1}(\lambda_1)\, d\nu_{x_2}(\lambda_2)\, dx_1 dx_2 \tag{2.7}$$

and $\overline{\mathcal{A}}$ is the set of Young measures $\nu = \{\nu_x\}_{x\in I}$ such that

$$\int_I\int_{\mathbf{R}} |\lambda|^p\, d\nu_x(\lambda)\, dx < \infty, \quad \int_I\int_{\mathbf{R}} \lambda d\nu_x(\lambda)\, dx = \gamma. \tag{2.8}$$

Here, we follow [22] from which we stress the following results. First, Theorem 2.1 characterizes the Young measures generated by weakly convergent sequences of the form $\{(u'(x_1), u'(x_2))\}$. (This result can also be directly deduced using denseness results of Dirac Young measures (cf. [29], [3]).) By using Theorem 2.1 we can easily prove the second result, Theorem 2.2, which guarantees that under the preceding hypotheses, (2.6)–(2.8) is a generalized version of (2.5).

THEOREM 2.1. *$\Lambda_{(x_1,x_2)}$ is the Young measure generated by $\left\{\left(u_j'(x_1), u_j'(x_2)\right)\right\}$, $\{u_j(x)\}$ a bounded sequence in $W^{1,p}(I)$, if and only if*

$$\Lambda_{(x_1,x_2)} = \nu_{x_1}\otimes\nu_{x_2}$$

and

$$\int_I\int_{\mathbf{R}} |\lambda|^p\, d\nu_x(\lambda)\, dx < \infty,$$

where $\nu = \{\nu_x\}_{x\in I}$ is the Young measure generated by $\left\{u_j'(x)\right\}$.

THEOREM 2.2. *Under (2.2) there exists $\upsilon \in \overline{\mathcal{A}}$ such that*

$$m = \overline{J}(\upsilon) = \inf\left\{\overline{J}(\nu) : \nu \in \overline{\mathcal{A}}\right\}, \tag{2.9}$$

where m is the infimum given in (1.1).

Regarding Theorem 2.1 and the fundamental theorem of Young measures [5], [23], [26], we have the representation

$$\lim_{j\to\infty}\int_{I\times I} \psi\left(u_j'(x_1), u_j'(x_2)\right) dx_1 dx_2$$
$$= \int_{I\times I}\int_{\mathbf{R}\times\mathbf{R}} \psi(\lambda_1,\lambda_2)\, d\nu_{x_1}(\lambda_1)\, d\nu_{x_2}(\lambda_2)\, dx_1 dx_2,$$

where $\nu = \left\{\nu_{(x_1,x_2)}\right\}_{(x_1,x_2)\in I\times I}$ is the Young measure generated by the sequence of pairs $\left\{\left(u_j'(x_1), u_j'(x_2)\right)\right\}$, provided ψ is continuous and $\left\{\psi\left(u_j'(x_1), u_j'(x_2)\right)\right\}$ converges weakly in $L^1(I\times I)$. Besides, Theorem 2.2 guarantees at least the existence of a minimizer $\upsilon \in \overline{\mathcal{A}}$ for the generalized functional $\overline{J}$. Thus, for any minimizing sequence $\{w_j\} \subset \mathcal{A}$ there is a Young measure $\upsilon \in \overline{\mathcal{A}}$ such that $m = \overline{J}(\upsilon) = \lim_{i\to\infty} J(w_j)$.

Let us now consider the following: ν, a homogeneous Young measure in $\overline{\mathcal{A}}$; $\{u_j(x)\}$, a sequence in $\mathcal{A}$ such that $\left\{u_j'(x)\right\}$ generates ν; and $\{\psi_j(x)\}$, a bounded

sequence in $W_0^{1,p}(I)$ generating a homogeneous Young measure. Let $\mu = \{\mu_x\}_{x\in I}$ be the Young measure generated by the sequence of pairs $\left\{\left(u_j'(x), \psi_j'(x)\right)\right\}$. Notice that $\left\{\left(u_j'(x), \psi_j'(x)\right)\right\}$ does not necessarily generate a homogeneous Young measure even if each one of its components does.

THEOREM 2.3 (see [14], [27]). *Under the above circumstances, for any $x \in I$ and any $(\lambda_1, \lambda_2) \in \operatorname{supp} \mu_x$,*

$$\mu_x(\lambda_1, \lambda_2) = \mu_x(\lambda_2|\lambda_1) \otimes \nu(\lambda_1), \tag{2.10}$$

where $\mu_x(\cdot|\lambda_1)$ is a probability measure for any $\lambda_1 \in \operatorname{supp}\nu$ and the map

$$\lambda_1 \to \int_{\mathbf{R}} f(\lambda_1, \lambda_2)\, d\mu_x(\lambda_2|\lambda_1)$$

is ν-measurable, provided f is integrable with respect to μ_x.

The decomposition (2.10)[1] implies

$$\int_{\mathbf{R}\times\mathbf{R}} f(\lambda_1, \lambda_2)\, d\mu_x(\lambda_1, \lambda_2) = \int_{\mathbf{R}} \left(\int_{\mathbf{R}} f(\lambda_1, \lambda_2)\, d\mu_x(\lambda_2|\lambda_1) \right) d\nu(\lambda_1).$$

We also need the following result.

PROPOSITION 2.4 (see [24]).

(i) *If H is continuous and is assumed to verify $|H(\lambda)| \le C\left(|\lambda|^{p-1} + 1\right)$, $p > 1$, $C > 0$, and*

$$\int_{\mathbf{R}} H(\lambda)\, \Upsilon(\lambda)\, d\nu(\lambda_1) = 0$$

for any ν-measurable function Υ such that

$$\int_{\mathbf{R}} \Upsilon(\lambda)\, d\nu(\lambda) = 0 \quad \text{and} \quad \int_{\mathbf{R}} |\Upsilon(\lambda)|^p\, d\nu(\lambda) < \infty, \tag{2.11}$$

then

$$H(\lambda) = \int_{\mathbf{R}} H(\lambda)\, d\nu(\lambda)$$

for any $\lambda \in \operatorname{supp}\nu$.

(ii) *If G is continuous and verifies $|G(\lambda)| \le C\left(|\lambda|^{p-2} + 1\right)$, $p > 2$, $C > 0$, and*

$$\int_{\mathbf{R}} G(\lambda)\, \Gamma(\lambda)\, d\nu(\lambda) \ge 0$$

for any ν-measurable and positive function Γ such that

$$\int_{\mathbf{R}} (\Gamma(\lambda))^{p/2}\, d\nu(\lambda) < \infty, \tag{2.12}$$

then

$$G(\lambda) \ge 0$$

for any $\lambda \in \operatorname{supp}\nu$.

[1] In Young measure theory this is usually written as $\mu_x(d\lambda_1, d\lambda_2) = \mu_x(d\lambda_2|\lambda_1) \otimes \nu(d\lambda_1)$, which distinguishes parameters from the integration variable.

3. Generalized equilibrium conditions. We study the generalized version of the optimization principle (2.5) under the hypotheses (2.2)–(2.4). The generalized version of (2.5) is the minimization of the functional

$$\overline{J}(\nu) = \int_{I\times I}\int_{\mathbf{R}\times\mathbf{R}} W(\lambda_1,\lambda_2)\, d\nu_{x_1}(\lambda_1)\, d\nu_{x_2}(\lambda_2)\, dx_1 dx_2, \tag{3.1}$$

with $\overline{\mathcal{A}}$ as the set of admissibility. The independence on the x_i permits us to simplify the problem. We can restrict $\overline{\mathcal{A}}$ to the subset composed by its homogeneous members. To be more precise, given $\nu = \{\nu_x\}_{x\in I} \in \overline{\mathcal{A}}$, we can consider its homogenization $\overline{\nu}$, also a probability measure in $\overline{\mathcal{A}}$, such that $\overline{J}(\nu) = \overline{J}(\overline{\nu})$. The proof of this statement is easy: recall that $\overline{\nu}$ is defined via the formula

$$\langle \overline{\nu}, \chi\rangle = \fint_I \int_{\mathbf{R}} \chi(\lambda)\, d\nu_x(\lambda)\, dx$$

for any continuous function χ,[2] and consequently

$$\int_{\mathbf{R}} \lambda d\overline{\nu}(\lambda) = \gamma. \tag{3.2}$$

Then

$$\overline{J}(\nu) = \int_0^1 \int_{\mathbf{R}} L(\lambda_2)\, d\nu_{x_2}(\lambda_2)\, dx_2,$$

where

$$\begin{aligned} L(\lambda_2) &= \int_0^1 \left[\int_{\mathbf{R}} W(\lambda_1,\lambda_2)\, d\nu_{x_1}(\lambda_1)\right] dx_1 \\ &= \int_0^1 (\overline{\nu}, W(\cdot,\lambda_2))\, dx_1 \\ &= \int_0^1 \int_{\mathbf{R}} W(\lambda_1,\lambda_2)\, d\overline{\nu}(\lambda_1)\, dx_1. \end{aligned}$$

In the same way we have

$$\begin{aligned} \overline{J}(\nu) &= \int_0^1 \int_{\mathbf{R}} L(\lambda_2)\, d\nu_{x_2}(\lambda_2)\, dx_2 \\ &= \int_0^1 \int_{\mathbf{R}} L(\lambda_2)\, d\overline{\nu}(\lambda_2)\, dx_2 \\ &= \int_0^1 \int_0^1 \left[\int_{\mathbf{R}}\int_{\mathbf{R}} W(\lambda_1,\lambda_2)\, d\overline{\nu}(\lambda_1)\, d\overline{\nu}(\lambda_2)\right] dx_1 dx_2 \\ &= \overline{J}(\overline{\nu}). \end{aligned}$$

We analyze (3.1) assuming that the elements competing in the principle are only the homogeneous Young measures ν of $\overline{\mathcal{A}}$ such that $\int_{\mathbf{R}} \lambda d\nu(\lambda) = \gamma$. We denote this set

[2]That is the definition for the homogenization of a Young measure; however, the average does not affect the integral because $|I| = 1$.

of admissibility by $\overline{\mathcal{A}}_h^{\gamma}$. Let the framework be the one introduced before Theorem 2.3. We consider the gradients $w_j'(x) = u_j'(x) + t\psi_j'(x)$, where $\{u_j'(x)\}$ generates the homogeneous Young measure ν, a minimizer of the general principle (3.1). If we take $\{w_j'\}$ and consider μ^t, its homogeneous Young measure in $\overline{\mathcal{A}}_h^{\gamma}$, we set the function

$$h(t) \doteq \overline{J}(\mu^t) = \int_{\mathbf{R}} \int_{\mathbf{R}} W(\alpha_1, \alpha_2)\, d\mu^t(\alpha_1)\, d\mu^t(\alpha_2), \quad t \in \mathbf{R}.$$

Consequently,

$$h(t) = \int_0^1 \int_0^1 \int_{\mathbf{R}^2} \int_{\mathbf{R}^2} W(\lambda_1 + t\delta_1, \lambda_2 + t\delta_2)\, d(\gamma_{x_1}(\lambda_1, \delta_1) \otimes \gamma_{x_2}(\lambda_2, \delta_2))\, dx_1 dx_2,$$

where $\gamma = \{\gamma_{x_i}(\lambda_i, \delta_i)\}_{x_i \in I}$ is the Young measure generated by the sequences of pairs $\{(u_j'(x_i), \psi_j'(x_i))\}$, $i = 1, 2$.

The function h has a minimum for $t = 0$. By the smoothness assumptions on W we have the classical equilibrium conditions $h'(0) = 0$ and $h''(0) \geq 0$. Henceforth the point is to express these conditions in a more transparent way.

By Theorem 2.3 we have the decomposition

$$\gamma_{x_i}(\lambda_i, \delta_i) = \gamma_{x_i}(\delta_i | \lambda_i) \otimes \nu(\lambda_i). \tag{3.3}$$

We define the homogenized Young measure of γ by

$$\langle \overline{\gamma}, \chi(\lambda_i, \delta_i) \rangle = \fint_0^1 \int_{\mathbf{R}^2} \chi(\lambda_i, \delta_i)\, d\gamma_{x_i}(\lambda_i, \delta_i)\, dx_i.$$

Then

$$\langle \overline{\gamma}, \chi(\cdot) \rangle = \int_{\mathbf{R}} \chi(\lambda_i)\, d\nu(\lambda_i) = \langle \nu, \chi(\cdot) \rangle,$$

which implies that ν is the canonical projection onto $\mathbf{R}$ of $\overline{\gamma}$ ($\nu(E) = \overline{\gamma}(E \times \mathbf{R})$). Now, defining $\overline{\gamma}(\delta_i | \lambda_i)$ via the formula

$$\langle \overline{\gamma}(\delta_i | \lambda_i), \chi(\cdot) \rangle = \fint_0^1 \int_{\mathbf{R}} \chi(\delta_i)\, \gamma_{x_i}(\delta_i | \lambda_i)\, dx_i$$

(χ continuous) and using (3.3) we see

$$\begin{aligned} \langle \overline{\gamma}, \chi(\lambda_i, \delta_i) \rangle &= \int_{\mathbf{R}} \left[\fint_0^1 \int_{\mathbf{R}} \chi(\lambda_i, \delta_i)\, d\gamma_{x_i}(\delta_i | \lambda_i)\, dx_i \right] d\nu(\lambda_i) \\ &= \int_{\mathbf{R}^2} \chi(\lambda_i, \delta_i)\, d(\overline{\gamma}(\delta_i | \lambda_i) \otimes \nu(\lambda_i)). \end{aligned}$$

Therefore,

$$\overline{\gamma}(\lambda_i, \delta_i) = \overline{\gamma}(\delta_i | \lambda_i) \otimes \nu(\lambda_i), \tag{3.4}$$

and consequently h can be read as

$$h(t) = \int_{\mathbf{R}^2} \int_{\mathbf{R}^2} W(\lambda_1 + t\delta_1, \lambda_2 + t\delta_2)\, d(\overline{\gamma}(\lambda_1, \delta_1) \otimes \overline{\gamma}(\lambda_2, \delta_2)).$$

We should only take into account that $h'(t) = 0$ ((2.3) and (2.4) permit us the interchange of derivation and integration; see [6, p. 215]) to arrive at

$$0 = \int_{\mathbf{R}^2} \int_{\mathbf{R}^2} \left\{ \frac{d}{dt} W(\lambda_1 + t\delta_1, \lambda_2 + t\delta_2) \Big|_{t=0} \right\} d(\overline{\gamma}(\lambda_1, \delta_1) \otimes \overline{\gamma}(\lambda_2, \delta_2)),$$

which means

$$0 = \int_{\mathbf{R}^2} \int_{\mathbf{R}^2} \left\{ \frac{\partial W(\lambda_1, \lambda_2)}{\partial A_1} \delta_1 + \frac{\partial W(\lambda_1, \lambda_2)}{\partial A_2} \delta_2 \right\} d(\overline{\gamma}(\lambda_1, \delta_1) \otimes \overline{\gamma}(\lambda_2, \delta_2)),$$

so that by (3.4) we can write

$$\begin{aligned} 0 &= \int_{\mathbf{R}^2 \times \mathbf{R}^2} \left\{ \frac{\partial W(\lambda_1, \lambda_2)}{\partial A_1} \delta_1 + \frac{\partial W(\lambda_1, \lambda_2)}{\partial A_2} \delta_2 \right\} d(\overline{\gamma}(\delta_1 | \lambda_1) \otimes \nu(\lambda_1)) \, d(\overline{\gamma}(\delta_2 | \lambda_2) \otimes \nu(\lambda_2)) \\ &= \int_{\mathbf{R}^2} \left\{ \frac{\partial W(\lambda_1, \lambda_2)}{\partial A_1} \int_{\mathbf{R}} \delta_1 d\overline{\gamma}(\delta_1 | \lambda_1) + \frac{\partial W(\lambda_1, \lambda_2)}{\partial A_2} \int_{\mathbf{R}} \delta_2 d\overline{\gamma}(\delta_2 | \lambda_2) \right\} d\nu(\lambda_1) \, d\nu(\lambda_2). \end{aligned}$$

If we denote $\Upsilon(\lambda_j) \doteq \int_{\mathbf{R}} \delta_j d\overline{\gamma}(\delta_j | \lambda_j)$, $j = 1, 2$, the above equilibrium conditions read as

$$0 = \int_{\mathbf{R}^2} \left\{ \frac{\partial W(\lambda_1, \lambda_2)}{\partial A_1} \Upsilon(\lambda_1) + \frac{\partial W(\lambda_1, \lambda_2)}{\partial A_2} \Upsilon(\lambda_2) \right\} d\nu(\lambda_1) \, d\nu(\lambda_2),$$

and by change of variables we have

$$\begin{aligned} 0 &= \int_{\mathbf{R}^2} \left\{ \frac{\partial W(\lambda_1, \lambda_2)}{\partial A_1} + \frac{\partial W(\lambda_2, \lambda_1)}{\partial A_2} \right\} \Upsilon(\lambda_1) \, d\nu(\lambda_1) \, d\nu(\lambda_2) \\ &= \int_{\mathbf{R}} \left\{ \int_{\mathbf{R}} \left\{ \frac{\partial W(\lambda_1, \lambda_2)}{\partial A_1} + \frac{\partial W(\lambda_2, \lambda_1)}{\partial A_2} \right\} d\nu(\lambda_2) \right\} \Upsilon(\lambda_1) \, d\nu(\lambda_1) \\ &= \int_{\mathbf{R}} H(\lambda_1) \Upsilon(\lambda_1) \, d\nu(\lambda_1), \end{aligned}$$

where

$$H(\lambda_1) \doteq \int_{\mathbf{R}} \left\{ \frac{\partial W}{\partial A_1}(\lambda_1, \lambda_2) + \frac{\partial W}{\partial A_2}(\lambda_2, \lambda_1) \right\} d\nu(\lambda_2).$$

We observe

$$\begin{aligned} \int_{\mathbf{R}} \Upsilon(\lambda_j) \, d\nu(\lambda_j) &= \int_{\mathbf{R}^2} \int_0^1 \delta_j d\gamma_x(\delta_j | \lambda_j) \, dx d\nu(\lambda_j) \\ &= \int_0^1 \int_{\mathbf{R}^2} \delta_j d\gamma_x(\delta_j | \lambda_j) \, d\nu(\lambda_j) \\ &= \int_0^1 \int_{\mathbf{R}^2} \delta_j d\gamma_x(\lambda_j, \delta_j) \, dx = 0, \end{aligned}$$

and by Jensen's inequality

$$\begin{aligned} \int_{\mathbf{R}} |\Upsilon(\lambda_j)|^p \, d\nu(\lambda_j) &= \int_{\mathbf{R}} \left| \int_{\mathbf{R}} \delta_j d\overline{\gamma}(\delta_j | \lambda_j) \right|^p d\nu(\lambda_j) \\ &\leq \int_{\mathbf{R}} \int_{\mathbf{R}} |\delta_j|^p \, d\overline{\gamma}(\delta_j | \lambda_j) \, d\nu(\lambda_j) \\ &= \int_0^1 \int_{\mathbf{R}^2} |\delta_j|^p \, d\gamma_x(\lambda_j, \delta_j) \, dx < \infty. \end{aligned}$$

Reciprocally, for any field Υ fulfilling (2.11), we can find a sequence $\left\{\left(u_j'\left(x_i\right),\psi_j'\left(x_i\right)\right)\right\}$, where $\left\{u_j'\left(x\right)\right\}$generates ν and $\left\{\psi_j\left(x\right)\right\}$ is a bounded sequence in $W_0^{1,p}\left(I\right)$ such that its Young measure can be written as

$$\mu\left(\lambda,\delta\right)=\mu\left(\delta|\,\lambda\right)\otimes\nu\left(\lambda\right)$$

and $\int_{\mathbf{R}}\delta d\mu\left(\delta|\,\lambda\right)=\Upsilon$ (see [22] and [23] for a complete discussion). Therefore, we can now use Proposition 2.4(i) to state the following result.

THEOREM 3.1. *Under the above circumstances*

$$\operatorname{supp}\nu\subset\left\{\lambda\in\mathbf{R}:H\left(\lambda\right)=\int_{\mathbf{R}}H\left(\lambda_1\right)d\nu\left(\lambda_1\right)\right\}.$$

Now, our investigation is concerned with the condition $h''(0)\geq 0$. According to this inequality and by some simple computations we have

$$\begin{aligned}0\leq&\int_{\mathbf{R}^2}\left\{\frac{\partial^2W}{\partial A_1^2}\left(\lambda_1,\lambda_2\right)\int_{\mathbf{R}}\delta_1^2d\overline{\gamma}\left(\delta_1|\,\lambda_1\right)+\frac{\partial^2W}{\partial A_2^2}\left(\lambda_1,\lambda_2\right)\int_{\mathbf{R}}\delta_2^2d\overline{\gamma}\left(\delta_2|\,\lambda_2\right)\right.\\&\left.+2\frac{\partial^2W}{\partial A_1\partial A_2}\left(\lambda_1,\lambda_2\right)\int_{\mathbf{R}}\delta_1d\overline{\gamma}\left(\delta_1|\,\lambda_1\right)\int_{\mathbf{R}}\delta_2d\overline{\gamma}\left(\delta_2|\,\lambda_2\right)\right\}d\nu\left(\lambda_1\right)d\nu\left(\lambda_2\right)\\=&\int_{\mathbf{R}^2}\left\{\frac{\partial^2W}{\partial A_1^2}\left(\lambda_1,\lambda_2\right)\Gamma\left(\lambda_1\right)+\frac{\partial^2W}{\partial A_2^2}\left(\lambda_1,\lambda_2\right)\Gamma\left(\lambda_2\right)\right.\\&\left.+2\frac{\partial^2W}{\partial A_1\partial A_2}\left(\lambda_1,\lambda_2\right)\Upsilon\left(\lambda_1\right)\Upsilon\left(\lambda_2\right)\right\}d\nu\left(\lambda_1\right)d\nu\left(\lambda_2\right),\end{aligned}$$

where

$$\Gamma\left(\lambda_i\right)=\int_{\mathbf{R}}\delta_i^2d\overline{\gamma}\left(\delta_i|\,\lambda_i\right).$$

By Jensen's inequality, if

$$2\Upsilon\left(\lambda_1\right)\Upsilon\left(\lambda_2\right)\leq\Upsilon\left(\lambda_1\right)^2+\Upsilon\left(\lambda_2\right)^2\leq\Gamma\left(\lambda_1\right)+\Gamma\left(\lambda_2\right),$$

then

$$\begin{aligned}0\leq&\int_{\mathbf{R}^2}\left\{\frac{\partial^2W}{\partial A_1^2}\left(\lambda_1,\lambda_2\right)\Gamma\left(\lambda_1\right)+\frac{\partial^2W}{\partial A_2^2}\left(\lambda_1,\lambda_2\right)\Gamma\left(\lambda_2\right)\right.\\&\left.+\frac{\partial^2W}{\partial A_1\partial A_2}\left(\lambda_1,\lambda_2\right)\Gamma\left(\lambda_1\right)+\frac{\partial^2W}{\partial A_1\partial A_2}\left(\lambda_1,\lambda_2\right)\Gamma\left(\lambda_2\right)\right\}d\nu\left(\lambda_1\right)d\nu\left(\lambda_2\right),\end{aligned}$$

and by changing the variables we have

$$\begin{aligned}0\leq&\int_{\mathbf{R}}\left\{\int_{\mathbf{R}}\left\{\frac{\partial^2W}{\partial A_1^2}\left(\lambda_1,\lambda_2\right)+\frac{\partial^2W}{\partial A_2^2}\left(\lambda_2,\lambda_1\right)\right.\right.\\&\left.\left.+\frac{\partial^2W}{\partial A_1\partial A_2}\left(\lambda_1,\lambda_2\right)+\frac{\partial^2W}{\partial A_1\partial A_2}\left(\lambda_2,\lambda_1\right)\right\}d\nu\left(\lambda_2\right)\right\}\Gamma\left(\lambda_1\right)d\nu\left(\lambda_1\right).\end{aligned}$$

Obviously $\Gamma\left(\lambda_i\right)\geq 0$, and again by Jensen's inequality ($p\geq 2$),

$$\begin{aligned}\int_{\mathbf{R}}\left|\Gamma\left(\lambda_i\right)\right|^{p/2}d\nu\left(\lambda_i\right)&=\int_{\mathbf{R}}\left|\int_{\mathbf{R}}\delta_i^2d\overline{\gamma}\left(\delta_i|\,\lambda_i\right)\right|^{p/2}d\nu\left(\lambda_i\right)\\&\leq\int_{\mathbf{R}}\int_{\mathbf{R}}\left|\delta_i\right|^p d\overline{\gamma}\left(\delta_i|\,\lambda_i\right)d\nu\left(\lambda_i\right)\\&=\int_0^1\int_{\mathbf{R}^2}\left|\delta_i\right|^p d\gamma_x\left(\lambda_i,\delta_i\right)<\infty.\end{aligned}$$

As before, given any positive Γ verifying (2.12), we can build a sequence of pairs whose Young measure is written in the form $\mu(\lambda,\delta)=\mu(\delta|\lambda)\otimes\nu(\lambda)$ and such that the second moment $\int_{\mathbf{R}}\delta^2 d\mu(\delta|\lambda)$ coincides with Γ. So, by Proposition 2.4(ii)

$$G(\lambda_1)\geq 0$$

for any $\lambda_1\in\operatorname{supp}\nu$, where

$$G(\lambda_1)\doteq\int_{\mathbf{R}}\left\{\frac{\partial^2 W}{\partial A_1^2}(\lambda_1,\lambda_2)+\frac{\partial^2 W}{\partial A_2^2}(\lambda_2,\lambda_1)+\frac{\partial^2 W}{\partial A_1\partial A_2}(\lambda_1,\lambda_2)+\frac{\partial^2 W}{\partial A_1\partial A_2}(\lambda_2,\lambda_1)\right\}d\nu(\lambda_2).$$

THEOREM 3.2.

$$\operatorname{supp}\nu\subset\{\lambda\in\mathbf{R}:G(\lambda)\geq 0\}.$$

4. One example. We show how Theorems 3.1 and 3.2 may be applied to solve the generalized problem and to decide about the existence of solutions to the original problem (2.5), in some simple situations.

Let us take

$$W(\lambda_1,\lambda_2)=\left(\lambda_1^2+\lambda_2^2-1\right)^2$$

and suppose ν is a minimizer for (3.1). The equilibrium condition of Theorem 3.1 gives in this case the identity

$$\int_{\mathbf{R}}\int_{\mathbf{R}}\left\{8\lambda_1\left(\lambda_1^2+\lambda_2^2-1\right)\right\}d\nu(\lambda_2)\,d\nu(\lambda_1)=\int_{\mathbf{R}}\left\{8\lambda_1\left(\lambda_1^2+\lambda_2^2-1\right)\right\}d\nu(\lambda_2)$$

for any $\lambda_1\in\operatorname{supp}\nu$. That is,

$$\begin{aligned}\int_{\mathbf{R}}\lambda_1^3 d\nu(\lambda_1)+\int_{\mathbf{R}}\lambda_1 d\nu(\lambda_1)\int_{\mathbf{R}}\lambda_2^2 d\nu(\lambda_2)-\int_{\mathbf{R}}\lambda_1 d\nu(\lambda_1)\\ =\lambda_1^3+\lambda_1\int_{\mathbf{R}}\lambda_2^2 d\nu(\lambda_2)-\lambda_1.\end{aligned}\tag{4.1}$$

This fact shows that at most there exist three different mass points for ν; let us denote them by γ_1,γ_2, and γ_3, and suppose $\nu=\alpha_1\delta_{\gamma_1}+\alpha_2\delta_{\gamma_2}+\alpha_3\delta_{\gamma_3}$, where $\alpha_i\in[0,1]$, $\sum_{i=1}^3\alpha_i=1$. If we write (4.1) in terms of the moments for ν, we obtain

$$\gamma_j^3+(m_2-1)\gamma_j=m_3+m_1(m_2-1),$$

where $m_k=\sum_{j=1}^3\alpha_j\gamma_j^k$, $k=1,2,3$. We supply the above identity with the constraint $u(x)=\gamma\cdot x$ on $\partial(0,1)$ (see (3.2)), thus $m_1=\gamma$; and finally, we arrive at the system of equations

$$\gamma_j^3+\left(\sum_{j=1}^3\alpha_j\gamma_j^2-1\right)\gamma_j=\left(\sum_{j=1}^3\alpha_j\gamma_j^3\right)+\left(\sum_{j=1}^3\alpha_j\gamma_j^2-1\right)\gamma,\tag{4.2}$$

$$\sum_{j=1}^3\alpha_j\gamma_j=\gamma,\tag{4.3}$$

whose unknowns are γ_1, γ_2, γ_3, α_1, and α_2. Moreover, by Theorem 3.2 we have for any γ_j, $j = 1, 2, 3$, the constraints

$$\Upsilon(\gamma_j) = 3\gamma_j^2 + 2m_1\gamma_j + m_2 - 1 \geq 0. \tag{4.4}$$

Notice that the information about minimizing measures has been reduced to the study of the nonlinear system (4.2)–(4.3) with the constraints (4.4). In general the resulting system has a family of solutions. Once these solutions are found and substituted in (3.1), we would have reduced the difficulties in the task of finding minimizers.

In the present example we get the following solutions to (4.2)–(4.3).

R-1. If $\gamma \in [-\frac{\sqrt{2}}{2}, \frac{\sqrt{2}}{2}]$, we obtain
(i) $\nu = \delta_c$, with $c \in [-\frac{\sqrt{2}}{2}, \frac{\sqrt{2}}{2}]$,
(ii) $\nu = d\delta_{\frac{\sqrt{2}}{2}} + (1-d)\delta_{-\frac{\sqrt{2}}{2}}$, and
(iii) $\nu = d\delta_{\frac{-1}{2}} + (1-d)\delta_{\frac{1}{2}}$.

R-2. If $\gamma > \frac{\sqrt{2}}{2}$ or $\gamma < -\frac{\sqrt{2}}{2}$, we have $\nu = \delta_c$, where $c > \frac{\sqrt{2}}{2}$ or $c < -\frac{\sqrt{2}}{2}$, respectively.

In the range R-1 the unique solution is $\nu = \frac{1+\gamma\sqrt{2}}{2}\delta_{\frac{\sqrt{2}}{2}} + \frac{1-\gamma\sqrt{2}}{2}\delta_{-\frac{\sqrt{2}}{2}}$. We discard $\nu = d\delta_{-\frac{1}{2}} + (1-d)\delta_{\frac{1}{2}}$ because by using (4.4) we necessarily get that $d = \frac{1}{2}$ and $\overline{J}(\frac{1}{2}\delta_{-\frac{1}{2}} + \frac{1}{2}\delta_{\frac{1}{2}}) > 0$. In R-2 there exists only one solution, $\nu = \delta_\gamma$. Consequently we can write the infimum of the problem as a function of γ:

$$\overline{J}_W(\gamma) = \begin{cases} 0 & \text{if } -\frac{\sqrt{2}}{2} \leq \gamma \leq \frac{\sqrt{2}}{2}, \\ \left(2\gamma^2 - 1\right)^2 & \text{otherwise.} \end{cases}$$

There are some other examples for which the optimality conditions can be applied. For instance, we can take

$$W(\lambda_1, \lambda_2) = ((\lambda_1 + \lambda_2)^2 - 1)^2 + (\lambda_1 - \lambda_2)\arctan(\lambda_1 + \lambda_2)^2$$

or

$$W(\lambda_1, \lambda_2) = (\lambda_1^2 + \lambda_2^2 - 1)^2 + \arctan\lambda_1^2 + \arctan\lambda_2^2.$$

5. Relaxation. The purpose of this section is to suggest a new relaxation of the homogeneous problem

$$\inf\left\{\int_{I\times I} W(u'(x), u'(y))\,dxdy : u \in W_0^{1,p}(I) + \gamma.x\right\}. \tag{5.1}$$

The idea is to maintain the density energy in all regions where the infimum of the principle (3.1) is attained by a Dirac delta. In those regions where the solutions give rise to oscillations we consider a new density energy whose definition is similar to the traditional convexification. Specifically, for every function W satisfying the conditions (2.2) we define its *nonlocal convexification* as

$$\widetilde{C}W = \sup\left\{S : S \leq W,\ \overline{J}_S(\gamma) = S(\gamma,\gamma) = \overline{J}_W(\gamma) \qquad \text{for any } \gamma \in \mathbf{R}\right\},$$

where

$$\overline{J}_G(\gamma) \doteq \inf\left\{\int_{\mathbf{R}\times\mathbf{R}} G(\lambda_1, \lambda_2)\,d\mu(\lambda_1)\,d\mu(\lambda_2) : \mu \in \overline{\mathcal{A}}_h^\gamma\right\}.$$

The leading idea is to find a new variational principle whose infimum is attained and coincides with m_γ, the infimum of the original problem (5.1). To achieve this we use the variational principle associated with the nonlocal convexification (see (5.2) below).

Unfortunately we cannot ensure that

$$\mathcal{S} \doteq \left\{S : S \leq W,\ \overline{J}_S(\gamma) = S(\gamma,\gamma) = \overline{J}_W(\gamma) \qquad \text{for any } \gamma \in \mathbf{R}\right\}$$

is a nonempty set. Even though we can consider separately the convex functions S such that $S \leq W$ and $\overline{J}_S(\gamma) = S(\gamma,\gamma)$,[3] we will not be able to state $S(\gamma,\gamma) = \overline{J}_W(\gamma)$ for any $\gamma \in \mathbf{R}$. If we suppose this obstacle has been overcome and we can get the corresponding generalized formulation for the principle

$$\inf\left\{\int_{I\times I} \widetilde{C}W(u'(x),u'(y))\,dxdy : u \in W_0^{1,p}(I) + \gamma.x\right\}, \tag{5.2}$$

then we would have that for any $\gamma \in \mathbf{R}$, $m_\gamma = \widetilde{m}_\gamma$ and there would exist $v \in W_0^{1,p}(I) + \gamma.x$ such that $\int_{I\times I} \widetilde{C}W(v'(x),v'(y))\,dxdy = \widetilde{m}_\gamma$, where $\widetilde{m}_\gamma$ is the infimum of (5.2). The proof is straightforward.

Notice that for the obtainment of the nonlocal convexification we have to find $\overline{J}_W(\gamma)$ for any γ and after getting that, we must investigate the kind of Young measures which are solutions to this principle, making the emphasis on when they are Dirac deltas. In spite of the fact that this process could be carried out, we shall not be able to define $\widetilde{C}W$ in all $\mathbf{R}^2$. That reduces the significance of the nonlocal convex envelope. We illustrate this by examining the example dealt in section 4: we know that $\nu = \frac{1+\gamma\sqrt{2}}{2}\delta_{\frac{\sqrt{2}}{2}} + \frac{1-\gamma\sqrt{2}}{2}\delta_{\frac{-\sqrt{2}}{2}}$ is the minimizer if $\gamma \in [\frac{-1}{\sqrt{2}}, \frac{1}{\sqrt{2}}]$ such that $\overline{J}_W(\gamma) = 0$. If the Young measure is a delta, $\overline{J}_W(\gamma) = W(\gamma,\gamma) = \left(2\gamma^2 - 1\right)^2$, $\gamma \notin [\frac{-1}{\sqrt{2}}, \frac{1}{\sqrt{2}}]$. Those facts induce us to define $\widetilde{C}W$, only in a subset of $\mathbf{R}^2$, as

$$\widetilde{C}W(\lambda_1,\lambda_2) = \begin{cases} W(\lambda_1,\lambda_2) & \text{if } |\lambda_1|, |\lambda_2| \geq \frac{\sqrt{2}}{2}, \\ 0 & \text{if } |\lambda_1|, |\lambda_2| < \frac{\sqrt{2}}{2}. \end{cases} \tag{5.3}$$

To find a complete definition of $\widetilde{C}W$ in $\mathbf{R}^2$ remains an open question. The difficulties are obvious. We observe that basically the obstacle is the nonemptiness of $\mathcal{S}$. Somehow, this task involves a nonlocal problem which clearly turns up when we try to define the convexification out of the diagonal $\{\lambda_1 = \lambda_2\}$. In any case, further work must be done in order to establish rigorously any kind of relaxation, perhaps like (5.2), which could be used successfully in some examples.

6. Final remarks. The analysis exhibited in sections 2 and 3 can be carried out if W depends on $(u'(x_1), \ldots, u'(x_n))$. Nevertheless, when $W = W(x_1, \ldots, x_n, u(x_1), \ldots, u(x_2), u'(x_1), \ldots, u'(x_n))$ this sort of analysis is not useful to derive necessary conditions. In this case a possible variational analysis can be performed taking Young measures generated by sequences of the type $\left\{u_j'(x_1) + t\psi'(x_1)\right\}$, where $\psi \in W_0^{1,p}(0,1)$ and $\left\{u_j'(x_1)\right\}$ is any sequence generating a minimizer (see [20]). If $\nu = \{\nu_x\}_{x\in(0,1)}$ is any minimizer of (2.6)–(2.8) with $W = W(x_1,x_2,\lambda_1,\lambda_2)$, then

$$\frac{\partial}{\partial x_1}\left\{\int_0^1 \left[A(x_1,x_2) + B(x_2,x_1)\right] dx_2\right\} = 0 \tag{6.1}$$

[3]Recall that separated convexity is a sufficient condition for the weak lower semicontinuity (cf. [22]).

in a weak sense, and

$$\int_0^1 C(x_1,x_2)\,dx_2 \geq 0 \quad \text{almost everywhere } x_1 \in (0,1), \tag{6.2}$$

where

$$A(x_1,x_2) = \int_{\mathbf{R}^2} \frac{\partial W}{\partial A_1}(x_1,x_2,\lambda_1,\lambda_2)\,d\nu_{x_1}(\lambda_1)\,d\nu_{x_2}(\lambda_2),$$
$$B(x_1,x_2) = \int_{\mathbf{R}^2} \frac{\partial W}{\partial A_2}(x_1,x_2,\lambda_1,\lambda_2)\,d\nu_{x_1}(\lambda_1)\,d\nu_{x_2}(\lambda_2),$$

and

$$C(x_1,x_2) = \int_{\mathbf{R}^2} \left\{ \frac{\partial^2 W}{\partial A_1^2}(x_1,x_2,\lambda_1,\lambda_2) + \frac{\partial^2 W}{\partial A_1 \partial A_2}(x_1,x_2,\lambda_1,\lambda_2) + \frac{\partial^2 W}{\partial A_1 \partial A_2}(x_2,x_1,\lambda_2,\lambda_1) + \frac{\partial^2 W}{\partial A_2^2}(x_2,x_1,\lambda_2,\lambda_1) \right\} d\nu_{x_1}(\lambda_1)\,d\nu_{x_2}(\lambda_2).$$

Clearly, the equilibrium conditions (6.1) and (6.2) do not seem to be the appropriate way to detect generalized minimizers. Even for the homogeneous case, the equilibrium conditions (Theorems 3.1 and 3.2) could be imprecise or not explicit enough to obtain solutions. In either case, it seems to be necessary to implement some additional information: for instance, we might, in some problems, complete the obtained necessary conditions with some sort of result limiting the number of Dirac deltas appearing (as a convex combination) in the representation of the solution. This approach has been used in Balder [4] within the frame of nonconvex optimal control problems, Winkler [30] for the minimization of affine functionals defined on moments sets, and also Bonnetier and Conca [7], [8] for a problem in optimal design.

Finally, we want to point out that the analysis of section 3 in higher dimensions becomes tremendously difficult. The crucial point is that we have to consider Young measures $\mu(\lambda_1,\lambda_2)$ coming from sequences of gradients of $\{(u_j(x),\psi_j(x))\}$, $x \in \Omega$, which are vector valued functions. This fact implies profound restrictions on the measure μ, and therefore on the fields Γ and Υ. The difficulties we have to face require a deeper analysis than the one performed here.

Acknowledgments. The author would like to thank Pablo Pedregal for several useful discussions on various issues related to this paper and would also like to acknowledge the referees for their careful reading of the original manuscript and their valuable suggestions and criticism.

REFERENCES

[1] G. Alberti and G. Bellettini, *A Nonlocal Anisotropic Model for Phase Transitions. Part* I: *The Optimal Profile Problem*, preprint, 1991.

[2] E. J. Balder, *A general approach to lower semicontinuity and lower closure in optimal control theory*, SIAM J. Control Optim., 22 (1984), pp. 570–598.

[3] E. J. Balder, *Consequences of denseness of Dirac Young measures*, J. Math. Anal. Appl., 207 (1997), pp. 536–540.

[4] E. J. Balder, *New existence results for optimal controls in the absence of convexity: The importance of extremality*, SIAM J. Control Optim., 32 (1994), pp. 890–916.

[5] J. M. Ball, *A version of the fundamental theorem for Young measures*, in PDEs and Continuum Models of Phase Transitions, M. Rascle, D. Serre, and M. Slemrod, eds., Lecture Notes in Phys. 344, Springer, Berlin, New York, 1989, pp. 207–215.

[6] P. Billingsley, *Probability and Measure*, Wiley, New York, 1986.
[7] E. Bonnetier and C. Conca, *Relaxation totale d'un problème d'optimisation de plaques*, C. R. Acad. Sci. Paris Sér. I Math., 317 (1994), pp. 931–936.
[8] E. Bonnetier and C. Conca, *Approximation of Young measures by functions and application to a problem of optimal design for plates with variable thickness*, Proc. Roy. Soc. Edinburgh Sect. A, 124 (1994), pp. 399–422.
[9] E. Bonnetier and C. Conca, *Optimality conditions for a relaxed layout optimization problem*, C. R. Acad. Sci. Paris Sér. I Math., 327 (1998), pp. 1005–1010.
[10] H. Berliocchi and J. M. Lasry, *Intégrandes normales et mesure pramétrées en calcul des variations*, Bull. Soc. Math. France, 101 (1973), pp. 129–184.
[11] D. Brandon and R. Rogers, *The coercivity and nonlocal ferromagnetism*, Contin. Mech. Thermodyn., 4 (1992), pp. 1–21.
[12] D. Brandon and R. Rogers, *Nonlocal regularization of L. C. Young's tacking problem*, Appl. Math. Opt., 25 (1992), pp. 287–301.
[13] B. Dacorogna, *Direct Methods in the Calculus of Variations*, Springer-Verlag, Berlin, New York, 1989.
[14] L. C. Evans, *Weak Convergence Methods for Nonlinear Partial Differential Equations*, CBMS Reg. Conf. Ser. Math. 74, American Mathematical Society, Providence, RI, 1990.
[15] G. Friesecke, *A necessary and sufficient condition for nonatttainment and formation of microstructure almost everywhere in scalar variational problems*, Proc. Roy. Soc. Edinburgh Sect., A, 124 (1994), pp. 437–471.
[16] R. V. Gamkrelidze, *Principles of Optimal Control Theory*, Plenum Press, New York, 1978.
[17] R. V. Kohn and G. Strang, *Optimal design and relaxation of variational problems:* I, II *and* III, Comm. Pure Appl. Math., 39 (1986), pp. 113-137, 139–182, and 353–377.
[18] E. J. McShane, *Generalized curves*, Duke Math. J., 6 (1940), pp. 512–536.
[19] E. J. McShane, *Necessary conditions in the generalized curve problem of calculus of variations*, Duke Math. J., 7 (1940), pp. 1–27.
[20] J. Muñoz, *Sobre algunos problemas insolubles de optimización*, Ph.D. thesis, Dpto. Mat. Apl., Facultad de Matemáticas, Universidad Complutense de Madrid, Spain, 1997.
[21] J Muñoz and P. Pedregal, *On the relaxation of an optimal design problem for plates*, Asymptotic Anal., 16 (1998), pp. 125–140.
[22] P. Pedregal, *Nonlocal variational principles*, Nonlinear Anal., 12 (1997), pp. 1379–1392.
[23] P. Pedregal, *Parametrized Measures and Variational Principles*, Birkhäuser–Verlag, Basel, Switzerland, 1997.
[24] P. Pedregal, *Equilibrium conditions for Young measures*, SIAM J. Control Optim., 36 (1998), pp. 797–813.
[25] R. Rogers, *A nonlocal model for the exchange energy in ferromagnetic materials*, J. Integral Equations Appl., 3 (1988), pp. 85–127.
[26] T. Roubicek, *Relaxation in Optimization Theory and Variational Calculus*, Walter de Gruyter, Berlin, 1997.
[27] T. Valadier, *Desintégration d'une mesure sur un produit*, C. R. Acad. Sci. Paris Sér. A, 276 (1973), pp. 33–35.
[28] J. Warga, *Relaxed variational problems*, J. Math. Appl., 4 (1962), pp. 111–128.
[29] J. Warga, *Optimal Control of Differential and Functional Equations*, Academic Press, New York, 1972.
[30] G. Winkler, *Extreme points of moment sets*, Math. Oper. Res., 13, (1988), pp. 581–587.
[31] L. C. Young, *Generalized curves and the existence of an attained absolute minimum in the calculus of variations*, C. R. Soc. Sci. Lett. Varsovie, Classe II, 30 (1937), pp. 212–234.
[32] L. C. Young, *Generalized surfaces in the calculus of variations* I *and* II, Ann. of Math. II, 43 (1937), pp. 84–103 and 530–544.
[33] L. C. Young, *Lectures on Calculus of Variations and Optimal Control Theory*, W. B. Saunders, Philadelphia, 1969.

SIAM J. CONTROL OPTIM.
Vol. 38, No. 5, pp. 1534–1553

THEORETICAL AND NUMERICAL ANALYSIS OF AN OPTIMAL CONTROL PROBLEM RELATED TO WASTEWATER TREATMENT*

A. MARTÍNEZ†, C. RODRÍGUEZ‡, AND M. E. VÁZQUEZ-MÉNDEZ‡

Abstract. In this work we deal with the design and management of wastewater treatment systems, mainly the disposal of sea outfalls discharging polluting effluent from a sewerage system. This problem is formulated as a pointwise optimal control problem with state and control constraints. The main difficulties arise from the lack of regularity of the second member in the state system and from the pointwise constraints on the state variables. We develop the theoretical analysis of the problem, we propose an algorithm for its numerical resolution, and finally, we give results for a realistic problem posed in the *ría* of Vigo, Spain.

Key words. pointwise control, pointwise state constraints, wastewater treatment, constrained optimization

AMS subject classifications. 49J20, 49K20, 49M37

PII. S0363012998345640

1. Introduction: The physical problem.

We consider a domain Ω with boundary Γ occupied by shallow water (as can be the case of a *ría*, an estuary, or a lake), where polluting wastewater is discharged through N_E outfalls, each one corresponding to a depuration plant. We also assume that inside the domain Ω there exist several zones ($A_i \subset \Omega,\ i = 1, \ldots, N_Z$) of bath, marine cultures, and so on, where it is necessary to ensure the water quality with pollution concentrations lower than an allowed maximum level.

In order to control the marine pollution some parameters are used that indicate the quality level of liquid media and its capacity to hold the aquatic life. Among these indicators we mention dissolved oxygen, temperature, and pH.

Oxygen is used by bacteria to decompose the organic matter. This process can be measured in terms of the need of oxygen, the so-called *biological oxygen demand* (BOD). If the pollution level is not too high this need can be satisfied by the *dissolved oxygen* (DO). If the quantity of organic matter increases beyond a maximum value the DO is not enough to decompose it, leading to important modifications (anaerobic processes) in the ecosystem. This fact means that in each protected area $A_i, i = 1, \ldots, N_Z$, a threshold value σ_i of BOD must not be exceeded and a minimum level of DO ζ_i must be guaranteed. To ensure this, one has to depure wastewater by chemical or biological treatments before discharging it into the sea.

*Received by the editors October 2, 1998; accepted for publication (in revised form) November 29, 1999; published electronically May 26, 2000. This research was supported by project XUGA20701A98 of Xunta de Galicia, Spain.

http://www.siam.org/journals/sicon/38-5/34564.html

†Departamento de Matemática Aplicada, ETSI Telecomunicacion, Universidade de Vigo, 36200 Vigo, Spain (aurea@dma.uvigo.es).

‡Departamento de Matemática Aplicada, Universidade de Santiago de Compostela, 15706 Santiago, Spain (carmen@zmat.usc.es, emesto@lugo.usc.es).

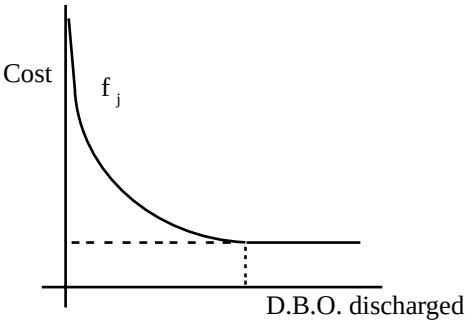


FIG. 1.1. *Standard function* $f_j(m)$.

The cost of the depuration in the jth plant, $j = 1, \ldots, N_E$, can be assumed to depend on the BOD in such a way that a lower level of BOD leads to a more intensive depuration and, consequently, to a higher cost. Taking in mind that absolute depuration is not feasible and that there exists a minimum cost, even in the case where no treatment is developed, the cost function f_j takes a form similar to the one shown in Figure 1.1.

The problem is then to determine the level of discharges in order to minimize the global depuration cost and to guarantee the above-mentioned constraints on the water quality in protected areas. Mathematically, this is a parabolic optimal control problem with pointwise state constraints and with pointwise control.

In recent years, several authors have studied parabolic optimal control problems with pointwise state constraints (see, for instance, Casas [4], Fattorini and Sritharan [8], Hu and Yong [12], Raymond and Zidani [18]). Lions [13], Saguez [19], Simon [20], and others have dealt with optimal control problems of parabolic type with pointwise control, but they did not include state constraints. Bermúdez, Martínez, and Rodríguez [2] studied a related stationary problem with pointwise state constraints and considered a control on the location. However, the dynamic problem with pointwise control and pointwise state constraints has not yet been, as far as we know, reported in the open literature.

From the theoretical point of view the main difficulties of the problem are due to the fact that the second member of the state system includes radon measures and to the presence of pointwise state constraints, which lead us to work in regular functional spaces in order to obtain the adjoint state.

Numerically, difficulties arise from the high number of constraints related to the time and space discretizations and to the demands imposed on the levels of BOD and DO.

2. Mathematical modeling. Let us suppose that the outfalls are located in the points $P_j \in \Omega,\ j = 1, \ldots, N_E$, and denote by $m_j(t),\ j = 1, \ldots, N_E$, the discharge of BOD in the point P_j at the time t. The evolution of the BOD and the DO in the domain $\Omega \subset R^2$ is governed, according to the distributed version of the model of Streter–Phelps, by the following partial differential equations system (cf. [1] and [16]), whose numerical resolution has been carried out in [21]. In this way, the concentrations

of BOD and DO in a point $x \in \Omega$ and a time $t \in (0,T)$, denoted $\rho_1(x,t)$ and $\rho_2(x,t)$, respectively, can be obtained as the solution of the boundary value problem

$$
\begin{aligned}
\frac{\partial \rho_1}{\partial t} + \vec{u}\nabla\rho_1 - \beta_1\Delta\rho_1 &= -\kappa_1\rho_1 + \frac{1}{h}\sum_{j=1}^{N_E} m_j\delta(x-P_j) && \text{in } \Omega\times(0,T),\\
\frac{\partial \rho_1}{\partial n} &= 0 && \text{on } \Gamma\times(0,T),\\
\rho_1(x,0) &= 0 && \text{in } \Omega,\\
\frac{\partial \rho_2}{\partial t} + \vec{u}\nabla\rho_2 - \beta_2\Delta\rho_2 &= -\kappa_1\rho_1 + \frac{1}{h}\kappa_2(d_s-\rho_2) && \text{in } \Omega\times(0,T),\\
\frac{\partial \rho_2}{\partial n} &= 0 && \text{on } \Gamma\times(0,T),\\
\rho_2(x,0) &= \rho_{20}(x) && \text{in } \Omega,
\end{aligned}
\tag{2.1}
$$

where $h(x,t)$ and $\vec{u}(x,t)$ denote, respectively, the height and the mean horizontal velocity of the fluid layer, obtained as a solution of the Saint-Venant equations (see [3]), $\delta(x-P_j)$ represents the Dirac measure in the point P_j, and parameters $\beta_1 > 0,\ \beta_2 > 0$ (horizontal viscosity coefficients involving dispersion and turbulence effects), $\kappa_1 > 0,\ \kappa_2 > 0$ (kinetic coefficients related to temperature and transference of oxygen through the surface), and d_s (oxygen saturation density) can be obtained from experimental measurements.

If we assume now that inside the domain Ω there are N_Z protected zones A_i, where a maximum level of BOD and a minimum level of DO must be ensured, that is,

$$
\rho_{1|A_i\times(0,T)} \le \sigma_i, \quad i = 1,\dots,N_Z, \tag{2.2}
$$

$$
\rho_{2|A_i\times(0,T)} \ge \zeta_i, \quad i = 1,\dots,N_Z, \tag{2.3}
$$

and we know the convex functions $f_j \in C^2(0,\infty)$ (the treatment cost of the discharge in the point $P_j,\ j = 1,\dots,N_E$) and, consequently, the global cost of the depuration system in a time interval $[0,T]$, which is given by

$$
J(m) = \sum_{j=1}^{N_E}\int_0^T f_j\left(m_j(t)\right)dt, \tag{2.4}
$$

then the problem $(\mathcal{P})$ of the optimal management of the depuration systems consists of finding the values of BOD $m_j(t) > 0,\ j = 1,\dots,N_E$, throughout the time interval in such a way that they satisfy the state system (2.1) and the constraints (2.2) and (2.3) and they minimize the cost function (2.4).

3. Analysis of the state system. Let $\Omega \subset R^2$ be a bounded domain with boundary Γ smooth enough. We make the following assumptions on the problem data:

$$
h \in C(\bar{\Omega}\times[0,T]) \quad h(x,t) \ge \alpha > 0 \quad \forall (x,t) \in \bar{\Omega}\times[0,T],
$$

$$
\vec{u} \in [L^\infty(\Omega\times(0,T))]^2, \quad \rho_{20} \in C^2(\bar{\Omega}), \quad m = (m_j)_{j=1}^{N_E} \in U_{ad},
$$

where

$$U_{ad} = \{m \in (L^\infty(0,T))^{N_E} \ : \ 0 < \underline{m}_j \le m_j(t) \le \overline{m}_j \text{ almost everywhere (a.e.) in } (0,T), \\ j = 1, \dots, N_E\}.$$

The weak solution of the system (2.1) can be defined by transposition techniques (see Lions and Magenes [14], Casas [4]) in the following way.

DEFINITION 3.1. *Given* $r, s \in [1,2), \frac{2}{r} + \frac{2}{s} > 3$, *we say that* $\rho = (\rho_1, \rho_2) \in [L^r(0,T; W^{1,s}(\Omega))]^2$ *is a solution of the system* (2.1) *if* $\forall\ \Phi = (\Phi_1, \Phi_2) \in [L^2(0,T; H^2(\Omega)) \cap H^1(0,T; L^2(\Omega))]^2 \cap [C^1(\bar{\Omega} \times [0,T])]^2$ *such that* $\Phi(.,T) = 0$, *it is verified that*

$$\begin{aligned} \int_0^T \int_\Omega \Big\{ & - \frac{\partial \Phi_1}{\partial t}\rho_1 - \frac{\partial \Phi_2}{\partial t}\rho_2 + \beta_1 \nabla\Phi_1 \nabla\rho_1 + \beta_2 \nabla\Phi_2 \nabla\rho_2 + \vec{u}\Phi_1 \nabla\rho_1 \\ & + \vec{u}\Phi_2 \nabla\rho_2 + \kappa_1 \Phi_1 \rho_1 + \kappa_1 \Phi_2 \rho_1 + \frac{1}{h(x,t)} \kappa_2 \Phi_2 \rho_2) \Big\} \, dx \, dt \\ & = \sum_{j=1}^{N_E} \int_0^T \frac{1}{h(P_j,t)} \Phi_1(P_j,t) m_j(t) \, dt \\ & + \int_0^T \int_\Omega \frac{1}{h(x,t)} \kappa_2 d_s \Phi_2(x,t) \, dx \, dt + \int_\Omega \Phi_2(x,0) \rho_{20}(x) \, dx. \end{aligned} \tag{3.1}$$

Let $\mathcal{A}$ be the operator defined by

$$\begin{aligned} \langle \mathcal{A}(w_1,w_2), (z_1,z_2) \rangle = \int_\Omega \Big(& - \beta_1 \Delta w_1 z_1 - \beta_2 \Delta w_2 z_2 \\ & + \vec{u} \cdot \nabla w_1 z_1 + \vec{u} \cdot \nabla w_2 z_2 + \kappa_1 w_1 z_1 + \kappa_1 w_1 z_2 + \frac{1}{h} \kappa_2 w_2 z_2 \Big) \, dx \end{aligned}$$

for $(w_1, w_2), (z_1, z_2)$ such that the previous expression makes sense. Then we have the following result.

THEOREM 3.2. *There exists a unique pair*

$$\rho = (\rho_1, \rho_2) \in [L^r(0,T; W^{1,s}(\Omega))]^2 \cap [L^2(0,T; L^2(\Omega))]^2$$

with

$$\frac{\partial \rho}{\partial t} = \left(\frac{\partial \rho_1}{\partial t}, \frac{\partial \rho_2}{\partial t} \right) \in [L^r(0,T; (W^{1,s'}(\Omega))')]^2$$

$\forall\ r, s \in [1,2), \frac{2}{r} + \frac{2}{s} > 3$, *such that* ρ *is the solution of* (2.1) *and satisfies*

$$\begin{aligned} \int_0^T \left\langle -\frac{\partial \Phi}{\partial t} + \mathcal{A}^*(\Phi), \rho \right\rangle dt = & \sum_{j=1}^{N_E} \int_0^T \frac{1}{h(P_j,t)} \Phi_1(P_j,t) m_j(t) \, dt \\ & + \int_0^T \int_\Omega \frac{1}{h(x,t)} \kappa_2 d_s \Phi_2(x,t) \, dx \, dt + \int_\Omega \Phi_2(x,0) \rho_{20}(x) \, dx \end{aligned} \tag{3.2}$$

$\forall\ \Phi = (\Phi_1, \Phi_2) \in \mathcal{B}$, *where*

$$\begin{aligned} \mathcal{B} = \Big\{ & \Phi = (\Phi_1, \Phi_2) \in [L^2(0,T; H^2(\Omega)) \cap H^1(0,T; L^2(\Omega))]^2 \ : \\ & - \frac{\partial \Phi}{\partial t} + \mathcal{A}^*(\Phi) \in [L^2(0,T; L^2(\Omega))]^2, \quad \frac{\partial \Phi}{\partial n_{\mathcal{A}^*}}_{|\Gamma \times (0,T)} = 0, \quad \Phi(.,T) = 0 \Big\}. \end{aligned}$$

Besides, there exist constants $C_k,\ k = 1, \dots, 6$, *only depending on data, such that*

$$\|\rho\|_{[L^r(0,T;W^{1,s}(\Omega))]^2} \le C_1 \sum_{i=1}^{N_E} \|m_i\|_{L^\infty(0,T)} + C_2 \|\rho_{20}\|_{C(\bar\Omega)} + C_3\, d_s \tag{3.3}$$

and

$$\|\rho\|_{[L^2(0,T;L^2(\Omega))]^2} \le C_4 \sum_{i=1}^{N_E} \|m_i\|_{L^\infty(0,T)} + C_5 \|\rho_{20}\|_{C(\bar\Omega)} + C_6\, d_s. \tag{3.4}$$

Proof. In the proof of this result we will adapt the one of Casas [4, Theorem 6.3] in order to obtain the new L^2-estimate (3.4) which is possible in this case due to the fact that the second member takes the form $\sum_{j=1}^{N_E} m_j(t)\delta(x - P_j)$.

Thus, let $\{f_j^k\}_{k\in N} \subset C(\bar\Omega),\ j = 1, \dots, N_E$, be weak* convergent sequences to $\delta(x - P_j)$ in the measures space $M(\bar\Omega) = C(\bar\Omega)'$, verifying $\|f_j^k\|_{L^1(\Omega)} \le 1$.

Let us consider $\rho^k = (\rho_1^k, \rho_2^k)$ the solution of system

$$\begin{aligned}
\frac{\partial \rho_1^k}{\partial t} + \vec u \nabla \rho_1^k - \beta_1 \Delta \rho_1^k + \kappa_1 \rho_1^k &= \frac{1}{h} \sum_{j=1}^{N_E} m_j f_j^k && \text{in } \Omega \times (0,T),\\
\frac{\partial \rho_1^k}{\partial n} &= 0 && \text{on } \Gamma \times (0,T),\\
\rho_1^k(x,0) &= 0 && \text{in } \Omega,\\
\frac{\partial \rho_2^k}{\partial t} + \vec u \nabla \rho_2^k - \beta_2 \Delta \rho_2^k + \kappa_1 \rho_1^k + \frac{1}{h}\kappa_2 \rho_2^k &= \frac{1}{h}\kappa_2 d_s && \text{in } \Omega \times (0,T),\\
\frac{\partial \rho_2^k}{\partial n} &= 0 && \text{on } \Gamma \times (0,T),\\
\rho_2^k(x,0) &= \rho_{20}(x) && \text{in } \Omega.
\end{aligned} \tag{3.5}$$

For all $\Psi = (\Psi^0, \Psi^1, \Psi^2) \in [\{\mathcal{D}(\Omega \times (0,T))\}^2]^3$ we denote by y^Ψ the solution of the system

$$\begin{aligned}
-\frac{\partial y_1}{\partial t} - \nabla.(\vec u y_1) - \beta_1 \Delta y_1 + \kappa_1 (y_1 + y_2) &= \Psi_1^0 - \frac{\partial \Psi_1^1}{\partial x_1} - \frac{\partial \Psi_1^2}{\partial x_2} && \text{in } \Omega \times (0,T),\\
\beta_1 \frac{\partial y_1}{\partial n} + \vec u.\vec n\, y_1 &= 0 && \text{on } \Gamma \times (0,T),\\
y_1(x,T) &= 0 && \text{in } \Omega,\\
-\frac{\partial y_2}{\partial t} - \nabla.(\vec u y_2) - \beta_2 \Delta y_2 + \frac{1}{h}\kappa_2 y_2 &= \Psi_2^0 - \frac{\partial \Psi_2^1}{\partial x_1} - \frac{\partial \Psi_2^2}{\partial x_2} && \text{in } \Omega \times (0,T),\\
\beta_2 \frac{\partial y_2}{\partial n} + \vec u.\vec n\, y_2 &= 0 && \text{on } \Gamma \times (0,T),\\
y_2(x,T) &= 0 && \text{in } \Omega.
\end{aligned}$$

Then we have

$$\begin{aligned}
&\int_0^T \int_\Omega \left\{ \sum_{i=1}^2 \Psi_i^0 \rho_i^k + \sum_{i,j=1}^2 \Psi_i^j \frac{\partial \rho_i^k}{\partial x_j} \right\} dx\, dt \int_0^T \left\langle -\frac{\partial y^\Psi}{\partial t} + \mathcal{A}^*(y^\Psi), \rho^k \right\rangle dt \\
&\quad = \int_0^T \left\langle \frac{\partial \rho^k}{\partial t} + \mathcal{A}(\rho^k), y^\Psi \right\rangle dt + \int_\Omega y_2^\Psi(x,0) \rho_{20}(x)\, dx
\end{aligned}$$

$$= \sum_{j=1}^{N_E} \int_0^T \int_\Omega \frac{1}{h(x,t)} f_j^k(x) m_j(t) y_1^\Psi(x,t)\, dx\, dt$$
$$+ \int_0^T \int_\Omega \frac{1}{h(x,t)} \kappa_2 d_s y_2^\Psi(x,t)\, dx\, dt + \int_\Omega y_2^\Psi(x,0)\rho_{20}(x)\, dx.$$

From this we deduce

$$\int_0^T \int_\Omega \left\{ \sum_{i=1}^2 \Psi_i^0 \rho_i^k + \sum_{i,j=1}^2 \Psi_i^j \frac{\partial \rho_i^k}{\partial x_j} \right\} dx\, dt$$
$$\le \frac{1}{\alpha} \sum_{i=1}^{N_E} \|m_i\|_{L^\infty(0,T)} \sum_{j=1}^{N_E} \int_0^T \int_\Omega |f_j^k(x) y_1^\Psi(x,t)|\, dx\, dt$$
$$+ \frac{1}{\alpha} \kappa_2 d_s \int_0^T \int_\Omega |y_2^\Psi(x,t)|\, dx\, dt + \int_\Omega |y_2^\Psi(x,0)\rho_{20}(x)|\, dx$$
$$\le \left\{ \tilde{C}_1 \sum_{i=1}^{N_E} \|m_i\|_{L^\infty(0,T)} + \tilde{C}_2 \|\rho_{20}\|_{C(\bar{\Omega})} + \tilde{C}_3\, d_s \right\} \|y^\Psi\|_{[C(\bar{\Omega}\times[0,T])]^2}.$$

On the other hand, from Ladyzhenskaja, Solonnikov, and Uraltseva [15] and Di Benedetto [6] we get

$$\int_0^T \int_\Omega \left\{ \sum_{i=1}^2 \Psi_i^0 \rho_i^k + \sum_{i,j=1}^2 \Psi_i^j \frac{\partial \rho_i^k}{\partial x_j} \right\} dx\, dt$$
$$\le \left\{ \tilde{C}_4 \sum_{i=1}^{N_E} \|m_i\|_{L^\infty(0,T)} + \tilde{C}_5 \|\rho_{20}\|_{C(\bar{\Omega})} + \tilde{C}_6\, d_s \right\} \sum_{j=0}^2 \|\Psi^j\|_{[L^{r'}(0,T;L^{s'}(\Omega))]^2}.$$

Due to the density of the space $\{\Psi^0 - \frac{\partial \Psi^1}{\partial x_1} - \frac{\partial \Psi^2}{\partial x_2} : \Psi \in [\{\mathcal{D}(\Omega \times (0,T))\}^2]^3\}$ in $[L^{r'}(0,T;(W^{1,s}(\Omega))')]^2$ we deduce the boundedness of $\{\rho^k\}$ in $[L^r(0,T;W^{1,s}(\Omega))]^2$.

Thus, taking a subsequence if necessary, we obtain that $\{\rho^k\}$ weakly converges to ρ in $[L^r(0,T;W^{1,s}(\Omega))]^2$ and that the limit satisfies the estimate (3.3). Moreover, ρ is independent on r and s (cf. Casas [4]). Finally, since $W^{1,s}(\Omega) \subset M(\bar{\Omega}) \subset (W^{1,s'}(\Omega))'$, we have that

$$\frac{\partial \rho}{\partial t} \in [L^r(0,T;(W^{1,s'}(\Omega))')]^2.$$

On the other hand, if we choose $\Psi = (\Psi^0, 0, 0)$ and argue as before, we obtain

$$\int_0^T \int_\Omega \sum_{i=1}^2 \Psi_i^0 \rho_i^k\, dx\, dt \le \left\{ \tilde{C}_7 \sum_{i=1}^{N_E} \|m_i\|_{L^\infty(0,T)} \right.$$
$$\left. + \tilde{C}_8 \|\rho_{20}\|_{C(\bar{\Omega})} + \tilde{C}_9\, d_s \right\} \|y^\Psi\|_{[L^2(0,T;H^2(\Omega))\cap H^1(0,T;L^2(\Omega))]^2}.$$

By the estimates of Ladyzhenskaja, Solonnikov, and Uraltseva [15] we know that

$$\|y^\Psi\|_{[L^2(0,T;H^2(\Omega))\cap H^1(0,T;L^2(\Omega))]^2} \le \tilde{C}_{10} \|\Psi^0\|_{[L^2(0,T;L^2(\Omega))]^2}$$

and we obtain the boundedness of $\{\rho^k\}$ in $[L^2(0,T;L^2(\Omega))]^2$, from which we can deduce that the solution ρ belongs to the space $[L^2(0,T;L^2(\Omega))]^2$ and satisfies the estimate (3.4).

Given $\Phi=(\Phi_1,\Phi_2)\in[L^2(0,T;H^2(\Omega))\cap H^1(0,T;L^2(\Omega))]^2\cap[C^1(\bar\Omega\times[0,T])]^2$ with $\Phi(.,T)=0$, multiplying (3.5) by Φ, and integrating by parts we have

$$\begin{aligned}\int_0^T\int_\Omega\Big\{&-\frac{\partial\Phi_1}{\partial t}\rho_1^k-\frac{\partial\Phi_2}{\partial t}\rho_2^k+\beta_1\nabla\Phi_1\nabla\rho_1^k+\beta_2\nabla\Phi_2\nabla\rho_2^k+\vec u\Phi_1\nabla\rho_1^k\\&+\vec u\Phi_2\nabla\rho_2^k+\kappa_1\Phi_1\rho_1^k+\kappa_1\Phi_2\rho_1^k+\frac{1}{h(x,t)}\kappa_2\Phi_2\rho_2^k)\Big\}\,dx\,dt\\&=\sum_{j=1}^{N_E}\int_0^T\int_\Omega\frac{1}{h(x,t)}\Phi_1(x,t)m_j(t)f_j^k(x)\,dx\,dt\\&+\int_0^T\int_\Omega\frac{1}{h(x,t)}\kappa_2 d_s\Phi_2(x,t)\,dx\,dt+\int_\Omega\Phi_2(x,0)\rho_{20}(x)\,dx.\end{aligned}$$

Passing to the limit we deduce (3.1) and, consequently, ρ is a solution of (2.1).

In order to prove (3.2), given $\Phi=(\Phi_1,\Phi_2)\in\mathcal{B}$, by multiplying (3.5) by Φ and integrating by parts we deduce

$$\begin{aligned}\int_0^T\Big\langle-\frac{\partial\Phi}{\partial t}+\mathcal{A}^*(\Phi),\rho^k\Big\rangle\,dt&=\sum_{j=1}^{N_E}\int_0^T\int_\Omega\frac{1}{h(x,t)}\Phi_1(x,t)m_j(t)f_j^k(x)\,dx\,dt\\&+\int_0^T\int_\Omega\frac{1}{h(x,t)}\kappa_2 d_s\Phi_2(x,t)\,dx\,dt+\int_\Omega\Phi_2(x,0)\rho_{20}(x)\,dx,\end{aligned}$$

and (3.2) is obtained by passing to the limit. The uniqueness of solution can be ensured from (3.2) (see Casas [4]). □

LEMMA 3.3. *Functions ρ_1 and ρ_2 are continuous in $\bar A_i\times[0,T]\ \forall\, i=1,\dots,N_Z$.*

Proof. Let $E\subset\Omega$ be a closed subset smooth enough such that

$$\begin{aligned}&P_j\in E\quad\forall j=1,\dots,N_E,\\&\cup_{i=1}^{N_Z}\bar A_i\subset\subset\Omega\backslash E.\end{aligned}$$

From the weak formulation (3.2) we deduce that

$$\begin{aligned}&\int_0^T\int_{\Omega\backslash E}\left(-\frac{\partial\Phi}{\partial t}+\mathcal{A}^*(\Phi)\right)\rho\,dx\,dt\\&\quad=\int_0^T\int_\Omega\left(-\frac{\partial\Phi}{\partial t}+\mathcal{A}^*(\Phi)\right)\rho\,dx\,dt=\sum_{j=1}^{N_E}\int_0^T\frac{1}{h(P_j,t)}\Phi_1(P_j,t)m_j(t)\,dt\\&\quad+\int_0^T\int_\Omega\frac{1}{h(x,t)}\kappa_2 d_s\Phi_2(x,t)\,dx\,dt+\int_\Omega\Phi_2(x,0)\rho_{20}(x)\,dx\\&\quad=\int_0^T\int_{\Omega\backslash E}\frac{1}{h(x,t)}\kappa_2 d_s\Phi_2(x,t)\,dx\,dt\quad\forall\Phi\in[\mathcal{D}((\Omega\backslash E)\times(0,T))]^2;\end{aligned}$$

that is, ρ satisfies

$$\int_0^T \int_{\Omega\backslash E} \left(-\frac{\partial \Phi}{\partial t} + \mathcal{A}^*(\Phi) \right) \rho \, dx \, dt$$
$$= \int_0^T \int_{\Omega\backslash E} \frac{1}{h(x,t)} \kappa_2 d_s \Phi_2(x,t) \, dx \, dt \quad \forall \Phi \in [\mathcal{D}((\Omega\backslash E) \times (0,T))]^2$$

with initial data $\rho_1(x,0) = 0$ and $\rho_2(x,0) = \rho_{20}$.

Thus, $\rho = (\rho_1, \rho_2) \in [L^2(0,T; H^1_{loc}(\Omega\backslash E))]^2 \cap [C(0,T; L^2_{loc}(\Omega\backslash E))]^2$ is a local solution of the problem

$$\begin{aligned}
\frac{\partial \rho_1}{\partial t} + \vec{u}\nabla\rho_1 - \beta_1 \Delta \rho_1 &= -\kappa_1 \rho_1 && \text{in } (\Omega\backslash E) \times (0,T), \\
\frac{\partial \rho_2}{\partial t} + \vec{u}\nabla\rho_2 - \beta_2 \Delta \rho_2 &= -\kappa_1 \rho_1 + \frac{1}{h}\kappa_2 (d_s - \rho_2) && \text{in } (\Omega\backslash E) \times (0,T), \\
\rho_1(x,0) &= 0 && \text{in } \Omega\backslash E, \\
\rho_2(x,0) &= \rho_{20} && \text{in } \Omega\backslash E.
\end{aligned} \tag{3.6}$$

From this we can obtain (see Ladyzhenskaja, Solonnikov, and Uraltseva [15, Chapter III, Theorem 10.1]) that ρ is continuous in the compact subsets of $(\Omega\backslash E) \times [0,T]$.

Particularly,

$$\rho \in [C(\cup_{i=1}^{N_Z} \bar{A}_i \times [0,T])]^2. \qquad \square$$

In order to obtain optimality conditions for the control problem it will be necessary to obtain the Gateaux derivative of the mappings:

$$F_1 : m \in (L^\infty(0,T))^{N_E} \longrightarrow F_1(m) = \rho_{1|\cup_{i=1}^{N_Z} \bar{A}_i \times [0,T]} \in C(\cup_{i=1}^{N_Z} \bar{A}_i \times [0,T]),$$
$$F_2 : m \in (L^\infty(0,T))^{N_E} \longrightarrow F_2(m) = \rho_{2|\cup_{i=1}^{N_Z} \bar{A}_i \times [0,T]} \in C(\cup_{i=1}^{N_Z} \bar{A}_i \times [0,T]).$$

First, we obtain the following continuity result.

LEMMA 3.4. *There exist constants $\hat{C}_1, \hat{C}_2, \hat{C}_3$ such that*

$$\|\rho\|_{[C(\cup_{i=1}^{N_Z} \bar{A}_i \times [0,T])]^2} \le \hat{C}_1 \sum_{i=1}^{N_E} \|m_i\|_{L^\infty(0,T)} + \hat{C}_2 \|\rho_{20}\|_{C(\bar{\Omega})} + \hat{C}_3 \, d_s.$$

Proof. As a consequence of Lemma 3.3 and the estimates for $\max|\rho(x,t)|$ in $(\Omega\backslash E) \times [0,T]$ (see Ladyzhenskaja, Solonnikov, and Uraltseva [15, Chapter III, Theorem 8.1 and Chapter II, Theorem 6.2]) we have

$$\begin{aligned}
\|\rho\|_{[C(\cup_{i=1}^{N_Z} \bar{A}_i \times [0,T])]^2} &= \|\hat{\rho}\|_{[C(\cup_{i=1}^{N_Z} \bar{A}_i \times [0,T])]^2} \\
&\le \bar{C}_1 \|\hat{\rho}\|_{[L^2(0,\hat{T}; L^2(\Omega\backslash E))]^2} + \bar{C}_2 \|\rho_{20}\|_{C(\overline{\Omega\backslash E})} + \bar{C}_3 \, d_s,
\end{aligned}$$

where $\hat{\rho}$ is the solution of the following boundary value problem:

$$
\begin{aligned}
\frac{\partial \hat{\rho}_1}{\partial t} + \vec{u}\nabla\hat{\rho}_1 - \beta_1\Delta\rho_1 = -\kappa_1\hat{\rho}_1 + \frac{1}{h}\chi_{[0,T]}\sum_{j=1}^{N_E} m_j\delta(x-P_j) \qquad & \text{in } \Omega\times(0,\hat{T}),\\
\frac{\partial \hat{\rho}_1}{\partial n} = 0 \qquad & \text{on } \Gamma\times(0,\hat{T}),\\
\hat{\rho}_1(x,0) = 0 \qquad & \text{in } \Omega,\\
\frac{\partial \hat{\rho}_2}{\partial t} + \vec{u}\nabla\hat{\rho}_2 - \beta_2\Delta\hat{\rho}_2 = -\kappa_1\hat{\rho}_1 + \frac{1}{h}\kappa_2(d_s-\hat{\rho}_2) \qquad & \text{in } \Omega\times(0,\hat{T}),\\
\frac{\partial \hat{\rho}_2}{\partial n} = 0 \qquad & \text{on } \Gamma\times(0,\hat{T}),\\
\hat{\rho}_2(x,0) = \rho_{20}(x) \qquad & \text{in } \Omega
\end{aligned}
\tag{3.7}
$$

for $\hat{T} > T$ and $\chi_{[0,T]}$ the characteristic function of $[0,T]$. Thus, by estimate (3.4) of Theorem 3.2 we deduce

$$
\|\rho\|_{[C(\cup_{i=1}^{N_Z}\bar{A}_i\times[0,T])]^2} \le \hat{C}_1\sum_{i=1}^{N_E}\|m_i\|_{L^\infty(0,T)} + \hat{C}_2\,\|\rho_{20}\|_{C(\bar{\Omega})} + \hat{C}_3\,d_s. \qquad \square
$$

Now, we can deduce the Gateaux differentiability.

LEMMA 3.5. *The mappings F_1 and F_2 are Gateaux differentiable. Moreover,*

$$
\begin{aligned}
DF_1(m)(n) &= \omega_{1\,|\cup_{i=1}^{N_Z}\bar{A}_i\times[0,T]},\\
DF_2(m)(n) &= \omega_{2\,|\cup_{i=1}^{N_Z}\bar{A}_i\times[0,T]},
\end{aligned}
$$

where ω_1 and ω_2 are the solutions of the linearized system:

$$
\begin{aligned}
\frac{\partial \omega_1}{\partial t} + \vec{u}\nabla\omega_1 - \beta_1\Delta\omega_1 = -\kappa_1\omega_1 + \frac{1}{h}\sum_{j=1}^{N_E} n_j\delta(x-P_j) \qquad & \text{in } \Omega\times(0,T),\\
\frac{\partial \omega_1}{\partial n} = 0 \qquad & \text{on } \Gamma\times(0,T),\\
\omega_1(x,0) = 0 \qquad & \text{in } \Omega,\\
\frac{\partial \omega_2}{\partial t} + \vec{u}\nabla\omega_2 - \beta_2\Delta\omega_2 = -\kappa_1\omega_1 - \frac{1}{h}\kappa_2\omega_2 \qquad & \text{in } \Omega\times(0,T),\\
\frac{\partial \omega_2}{\partial n} = 0 \qquad & \text{on } \Gamma\times(0,T),\\
\omega_2(x,0) = 0 \qquad & \text{in } \Omega.
\end{aligned}
\tag{3.8}
$$

Proof. Let (ω_1,ω_2) be the solution of system (3.8). Arguing as in Lemma 3.3 we can prove that

$$
\begin{aligned}
\omega_{1\,|\cup_{i=1}^{N_Z}\bar{A}_i\times[0,T]} &\in C(\cup_{i=1}^{N_Z}\bar{A}_i\times[0,T]),\\
\omega_{2\,|\cup_{i=1}^{N_Z}\bar{A}_i\times[0,T]} &\in C(\cup_{i=1}^{N_Z}\bar{A}_i\times[0,T]).
\end{aligned}
$$

Let us denote by $\rho(m)$ the solution of the state system (2.1) corresponding to a second member m. Thus, for $m,n\in L^\infty(0,T)$ and $\alpha\in(0,1)$ given, since ρ_1 is affine, it satisfies

$$
\rho_1(m+\alpha n)(x,t) - \rho_1(m)(x,t) = \alpha\omega_1(x,t) \quad \forall (x,t)\in\cup_{i=1}^{N_Z}\bar{A}_i\times[0,T].
$$

Then,

$$\lim_{\alpha\to 0}\frac{\rho_1(m+\alpha n)-\rho_1(m)}{\alpha}=\omega_1$$

in $C(\cup_{i=1}^{N_Z}\bar{A}_i\times[0,T])$. As a consequence of Lemma 3.4, the mapping taking n into ω_1 is linear and continuous. Thus,

$$DF_1(m)(n)=\omega_{1\,|\cup_{i=1}^{N_Z}\bar{A}_i\times[0,T]}.$$

Arguing in the same way with ρ_2 we also obtain that

$$DF_2(m)(n)=\omega_{2\,|\cup_{i=1}^{N_Z}\bar{A}_i\times[0,T]}. \qquad \square$$

4. Existence of a solution of the optimal control problem. We have the following theorem.

THEOREM 4.1. *If there exists a feasible control $\tilde{m}\in U_{ad}$ such that*

$$\tilde{\rho_1}_{|A_i\times(0,T)}\le\sigma_i,\quad i=1,\dots,N_Z,$$
$$\tilde{\rho_2}_{|A_i\times(0,T)}\ge\zeta_i,\quad i=1,\dots,N_Z,$$

then the optimal control problem has, at least, a solution.

Proof. Let $\{m^k\}_{n\in N}\in U_{ad}$ be a minimizing sequence. From the boundedness of the sequence we can deduce the existence of a subsequence (still denoted in the same way) that converges weakly in $(L^2(0,T))^{N_E}$ to an element $m\in U_{ad}$. From Theorem 3.2 we deduce that the sequence $\rho^k=(\rho_1^k,\rho_2^k)=(\rho_1(m^k),\rho_2(m^k))$ is bounded in $[L^r(0,T;W^{1,s}(\Omega))]^2$. Thus, for $r,s>1$, we have

$$\rho^k\longrightarrow\rho\quad\text{ weakly in }[L^r(0,T;W^{1,s}(\Omega))]^2.$$

So, passing to the limit, we obtain that ρ satisfies (3.1) and, consequently, $\rho=\rho(m)$.

Arguing as in Lemma 3.4 and taking into account the boundedness of m^k in $L^\infty(0,T)$, we obtain that the sequences ρ_j^k, $j=1,2$, are bounded, for some $\alpha\in(0,1)$, in $C^{\alpha,\alpha/2}(\cup_{i=1}^{N_Z}\bar{A}_i\times[0,T])$ (cf. [15, Chapter III, Theorem 10.1]). Then, using the compact imbedding of $C^{\alpha,\alpha/2}(\cup_{i=1}^{N_Z}\bar{A}_i\times[0,T])$ in $C(\cup_{i=1}^{N_Z}\bar{A}_i\times[0,T])$, we obtain the existence of subsequences, still denoted ρ_j^k, uniformly converging to ρ_j. Thus, from the direct pointwise convergence, we deduce

$$\rho_{1\,|A_i\times(0,T)}\le\sigma_i\quad\forall i=1,\dots,N_Z, \tag{4.1}$$
$$\rho_{2\,|A_i\times(0,T)}\ge\zeta_i\quad\forall i=1,\dots,N_Z. \tag{4.2}$$

Finally, since J is weakly lower semicontinuous (because of continuity and convexity of J), we have

$$J(m)\le\liminf J(m^k)$$

from which we deduce that m is a solution of the optimal control problem. $\square$

5. Optimality conditions. In this section we will obtain a first order optimality system satisfied by every solution of the optimal control problem. In order to express the optimality conditions in a simpler way we introduce functions p_1,p_2 as the solution (in the sense of Definition 5.1 below) of the following boundary value problem:

$$
\begin{aligned}
-\frac{\partial p_1}{\partial t} - \nabla.(\vec{u}p_1) - \beta_1\Delta p_1 + \kappa_1(p_1+p_2) &= \mu_{1|\Omega\times(0,T)} && \text{in } \Omega\times(0,T),\\
\beta_1\frac{\partial p_1}{\partial n} + \vec{u}.\vec{n}\,p_1 &= 0 && \text{on } \Gamma\times(0,T),\\
p_1(x,T) &= \mu_{1|\Omega\times\{T\}} && \text{in } \Omega,\\
-\frac{\partial p_2}{\partial t} - \nabla.(\vec{u}p_2) - \beta_2\Delta p_2 + \frac{1}{h}\kappa_2 p_2 &= \mu_{2|\Omega\times(0,T)} && \text{in } \Omega\times(0,T),\\
\beta_2\frac{\partial p_2}{\partial n} + \vec{u}.\vec{n}\,p_2 &= 0 && \text{on } \Gamma\times(0,T),\\
p_2(x,T) &= \mu_{2|\Omega\times\{T\}} && \text{in } \Omega,
\end{aligned}
\tag{5.1}
$$

where μ_1, μ_2 are regular Borel measures in $\bar{\Omega}\times[0,T]$. The weak solution of the system (5.1) can be defined by transposition techniques (Casas [4]) in the following way.

DEFINITION 5.1. *Given r, $s \in [1,2)$, $\frac{2}{r}+\frac{2}{s} > 3$, we say that $p = (p_1,p_2) \in [L^r(0,T;W^{1,s}(\Omega))]^2$ is a solution of the system (5.1) if $\forall\ z = (z_1,z_2) \in [L^2(0,T;H^1(\Omega)) \cap C^1(\bar{\Omega}\times[0,T])]^2$ such that $z(.,0) = 0$, it is verified that*

$$
\begin{aligned}
\int_0^T\int_\Omega \Big\{ &\frac{\partial z_1}{\partial t}p_1 + \frac{\partial z_2}{\partial t}p_2 + \rho_1\nabla z_1\nabla p_1 + \rho_2\nabla z_2\nabla p_2 + \vec{u}\nabla z_1 p_1 + \vec{u}\nabla z_2 p_2\\
&+\kappa_1 z_1 p_1 + \kappa_1 z_1 p_2 + \frac{1}{h(x,t)}\kappa_2 z_2 p_2\Big\}\,dx\,dt = \int_0^T\int_\Omega z_1 d\mu_1(x,t)\\
&+\int_0^T\int_\Omega z_1 d\mu_1(x,t) + \int_\Omega z_1(x,T)d\mu_{1T}(x) + \int_\Omega z_2(x,T)d\mu_{2T}(x).
\end{aligned}
$$

We denote by I_C the indicator function of a set C and by ∂f the subdifferential of a convex function f (see Ekeland and Temam [7]). If we define the sets S_1 and S_2 by

$$
\begin{aligned}
S_1 &= \{y \in C(\cup_{i=1}^{N_Z}\bar{A}_i\times[0,T]):\ y(x,t) \le \sigma_j \quad \forall (x,t)\in \overline{A}_j\times[0,T],\ j=1,\dots,N_Z\},\\
S_2 &= \{\omega \in C(\cup_{i=1}^{N_Z}\bar{A}_i\times[0,T]):\ \omega(x,t) \ge \zeta_j \quad \forall (x,t)\in \overline{A}_j\times[0,T],\ j=1,\dots,N_Z\},
\end{aligned}
$$

we have the following result.

THEOREM 5.2. *Let $m \in U_{ad}$ be an optimal control. Then, there exist two functions $\rho_1,\ \rho_2 \in L^r(0,T;W^{1,s}(\Omega)) \cap L^2(0,T;L^2(\Omega))\ \forall\ r,\ s \in [1,2),\ \frac{2}{r}+\frac{2}{s} > 3$, solving (2.1) and two functions $p_1,\ p_2 \in L^r(0,T;W^{1,s}(\Omega))$ solving (5.1), where $\mu_1,\ \mu_2$ are two Borel measures, with support in $\cup_{i=1}^{N_Z}\bar{A}_i\times[0,T]$, such that*

$$
\mu_{i_{|\cup_{i=1}^{N_Z}\bar{A}_i\times[0,T]}} \in \partial I_{S_i}(F_i(m)), \quad i=1,2, \tag{5.2}
$$

and the following relation is satisfied:

$$
\begin{aligned}
&\sum_{j=1}^{N_E}\Big\{\int_0^T f_j{}'(m_j(t))(n_j(t)-m_j(t))\,dt\\
&+\int_0^T \frac{1}{h(P_j,t)}p_1(P_j,t)(n_j(t)-m_j(t))\,dt\Big\} \ge 0 \quad \forall n \in U_{ad}.
\end{aligned}
\tag{5.3}
$$

Proof. Let m be a solution of the minimization problem:

$$\min_{n\in U_{ad}} J(n) + I_{S_1}(F_1(n)) + I_{S_2}(F_2(n)).$$

Since $U_{ad} \subset L^\infty(0,T)$ is a convex set and the functions F_1 and F_2 are affine and differentiable in m we obtain (see Ekeland and Temam [7]) the existence of two measures $\tilde{\mu}_1$, $\tilde{\mu}_2$, satisfying (5.2), such that

$$\begin{aligned} DJ(m)(n-m) + \langle (DF_1(m))^*(\tilde{\mu}_1), n-m\rangle \\ +\langle (DF_2(m))^*(\tilde{\mu}_2), n-m\rangle \geq 0 \quad \forall n \in U_{ad}, \end{aligned}$$

where

$$\begin{aligned} DJ(m)(n-m) &= \sum_{j=1}^{N_E} \int_0^T f_j{}'(m_j(t))(n_j(t)-m_j(t))\,dt, \\ \langle (DF_1(m))^*(\tilde{\mu}_1), n-m\rangle &= \langle \tilde{\mu}_1, DF_1(m)(n-m)\rangle, \\ \langle (DF_2(m))^*(\tilde{\mu}_2), n-m\rangle &= \langle \tilde{\mu}_2, DF_2(m)(n-m)\rangle. \end{aligned}$$

Then, we define $\mu_j = \tilde{\mu}_j\, \chi_{\cup_{i=1}^{N_Z} \bar{A}_i \times [0,T]}$, $j = 1,2$. We also introduce $p = (p_1, p_2) \in [L^r(0,T; W^{1,s}(\Omega))]^2$ for r, $s \in [1,2)$ with $\frac{2}{r} + \frac{2}{s} > 3$ such that it is solution of (5.1) and satisfies (cf. Casas [4])

$$\begin{aligned} &\int_0^T \left\langle p, \frac{\partial z}{\partial t} + \mathcal{A}(z) \right\rangle dt \\ &\quad = \int_{\cup_{i=1}^{N_Z} \bar{A}_i \times [0,T]} z_1\, d\mu_1(x,t) + \int_{\cup_{i=1}^{N_Z} \bar{A}_i \times [0,T]} z_2\, d\mu_2(x,t) \quad \forall z \in \mathcal{R}, \end{aligned}$$

where

$$\begin{aligned} \mathcal{R} = \Bigg\{ & z = (z_1, z_2) \in [L^2(0,T; H^1(\Omega)) \cap C(\bar{\Omega} \times [0,T])]^2 \ : \\ & \frac{\partial z}{\partial t} + \mathcal{A}(z) \in L^\infty(\Omega \times (0,T)),\ \frac{\partial z}{\partial n_{\mathcal{A}}}_{|\Gamma \times (0,T)} = 0,\ z(.,0) = 0 \Bigg\}. \end{aligned}$$

Finally, it can be shown that

$$\begin{aligned} &\langle \mu_1, DF_1(m)(n-m)\rangle + \langle \mu_2, DF_2(m)(n-m)\rangle \\ &\quad = \sum_{j=1}^{N_E} \int_0^T \frac{1}{h(P_j,t)} p_1(P_j,t)(n_j(t)-m_j(t))\,dt. \end{aligned}$$

In effect, let $\omega = (\omega_1, \omega_2)$ be the solution of the linearized problem (3.8). Then, as was shown in Lemma 3.5,

$$\begin{aligned} DF_1(m)(n-m) &= \omega_{1\,|\cup_{i=1}^{N_Z} \bar{A}_i \times [0,T]}, \\ DF_2(m)(n-m) &= \omega_{2\,|\cup_{i=1}^{N_Z} \bar{A}_i \times [0,T]}. \end{aligned}$$

Let $\{\theta^k\}$ be a mollifier sequence for the Dirac measure $\delta(x-0)$ with support in $B(0,\frac{1}{k})$ and let $\omega^k=(\omega_1^k,\omega_2^k)$ be the solution of the system

$$\begin{aligned}
\frac{\partial\omega_1^k}{\partial t}+\vec{u}\nabla\omega_1^k-\beta_1\Delta\omega_1^k=-\kappa_1\omega_1^k+\frac{1}{h}\sum_{j=1}^{N_E}(n_j-m_j)\theta^k(x-P_j) \qquad & \text{in } \Omega\times(0,T),\\
\frac{\partial\omega_1^k}{\partial n}=0 \qquad & \text{on } \Gamma\times(0,T),\\
\omega_1^k(x,0)=0 \qquad & \text{in } \Omega,
\end{aligned}$$

$$\begin{aligned}
\frac{\partial\omega_2^k}{\partial t}+\vec{u}\nabla\omega_2^k-\beta_2\Delta\omega_2^k=-\kappa_1\omega_1^k-\frac{1}{h}\kappa_2\omega_2^k \qquad & \text{in } \Omega\times(0,T),\\
\frac{\partial\omega_2^k}{\partial n}=0 \qquad & \text{on } \Gamma\times(0,T),\\
\omega_2^k(x,0)=0 \qquad & \text{in } \Omega.
\end{aligned}$$

From $\|\theta^k\|_{L^1(\Omega)}=1$ we obtain (see Theorem 3.2) that ω_1^k and ω_2^k are bounded in $L^r(0,T;W^{1,s}(\Omega))$ and $L^2(0,T;L^2(\Omega))$. Consequently,

$$\begin{aligned}
\omega_1^k &\longrightarrow \eta_1 \qquad \text{weakly in } L^r(0,T;W^{1,s}(\Omega)),\\
\omega_2^k &\longrightarrow \eta_2 \qquad \text{weakly in } L^r(0,T;W^{1,s}(\Omega)),
\end{aligned}$$

and, since θ^k is weak* convergent to $\delta(x-0)$, we deduce

$$\eta_1=\omega_1, \qquad \eta_2=\omega_2.$$

Due to the boundedness of ω_j^k, $j=1,2$, in $L^2(0,T;L^2(\Omega))$, we deduce that the sequences $\omega_{j\,|\cup_{i=1}^{N_Z}\bar{A}_i\times[0,T]}^k$, are bounded in $C^{\alpha,\alpha/2}(\cup_{i=1}^{N_Z}\bar{A}_i\times[0,T])$, for some $\alpha\in(0,1)$. Then, by the compact imbedding in $C(\cup_{i=1}^{N_Z}\bar{A}_i\times[0,T])$, we obtain the existence of subsequences uniformly converging to ω_j.

By the definition of p we have that

$$\begin{aligned}
&\int_0^T\left\langle p,\frac{\partial\omega^k}{\partial t}+\mathcal{A}(\omega^k)\right\rangle dt\\
&\quad=\int_{\cup_{i=1}^{N_Z}\bar{A}_i\times[0,T]}\omega_1^k\,d\mu_1(x,t)+\int_{\cup_{i=1}^{N_Z}\bar{A}_i\times[0,T]}\omega_2^k\,d\mu_2(x,t)\\
&\quad=\langle\mu_1,\omega_1^k\rangle+\langle\mu_2,\omega_2^k\rangle.
\end{aligned}$$

Thus, by the characterization of ω^k:

$$\begin{aligned}
&\langle\mu_1,\omega_1^k\rangle+\langle\mu_2,\omega_2^k\rangle\\
&\quad=\sum_{j=1}^{N_E}\int_0^T\int_\Omega\frac{1}{h(x,t)}p_1(x,t)(n_j(t)-m_j(t))\theta^k(x-P_j)\,dx\,dt\\
&\quad=\sum_{j=1}^{N_E}\int_0^T\int_{B(P_j,\frac{1}{k})}\frac{1}{h(x,t)}p_1(x,t)(n_j(t)-m_j(t))\theta^k(x-P_j)\,dx\,dt.
\end{aligned}$$

So, if we pass to the limit, taking into account that ω^k is smooth inside $\cup_{i=1}^{N_Z}\bar{A}_i\times[0,T]$ and that p_1 is smooth outside $\cup_{i=1}^{N_Z}\bar{A}_i\times[0,T]$, we obtain that

$$\langle\mu_1,\omega_1\rangle+\langle\mu_2,\omega_2\rangle=\sum_{j=1}^{N_E}\int_0^T\frac{1}{h(P_j,t)}p_1(P_j,t)(n_j(t)-m_j(t))\,dt,$$

which concludes the proof. ∎

6. The discretized problem. The first step in the numerical resolution of the problem is solving the state system. In order to do it, we carry out a time discretization, treating the convective term with a method of characteristics. For the time interval $[0,T]$ we choose an integer number N and we define $\Delta t = \frac{T}{N} > 0$ and $t_n = n\Delta t \ \forall n = 0,\dots,N$. Then, for the semidiscretized problem, we consider a variational formulation and approximate it by a space discretization with finite elements. We approximate Ω by the polygonal set Ω_h and choose an admissible triangulation τ_h of it (see [5]) with triangles of diameter $\le h$ and such that the vertices on the boundary of Ω_h also lie on the boundary of Ω. We define V_h as the following finite element space:

$$V_h = \left\{ v_h \in C^0(\bar{\Omega}),\ v_{h|K} \in P_1,\ K \in \tau_h \right\}.$$

The velocity and height fields, necessary for the discretized problem, are obtained by solving the Saint-Venant equations [3]. So, the resultant discretized problem is equivalent to the linear system

$$\left\{ \begin{array}{ll} M_{1h}\rho_{1h}^{n+1} = b_{1h}^n & \\ & \forall\ n = 0,\dots,N-1, \\ M_{2h}\rho_{2h}^{n+1} = b_{2h}^n & \end{array} \right. \tag{6.1}$$

for ρ_{1h}^0 and ρ_{2h}^0 given, and with

- $\rho_{1h}^{n+1} = \left(\rho_{1h}^{n+1}(x_1),\dots,\rho_{1h}^{n+1}(x_{N_v})\right)^t$,
- $(M_{1h})_{ij} = \beta_1 \int_\Omega \nabla\tilde{v}_i \nabla\tilde{v}_j dx + \left(\frac{1}{\Delta t} + \kappa_1\right)\int_\Omega \tilde{v}_i\tilde{v}_j dx$,
- $(b_{1h}^n)_l = \frac{1}{\Delta t}\int_\Omega (\rho_{1h}^n\ X_h^n)\tilde{v}_l dx + \int_\Omega \frac{1}{h^n}\sum_{j=1}^{N_E} m_j(t_n)\delta(x-P_j)\tilde{v}_l dx$,
- $\rho_{2h}^{n+1} = \left(\rho_{2h}^{n+1}(x_1),\dots,\rho_{2h}^{n+1}(x_{N_v})\right)^t$,
- $(M_{2h})_{ij} = \beta_2 \int_\Omega \nabla\tilde{v}_i \nabla\tilde{v}_j dx + \frac{1}{\Delta t}\int_\Omega \tilde{v}_i\tilde{v}_j dx$,
- $(b_{2h}^n)_l = -\kappa_2 \int_\Omega \frac{1}{h^n}\rho_{2h}^n\tilde{v}_l dx + \frac{1}{\Delta t}\int_\Omega (\rho_{2h}^n\ X_h^n)\tilde{v}_l dx - \kappa_1\int_\Omega \rho_{1h}^{n+1}\tilde{v}_l dx + \kappa_2\int_\Omega \frac{1}{h^n} d_s\tilde{v}_l dx$,

where $\{\tilde{v}_1,\tilde{v}_2,\dots,\tilde{v}_{N_v}\}$ is a basis of V_h such that $\tilde{v}_i(x_j) = \delta_{ij}$, $h^n(x) = h(x,t_n)$, and $X^n(x)$ is the position at instant t_n of the particle that will be at point x in the instant t_{n+1} (see [21] for further details).

The discretized cost $\hat{J}$ and the discretized constraints g are given by

$$\begin{array}{rccl} \hat{J}: & R^{N\times N_E} & \longrightarrow & R, \\ & m & \longrightarrow & \hat{J}(m) = \Delta t \displaystyle\sum_{j=1}^{N_E}\sum_{n=0}^{N-1} C_{jn} f_j(m_{jn}), \end{array}$$

$$\begin{array}{rccl} g: & R^{N\times N_E} & \longrightarrow & R^{N\times N_{VZ}} \times R^{N\times N_{VZ}} \times R^{N\times N_E}, \\ & m & \longrightarrow & g(m) = (\underbrace{\tilde{\rho}_{1h} - \sigma,\ \zeta - \tilde{\rho}_{2h}}_{= g_1(m)},\ \underbrace{\underline{m} - m}_{= g_2(m)})^t, \end{array}$$

where

- m is the vector consisting of all of the discharges at all times,
- m_{jn} is the amount of BOD discharged in P_j at time t_n,
- C_{jn} are the weights of the quadrature formula,
- N_{VZ} is the number of vertices in the protected areas,
- $\tilde{\rho}_{ih}$ is a vector of values of ρ_{ih} at each vertex included in the protected areas and for all times.

We remark that the function g can be decomposed into g_1, which is putting together the constraints about the water quality, and g_2, which collects the control constraints.

Then, the optimal control problem is approached by the following discretized problem:

$$(\mathcal{P}_{\mathcal{F}}) \quad \begin{cases} \min\limits_{m \in R^{N \times N_E}} & \hat{J}(m) \\ \text{such that} & g(m) \leq 0. \end{cases}$$

If the constraints set is nonempty, then this problem has a solution because this set is compact and the cost function is convex.

7. Numerical resolution by a feasible points algorithm. We solve the discretized control problem $(\mathcal{P}_{\mathcal{F}})$ by using a feasible points method which is based on a globally convergent algorithm introduced by Herskovits [9, 10] and Panier, Tits, and Herskovits [17].

We use the following notation:

- p is the dimension of the control and q is the number of the constraints on the state,
- (λ, θ) is the vector of the dual variables,
- $L(m, \lambda, \theta)$ is the Lagrangian,
- $H(m, \lambda, \theta)$ is the Hessian.

So, the first order Karush–Kuhn–Tucker optimality conditions for the discretized problem can be written as follows:

$$\nabla \hat{J}(m) + \nabla g_1 \lambda - I\theta = 0, \tag{7.1}$$

$$G_1(m)\lambda = 0, \qquad G_2(m)\theta = 0, \tag{7.2}$$

$$\lambda \geq 0, \qquad \theta \geq 0, \tag{7.3}$$

$$g_1(m) \leq 0, \qquad g_2(m) \leq 0, \tag{7.4}$$

where $G_1(m)$ and $G_2(m)$ are diagonal matrices, of order q and p, respectively, and with diagonal elements being the values of the corresponding functions $g_i(m)$.

The basic idea of the feasible points algorithm consists of solving the system of (7.1)–(7.2) in (m, λ, θ) by using a fixed point method, in such a way that the conditions (7.3)–(7.4) hold at each iteration.

Thus, for a given point $(m^k, \lambda^k, \theta^k)^t$, the Newton's method applied to the previous system computes the next iteration $(m_0^{k+1}, \lambda_0^{k+1}, \theta_0^{k+1})^t$ by solving

$$\begin{pmatrix} m_0^{k+1} \\ \lambda_0^{k+1} \\ \theta_0^{k+1} \end{pmatrix} = \begin{pmatrix} m^k \\ \lambda^k \\ \theta^k \end{pmatrix} - \begin{pmatrix} H(m^k, \lambda^k, \theta^k) & \nabla g_1 & -I \\ \Lambda^k (\nabla g_1)^t & G_1^k & 0 \\ -\Theta^k & 0 & G_2^k \end{pmatrix}^{-1} \begin{pmatrix} \nabla \hat{J}(m^k) + \nabla g_1 \lambda^k - I\theta^k \\ G_1^k \lambda^k \\ G_2^k \theta^k \end{pmatrix},$$

where Λ^k and Θ^k are diagonal matrices whose elements are the coordinates of λ^k and θ^k, respectively.

Obviously, conditions (7.3) and (7.4) do not usually hold at this new point $(m_0^{k+1}, \lambda_0^{k+1}, \theta_0^{k+1})^t$. Then, we define d^k as a search direction in m and rewrite the previous equality by computing $(d^k, \lambda_0^{k+1}, \theta_0^{k+1})^t$ as the solution of the following linear system:

$$\begin{pmatrix} H(m^k, \lambda^k, \theta^k) & \nabla g_1 & -I \\ \Lambda^k(\nabla g_1)^t & G_1^k & 0 \\ -\Theta^k & 0 & G_2^k \end{pmatrix} \begin{pmatrix} d^k \\ \lambda_0^{k+1} \\ \theta_0^{k+1} \end{pmatrix} = \begin{pmatrix} -\nabla \hat{J}(m^k) \\ 0 \\ 0 \end{pmatrix}. \tag{7.5}$$

Now, in order to determine the new primal point m^{k+1}, we perform a line search along d^k (by using an extension of Armijo's rule [11]) for obtaining a step t^k which leads us to a new point where the cost reduction is satisfactory.

Finally, the new value of the dual variable $(\lambda^{k+1}, \theta^{k+1})^t$ can be computed by several updating methods. We use the following one, based upon an idea from Herskovits [9]:

(1) We choose positive numbers ξ_1, ξ_2, μ_1, μ_2, λ^I, θ^I.

(2) For $i = 1, \ldots, q$, and for $j = 1, \ldots, p$, we define

$$(\lambda^{k+1})_i = \sup\{(\lambda_0^{k+1})_i, \xi_1 \|d^k\|^2\},$$

$$(\theta^{k+1})_j = \sup\{(\theta_0^{k+1})_j, \xi_2 \|d^k\|^2\}.$$

(3) If $\left(g_1(m^{k+1})\right)_i \geq -\mu_1$ and $(\lambda^{k+1})_i < \lambda^I$, then $(\lambda^{k+1})_i = \lambda^I$.

(4) If $\left(g_2(m^{k+1})\right)_j \geq -\mu_2$ and $(\theta^{k+1})_j < \theta^I$, then $(\theta^{k+1})_j = \theta^I$.

In this algorithm the role played by the resolution of the linear system (7.5) must be noted. Since all the constraints are linear we have $H^k = H(m^k, \lambda^k, \theta^k) = \nabla^2 \hat{J}(m^k)$, which, due to the particular form of $\hat{J}$, is diagonal and easy to compute. On the other hand, the convexity of f_j, $j = 1, \ldots, N_E$, ensures that H^k is positive definite and, consequently (see [17]), the matrix of the linear system is nonsingular if the components of λ^k and θ^k related to active constraints are strictly positive.

Thus, the nonsingular matrix, of order $(2p + q)$, has diagonal blocks, except $\nabla g_1 \in \mathcal{M}_{p \times q}$ and $\Lambda^k(\nabla g_1)^t \in \mathcal{M}_{q \times p}$. If the values of p and q are not too large, the system (7.5) can be easily solved, for instance, by a preconditioned biconjugate method.

However, in a realistic problem where p and q are very large, the need of preconditioning each iteration causes this method to become very slow. In this case the followed strategy consists of solving the linear system by blocks, locating previously the active constraints in order to avoid ill-conditioning. The resultant system is full and nonsymmetric, but only of order ($p +$ number of active constraints), and can be then solved, for instance, by the QR method.

8. Numerical results. Multiple tests have been developed, solving the previous problem for several *rías* in Galicia, Spain, with different initial values in the optimization algorithm, and the achieved results have always been satisfactory.

In this section we present the numerical results obtained when solving the problem on a realistic situation: We have taken a two-dimensional mesh of the *ría* of Vigo as

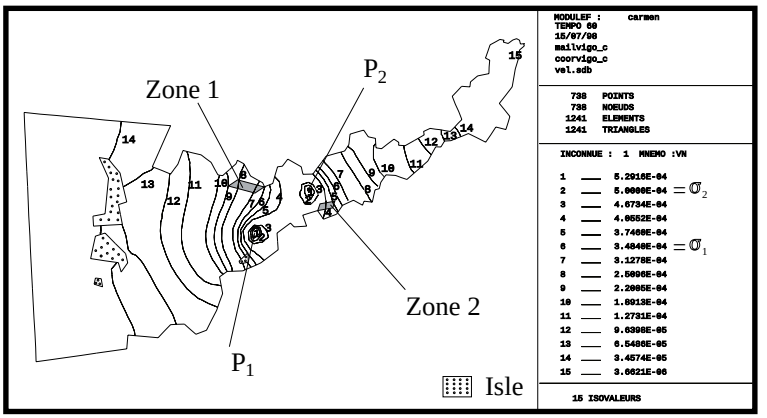


FIG. 8.1. *BOD concentration at high tide.*

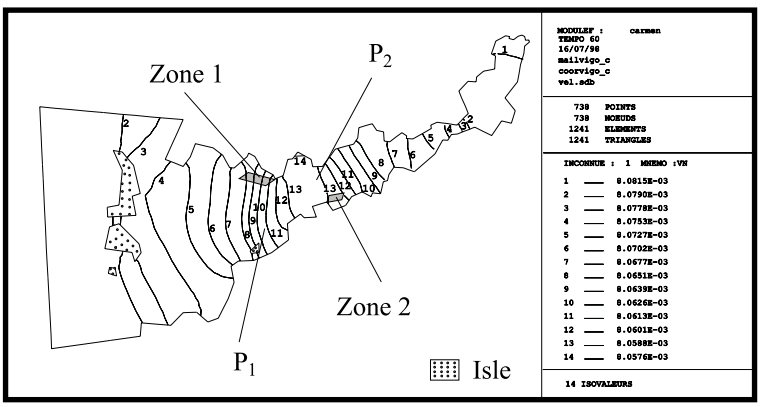


FIG. 8.2. *DO concentration at high tide.*

domain Ω, where we have considered two protected areas and two points of discharge (see Figures 8.1 and 8.2). We have also supposed that it is necessary to guarantee lower levels of pollution in zone 1 than in zone 2. (The values of the parameters can be seen in Table 8.1.)

The cost function is given in Figure 8.3: If we assume pollutant concentration of the sewage arriving to the sewage farm is 150 $\mathrm{Kg/m^3}$, then the depuration cost above this value is constant. The velocity and the height of the water have been obtained to solve the Saint-Venant equations on this domain. The numerical resolution of the Saint-Venant equations has been carried out in [3].

Figures 8.1 and 8.4 show the BOD concentration at high tide and at low tide, respectively. The constraints occur everywhere in the protected areas; at high tide, they saturate at one of the vertices in zone 1, but at low tide, after a tidal cycle, the saturation takes place at one of the vertices in zone 2.

The values of the optimal discharges, which produce this situation, can be seen in Figure 8.5. The discharge rate is greater during rising tide at point 2 than at point 1. However, during ebb tide (after $t = 60$) the flow rate decreases at P_2 and increases

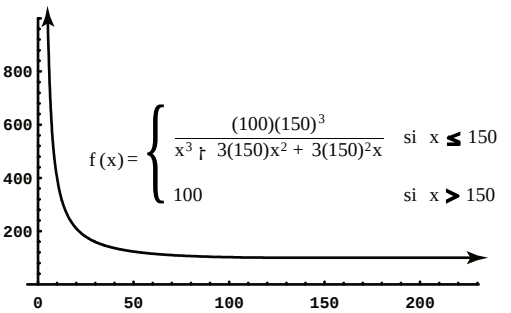


FIG. 8.3. *Cost function.*

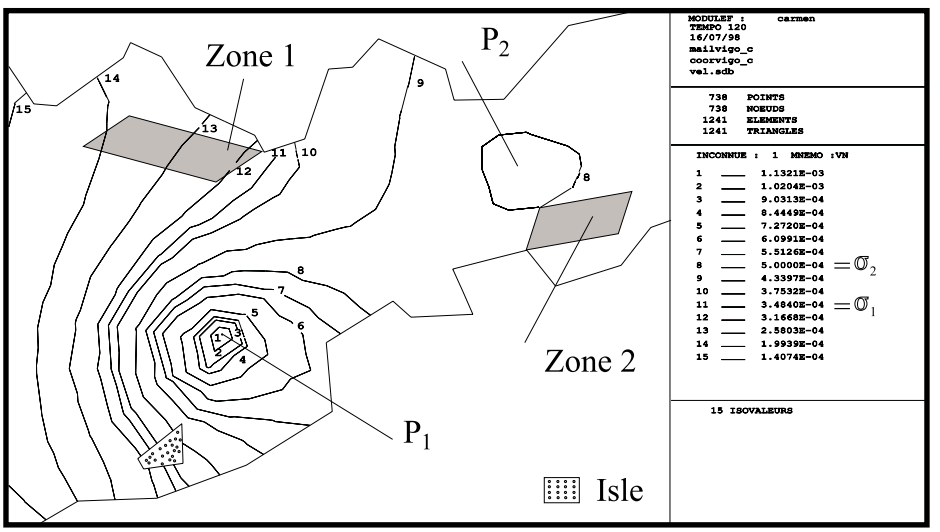


FIG. 8.4. *BOD concentration at low tide.*

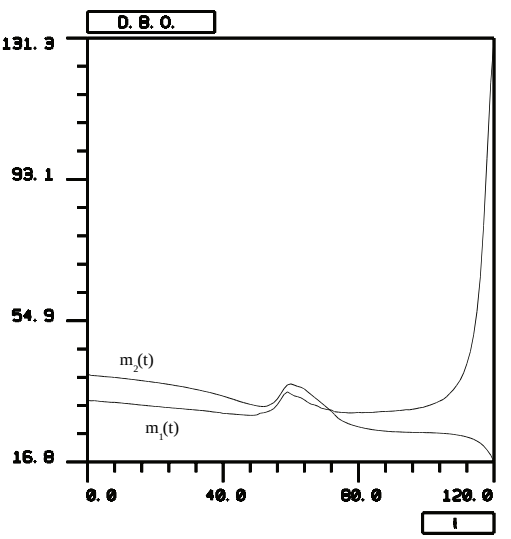


FIG. 8.5. *Optimal discharges during a tidal cycle.*

TABLE 8.1
Parameters for solving the problem ($\mathcal{P}_\mathcal{F}$) on the ría of Vigo.

Saint-Venant parameters	Empiric coefficients	Adimensional constants
Tidal cycle: T=12.4 h	$\beta_1 = \beta_2 = 2000$ m^2/s	$N = 120$
Tidal run: 2.8 m	$\kappa_1 = 1.15\,10^{-5}$ s^{-1}	$\xi_1 = 10^{-8}$
Water density: 1000 Kg/m^3	$\kappa_2 = 9\,10^{-12}$ s^{-1}	$\xi_2 = 10^{-8}$
Air density: 1.28 Kg/m^3	$d_s = 8.98\,10^{-3}$ Kg/m^3	$\mu_1 = 10^{-5}$
North latitude: 0.7326 rad	$\rho_{20} = 8.082\,10^{-3}$ Kg/m^3	$\mu_2 = 10^{-5}$
Wind direction: 3.9269 rad	$\sigma_1 = 3.48398\,10^{-4}$ Kg/m^3	$\lambda^I = 1$
Wind velocity: 10 Km/h	$\sigma_2 = 5\,10^{-4}$ Kg/m^3	$\theta^I = 1$
Angular velocity of Earth:	$\zeta_1 = 8.05255\,10^{-3}$ Kg/m^3	
7.92 10^{-5} rad/s	$\zeta_2 = 8.03218\,10^{-3}$ Kg/m^3	
	$\underline{m} = 0$ Kg/m^3	
	$\overline{m} = 150$ Kg/m^3	

at P_1. This is an obvious consequence of the outfalls position. In fact, during rising tide, P_2 is *better* than P_1, but during ebb tide P_1 is *the best* of them.

This work is being developed in the framework of a project sponsored by the government of Galicia, Spain, the final aim of which is the salubrity of the ecosystem formed by Galician *rías*. This is a very important purpose, from both an ecological and an economical point of view, because of its influence on shellfish production and touristic resources, and the local administration will provide data for validation of the models and optimization methods presented here.

Acknowledgments. The authors are grateful to Profs. L.J. Alvarez-Vázquez, A. Bermúdez and R. Muñoz-Sola for helpful discussions. The authors also thank the referees for their suggestions.

REFERENCES

[1] A. BERMÚDEZ, *Mathematical techniques for some environmental problems related to water pollution control*, in Mathematics, Climate and Environment, J. I. Díaz and J. L. Lions, eds., Masson, Paris, 1993.

[2] A. BERMÚDEZ, A. MARTÍNEZ, AND C. RODRÍGUEZ, *Un probleme de controle ponctuel lie a l'emplacement optimal d'emissaires d'evacuation sous-marine*, C. R. Acad. Sci. Paris Ser. I Math., 313 (1991), pp. 515–518.

[3] A. BERMÚDEZ, C. RODRÍGUEZ, AND M. A. VILAR, *Solving shallow water equations by a mixed implicit finite element method*, IMA J. Numer. Anal., 11 (1991), pp. 79–97.

[4] E. CASAS, *Pontryagin's principle for state-constrained boundary control problems of semilinear parabolic equations*, SIAM J. Control Optim., 35 (1997), pp. 1297–1327.

[5] P. G. CIARLET, *Basic error estimates for elliptic problems*, in Handbook of Numerical Analysis, Vol. II, P. G. Ciarlet and J. L. Lions, eds., North-Holland, Amsterdam, 1991.

[6] F. DI BENEDETTO, *On the local behaviour of solutions of degenerate parabolic equations with measurable coefficients*, Ann. Scuola Norm. Sup. Pisa Cl. Sci., 13 (1986), pp. 487–535.

[7] I. EKELAND AND R. TEMAM, *Convex Analysis and Variational Problems*, North-Holland, Amsterdam, 1976.

[8] H. O. FATTORINI AND S. S. SRITHARAN, *Optimal control problems with state constraints in fluid mechanics and combustion*, Appl. Math. Optim., 38 (1998), pp. 159–192.

[9] J. HERSKOVITS, *A two stage feasible directions algorithm for nonlinear constrained optimization*, Math. Programming, 36 (1986), pp. 19–38.

[10] J. HERSKOVITS, *A feasible directions interior point technique for nonlinear optimization*, J. Optim. Theory Appl., 99 (1998), pp. 121–146.

[11] J. B. HIRIART-URRUTY AND C. LEMARECHAL, *Convex Analysis and Minimization Algorithms*, Springer-Verlag, Berlin, 1993.

[12] B. HU AND J. YONG, *Pontryagin maximum principle for semilinear and quasilinear parabolic*

equations with pointwise state constraints, SIAM J. Control Optim., 33 (1995), pp. 1857–1880.

[13] J. L. Lions, *Pointwise control for distributed systems*, in Control and Estimation in Distributed Parameter Systems, H. T. Banks, ed., SIAM, Philadelphia, 1992.

[14] J .L. Lions and E. Magenes, *Problemes aux Limites non Homogenes et Applications*, Dunod, Paris, 1968.

[15] O. A. Ladyzhenskaja, V. A. Solonnikov, and N. N. Uraltseva, *Linear and Quasilinear Equations of Parabolic Type*, Transl. Math. Monogr. 23, Amer. Math. Soc., Providence, RI, 1968.

[16] G. I. Marchuk, *Mathematical Models in Environmental Problems*, North-Holland, Amsterdam, 1986.

[17] E. R. Panier, A. L. Tits, and J. Herskovits, *A QP-free, globally convergent, locally superlinearly convergent algorithm for inequality constrained optimization*, SIAM J. Control Optim., 26 (1988), pp. 788–811.

[18] J. P. Raymond and H. Zidani, *Pontryagin's principle for state-constrained control problems governed by parabolic equations with unbounded controls*, SIAM J. Control Optim., 36 (1998), pp. 1853–1879.

[19] C. Saguez, *Controle Ponctuel et Controle en Nombres Entieres de Systemes Distribués*, Ph.D. thesis, University of Paris VI, 1974.

[20] J. Simon, *Controle de la solution d'une equation parabolique avec donnée ponctuelle*, J. Systems Sci. Math. Sci., 3 (1983), pp. 1–27.

[21] M. E. Vázquez-Méndez, *Contribución a la resolución numérica de modelos para el estudio de la contaminación de aguas*, Master thesis, Departamento de Matemática Aplicada, Universidade de Santiago de Compostela, Spain, 1992.

SIAM J. CONTROL OPTIM.
Vol. 38, No. 5, pp. 1554–1580

BASIC CONTROL FOR THE VISCOUS MOORE–GREITZER PARTIAL DIFFERENTIAL EQUATION*

BJÖRN BIRNIR† AND HÖSKULDUR ARI HAUKSSON†

Abstract. The notions of basic controllability and basic control are defined. A quadratic optimal control of the linearized viscous Moore–Greitzer equation is presented, and it is confirmed that stall is uncontrollable in this model. A basic control is constructed for the nonlinear viscous Moore–Greitzer equation which can control both surge and stall. Some extensions of this construction are discussed. Numerical simulations of the basic control are presented, and its performance is compared to the performance of a backstepping control constructed by Banaszuk, Hauksson, and Mezić [*SIAM J. Control Optim.*, 37 (1999), pp. 1503–1537]. It is shown that the viscous Moore–Greizer equation with throttle control is not basically controllable, but under certain conditions, adding air injection control will make the equation basically controllable.

Key words. optimal control of PDE, Moore–Greitzer equation, basic control, basic attractor, global attractor

AMS subject classifications. 49K20, 93B05, 93B27, 35K40, 35K55

PII. S0363012998345184

1. Introduction.

In recent years a lot of attention has been devoted to the study of air flow through turbomachines. The main reason for this interest is that when a turbomachine, such as a jet engine, operates close to its optimal operating parameter values, the flow can become unstable. These instabilities put a large stress on the engine, and in some cases the engine needs to be turned off in order to recover original operation. For this reason jet engines are currently operated away from their optimal operating parameter values.

A jet engine can be thought of as a compressor, where the incoming air is compressed by alternating rings of rotating blades and stationary blades. The mixture of fuel and compressed air is then ignited, and the resulting combustion generates thrust that propels the aircraft. There are primarily two types of instabilities that occur in the flow through the compressor. They are called *surge* and *stall*. Surge is characterized by large oscillations of the mean mass flow through the engine. During part of the cycle, the mean mass flow may become reversed, thrusting air out the front end of the engine. This puts a large stress on the components of the engine and seriously impairs its performance. When stall occurs, there are regions of relatively low air flow that form at isolated locations around the rim of the compressor. Here too, the phenomenon can be so pronounced that the flow in these isolated regions is reversed. Again, this causes a large stress on the components of the engine and reduces its performance.

Moore and Greitzer published in 1986 a PDE model for turbomachines which has been very successful [26]. A substantial amount of work has been done on finite Galerkin approximations of that model since (see, e.g., [23], [20], and references

*Received by the editors September 28, 1998; accepted for publication (in revised form) November 23, 1999; published electronically May 26, 2000. This research was funded by NSF under grant number DMS-9704874 and in part by AFOSR under grant number F49620-95-1-0409. Equipment was funded by NSF grants DMS-9628606 and PHY-9601954.

http://www.siam.org/journals/sicon/38-5/34518.html

†Department of Mathematics, University of California Santa Barbara, Santa Barbara, CA 93106 and Science Institute, University of Iceland, 3 Dunhaga, IS-107 Reykjavik, Iceland (birnir@math.ucsb.edu, URL: www.math.ucsb/~birnir, hauksson@math.ucsb.edu).

therein). Banaszuk, Hauksson, and Mezić [2] considered the full PDE model of Moore and Greitzer.

Currently Mezić [24] has derived a model of the three dimensional flow in jet engine compressors. His model reduces to that presented in [6] when one assumes that the dependence of the flow on the radial direction is negligible. The viscous term in that equation, first introduced by Adomaitis and Abed [1], has, however, a new and better interpretation in Mezić's treatment. The term is not due to the viscosity of the air, but rather, it is a diffusion term due to the inviscid process of turbulent momentum transport via Reynolds stresses. The difference is several orders of magnitude of the constant ν, which now represents the eddy viscosity. It is this model with the additional assumption that the flow is independent of the radial coordinate that we use here. In this form the model is known as the viscous Moore–Greitzer equation (vMG equation).

Birnir and Hauksson [6] proved that the vMG equation is well posed in the Hilbert space $X = \bar{H}^1 \times \mathbf{R}^2$, where $\bar{H}^1$ denotes the Sobolev space with index one of functions on the unit circle with square integrable first derivative and zero mean. This solution is smooth in space and time variables, and this dynamical system has a global attractor with finite Hausdorff and fractal dimensions. In [7] the authors analyzed the basic attractor and found explicit solutions for stall for certain parameter values and showed that they are stable and persist under small perturbations of the parameters. Stall is a solitary wave that rotates around the annulus at half the rotor speed of the engine. They conclude that the basic attractor consists of design flow, surge, and one or more stall solutions. The analysis of the basic attractor (see subsection 1.1) was extended for all parameter values in [8], and there and in Hauksson [16] they derived a reduced order model that captures the dynamics of the vMG equation quantitatively as well as qualitatively. These results are in good agreement with experimental [12] and numerical results [14].

There have been numerous control related results reported for Galerkin truncations of the Moore–Greitzer equation. The first control result for the full system, and to our knowledge the only one until now, was presented by Banaszuk, Hauksson, and Mezić [2]. They constructed a feedback throttle control which made design flow attract all of the state space for the inviscid Moore–Greitzer equation. This equation is a hyperbolic equation and the nature of its solutions are quite different than that for the vMG equation. Current results by Mezić [24] indicate that the latter is a better physical model for the flow through the jet engine compression system. The controller obtained by Banaszuk, Hauksson, and Mezić is not very cost effective. In particular, it over-reacts to small-amplitude high-frequency disturbances which are naturally damped in the viscous model.

The backstepping control given by Banaszuk, Hauksson, and Mezić shows that one can eliminate stall and surge by using throttle control. The question we want to answer in this paper is how simple can we make the control design, and how efficient can the control be? The control philosophy we want to adopt is to construct a control strategy that can recover design flow operation after large disturbances, but this strategy is not necessarily good for regulating the design flow. For that a different strategy would be used.

This paper is arranged as follows. In the immediate subsection we present the equations of motion and assumptions. This section also explains the dynamics of the model briefly. Section 3 defines basic controllability and attractor controllability. Section 4 considers the simplest possibility, which is to linearize the system about the design flow and then apply the standard optimal control theory to obtain an

optimal control subject to quadratic cost. Here we confirm that stall is uncontrollable in the linearized model. Section 5 considers first the case when the throttle is moved adiabatically. We then use our knowledge of the basic attractor of the system to construct a basic control which recovers design flow operation from stall or surge. Numerical simulations are presented in section 6. Here we compare the basic control with the backstepping control. In section 7 we prove that the vMG equation with throttle control is not basically controllable. Moreover, if one in addition has air injection at one's disposal, the vMG equation is basically controllable.

1.1. The equation of motion and assumptions. Currently Mezić [24] has derived a model of the three dimensional flow through the compression system in jet engines. When one assumes that the flow does not depend on the radial direction, the equations reduce to the following:

$$\frac{\partial}{\partial t}\varphi = \nu\frac{\partial^2\varphi}{\partial\theta^2} - \frac{1}{2}\frac{\partial\varphi}{\partial\theta} + \psi_c(\Phi+\varphi) - \overline{\psi_c}, \quad \theta\in[0,2\pi], \tag{1.1}$$

$$\dot{\Phi} = \frac{1}{l_c}(\overline{\psi_c} - \Psi), \tag{1.2}$$

$$\dot{\Psi} = \frac{1}{4l_cB^2}(\Phi - \gamma F_T^{-1}(\Psi)), \tag{1.3}$$

where

$$\overline{\psi_c} := \frac{1}{2\pi}\int_0^{2\pi}\psi_c(\Phi(t)+\varphi(t,\theta))d\theta. \tag{1.4}$$

In this guise the equations are a special case of the original equations presented by Moore and Greitzer [26], except for an additional term due to the eddy viscosity. This term was first suggested by Adomaitis and Abed [1], but without the justification later provided by Mezić. The equation is known as the vMG equation. Here the dot represents the total derivative with respect to time.

The characteristic ψ_c is a cubic polynomial with a negative leading coefficient, and F_T^{-1} is a smooth function which is equal to $F_T^{-1}(\Psi) = \Psi|\Psi|^{-1/2}$ outside a small neighborhood of the origin.

In what follows, we will allow γ to depend on the state, but we will assume that it does so in a smooth way and that there exists a constant $\tilde{\gamma}$ such that

$$\gamma(\Phi,\Psi,\varphi) \geq \tilde{\gamma} > 0.$$

With these restrictions on γ, the results on existence of unique solutions and their regularity [6] still hold. In addition, the system will again have a global attractor whose fractal and Hausdorff dimensions can be bounded by the same bounds as in [6] with γ replaced by $\tilde{\gamma}$.

2. The basic attractor. Once the existence of a global attractor has been established, the natural question arises: How can one construct the global attractor, and can one obtain a system of ODEs that describe the evolution on the attractor? There are, for the most part, two main approaches that researchers have taken here.

The first one, and the more popular one, is to think of the attractor as a set embedded in a larger manifold, often called an inertial manifold (see, e.g., [11] and [13]). The problem of finding ODEs that describe the flow on this manifold or an approximate manifold is then solved by using a Galerkin projection onto a basis (see,

e.g., [27] and [21]). The number of basis vectors needed is often quite large. This can be due to either the fact that the bounds on the dimension of the attractors found by current methods tend to be rather conservative, or that the asymptotic dynamics of the system in question are in fact high dimensional. Hence the system of ODEs is not tractable for analytical analysis but lends itself better to numerical work.

The second approach is to consider only the core of the attractor called the basic attractor (see below). Here one constructs the particular solutions in the attractor which attract "almost all" of the phase space. For some systems, the asymptotic dynamics in this "almost every" sense are low dimensional, and one can completely determine the flow on the basic attractor *analytically*.

Here we adopt the second approach, but before we go further, let us clarify what we mean by the basic attractor and by "almost every."

2.1. Prevalence and basic attractors. We need to extend the measure theoretic terms measure zero and almost every to infinite dimensional Banach spaces. Furthermore, we want to do it in such a way that these definitions behave well under the operations of the vector space. It turns out that it suffices that they behave well under translations of the set. The problem here lies in that there do not exist any nontrivial translation invariant measures in infinite dimensional spaces. If a subset $U \subset X$ in an infinite dimensional Banach space is nonempty and μ is a translation invariant measure on X, then either $\mu(U) = 0$ or $\mu(U) = \infty$. Following Hunt, Sauer, and Yorke [17], the ideas of measure zero and almost every can be replaced by shy and prevalent.

DEFINITION 2.1. *Let X denote a separable Banach space. We denote by $S + v$ the translate of the set $S \subset X$ by a vector v. A measure μ is said to be* transverse *to a Borel set $S \subset X$ if the following two conditions hold.*

- *There exists a compact set $U \subset X$ for which $0 < \mu(U) < \infty$.*
- *$\mu(S + v) = 0$ for every $v \in X$.*

A Borel set $S \subset X$ is called shy *if there exists a compactly supported measure transverse to S. More generally, a subset of X is called shy if it is contained in a shy Borel set. The complement of a shy set is said to be a* prevalent *set.*

The basic attractor should be the smallest part of the global attractor $\mathcal{A}$ which attracts a prevalent set. Let us make this more precise.

DEFINITION 2.2. *An attractor $\mathcal{B}$ is a* basic attractor *if it satisfies the following two conditions.*

1. *The basin of attraction of $\mathcal{B}$ is prevalent.*
2. *$\mathcal{B}$ is minimal with respect to property* (1)*, i.e., there exists no strictly smaller $\mathcal{B}' \subset \mathcal{B}$ with* $\text{basin}(\mathcal{B}) \subset \text{basin}(\mathcal{B}')$*, up to shy sets.*

This means that every point of $\mathcal{B}$ is essential; no point can be removed without removing a portion of the basin that is not shy. In numerical simulations or in physical experiments one would therefore only expect to observe the basic attractor after a long-enough settling period.

In general, the basic attractor will be disconnected although the global attractor is connected. We can therefore speak of components of the basic attractor.

The following theorem, which is an extension of a finite dimensional version by Milnor [25], was proven in Birnir [4].

THEOREM 2.3. *Let $\mathcal{A}$ be the compact attractor of a continuous map $T(t)$ on a separable Banach space X. Then $\mathcal{A}$ can be decomposed into a basic attractor $\mathcal{B}$ and a remainder $\mathcal{C}$,*

$$\mathcal{A} = \mathcal{B} \cup \mathcal{C},$$

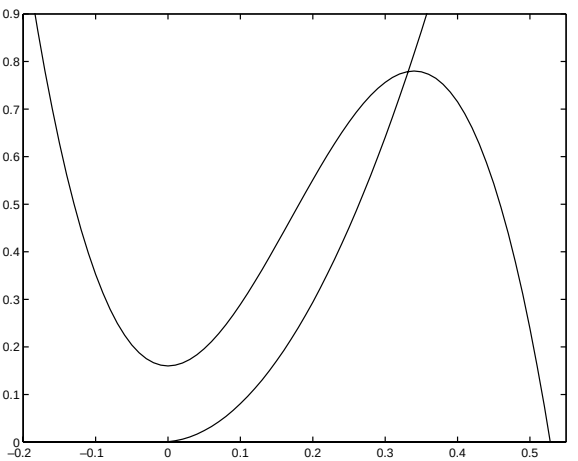


FIG. 2.1. *Two characteristics: the cubic compressor characteristic and the parabolic throttle characteristic. The intersection of these curves is a stationary solution for* $\varphi = 0$. *They are stable to the right of the peak.*

such that basin$(\mathcal{B})$ *is prevalent and basin*$(\mathcal{C})\backslash$*basin*$(\mathcal{B})$ *is shy.*

The basic attractor $\mathcal{B}$ for the vMG equation depends on the parameters in the equation; in particular, it depends on the throttle parameter $\mathcal{B} = \mathcal{B}(\gamma)$. It will prove useful later to have the following definition. Let $\tilde{\gamma} > 0$. The *basic union attractor* is the union of basic attractors

$$\cup_{\gamma \geq \tilde{\gamma}} \mathcal{B}(\gamma).$$

A complete description of the basic union attractor for the vMG equation has been given in [7] and [8]. It turns out that in the cases where an explicit description of $\mathcal{B}$ has been given (see also [5] and [9]), the dimension of $\mathcal{B}$ is small, whereas the dimension of $\mathcal{C}$ can be quite large.

2.2. The geometry of the basic attractor. Experimental and numerical evidence indicate that the basic attractor in axial compression systems is low dimensional [12], [14], [15], [30]. It can consist of a combination of axisymmetric design flow, surge, and stall. The design flow is a stationary solution, and surge is a periodic cycle which has been well studied (see [26] and [23]). It only involves the two ODEs in the system (1.1)–(1.3). Stall has been studied in low-order Galerkin truncations of the Moore–Greitzer equations [23], [20].

2.2.1. Design flow. Under normal conditions the engine operates in design flow. There the flow through the compressor is uniform in space and time and the pressure rise is relatively high. In particular, $\varphi = 0$, and Φ and Ψ are constant.

Figure 2.1 shows the (Φ, Ψ) plane. The parabola starting at the origin represents all stationary solutions for (1.3) and is called the throttle characteristic. The cubic curve represents all stationary solutions for (1.2), given that $\varphi = 0$, and is called the compressor characteristic. Since $\varphi = 0$ is a stationary solution for (1.1), we can conclude that the intersection of the two curves in Figure 2.1 is a stationary solution for the full system (1.1)–(1.3). This stationary solution is called design flow.

To analyze the stability of design flow $(0, \Phi_0, \Psi_0)$, we linearize the system about the design flow solution. Let us define the variable $y = (y_1, y_2, y_3) = (\varphi, \Phi - \Phi_0, \Psi -$

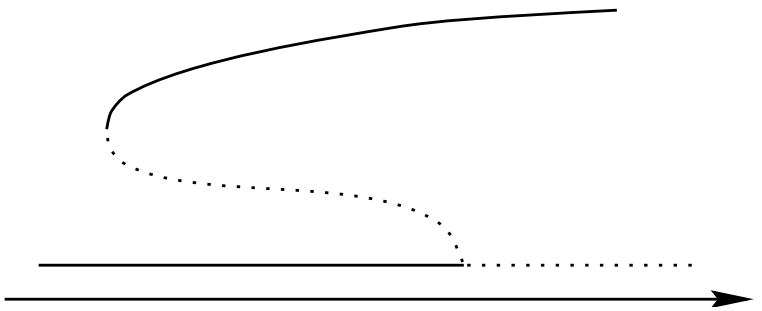


FIG. 2.2. *The design flow undergoes a series of subcritical Hopf bifurcation to surge and stall. The amplitude of the periodic orbits is plotted as a function of the bifurcation parameter.*

Ψ_0). The linearized system is

$$\dot{y} = Ay,$$

where

$$A = \begin{bmatrix} \nu\partial_\theta^2 + \psi_c'(\Phi_0) & 0 & 0 \\ 0 & \frac{1}{l_c}\psi'(\Phi_0) & -\frac{1}{l_c} \\ 0 & \frac{1}{4l_cB^2} & \frac{-\gamma}{4l_cB^2}(F_T^{-1})'(\Psi_0) \end{bmatrix}.$$

This is a block diagonal system, and its stability and bifurcations have been analyzed completely (see [23] and [10]). When the throttle parameter γ is decreased, the average flow Φ decreases as well and the design flow undergoes a series of subcritical Hopf bifurcations. If the parameter B is large enough, there will be one Hopf bifurcation originating from the two dimensional block related to the two ODEs (1.2)–(1.3). This bifurcation gives rise to surge. Furthermore, when $\psi_c'(\Phi) = \nu n^2$, then the nth Fourier harmonic will become unstable and another subcritical Hopf bifurcation occurs. The number of these Hopf bifurcations, which give rise to stall, is bounded by

$$n_{max} \leq \sqrt{\frac{1}{\nu}\max_{\Phi}\psi_c'(\Phi)}. \tag{2.1}$$

Figure 2.2 shows how the amplitude of the periodic orbits originating from the subcritical Hopf bifurcations varies with the bifurcation parameter.

Design flow is stable to the right of the peak of the compressor characteristic. It is desirable to operate the engine on the right side of the peak with as high a pressure rise as possible without risking the system being thrown over to the unstable side by disturbances.

2.2.2. Surge. Surge is a limit cycle in the two ODEs (1.2)–(1.3), where the nonaxisymmetric disturbance is zero, $\varphi = 0$. It has been studied by many authors, among them Greitzer [15] and McCaughan [22], [23]. It arises as a subcritical Hopf bifurcation in the system

$$\dot{\Phi} = \frac{1}{l_c}(\psi_c(\Phi) - \Psi),$$
$$\dot{\Psi} = \frac{1}{4B^2l_c}(\Phi - \gamma F_T^{-1}(\Psi)),$$

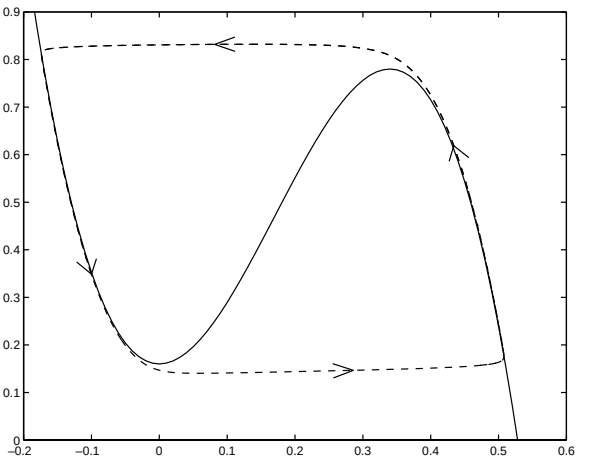


FIG. 2.3. *The surge limit cycle in the* (Φ, Ψ) *plane where* $\varphi = 0$.

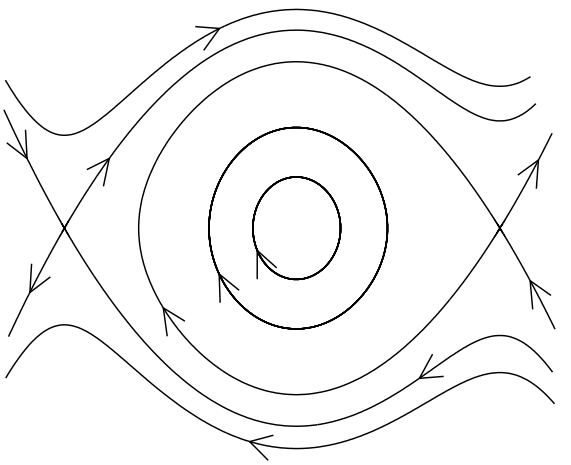


FIG. 2.4. *The phase portrait for Duffing's equation in the general case when* $\gamma \neq \gamma_*$.

which occurs for a large enough B when the throttle parameter γ is decreased. Since the bifurcation is subcritical, we have a one-parameter family of unstable surge cycles that originates from the bifurcation point. This branch bends on itself, and the cycles become stable [23], [10]. These stable cycles are fairly large, and a simulation of one is shown in Figure 2.3.

The solution spends most of its time on the two vertical sides of the cycle. There the slope of the compressor characteristic is negative, so all nonaxisymmetric disturbances are damped.

2.2.3. Stall. Stall is a solitary wave solution. The wave rotates around the unit circle, and the average flow Φ and pressure rise Ψ are constant. When one looks for traveling wave solutions of the vMG equation, the problem can be reduced to finding periodic solutions of Duffing's equation with the correct periods [7], [8]. These periodic solutions lie inside a homoclinic (or heteroclinic) orbit (see Figure 2.4), and since the compressor characteristic is a cubic polynomial, these solutions can be found explicitly

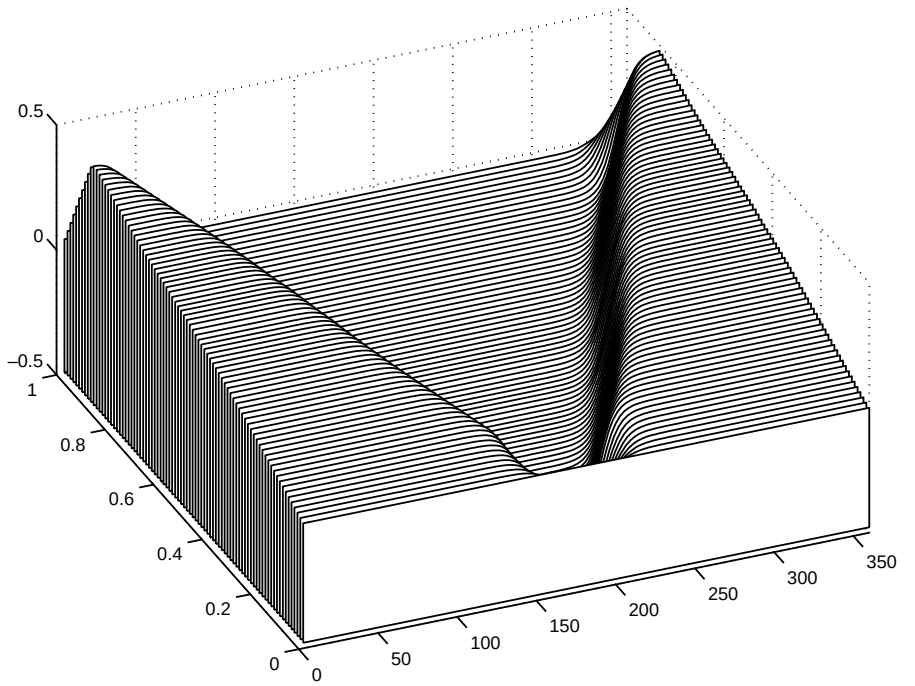


FIG. 2.5. *The one-parameter family of stall cells. As the parameter β varies from zero to one, the stall cell grows from a constant zero solution to a narrow stall cell that slowly widens until it fills out the annulus and the stall disturbance becomes zero again.*

with quadratures. They can be expressed as rational functions of the Jacobi elliptic function ns [29].

$$\varphi(\eta) = \frac{p - q\sqrt{\frac{-B_2}{A_2}}ns\left(\eta(p-q)\sqrt{-B_1A_2} + i\hat{\omega}_2, \sqrt{\frac{B_2A_1}{B_1A_2}}\right)}{1 - \sqrt{\frac{-B_2}{A_2}}ns\left(\eta(p-q)\sqrt{-B_1A_2} + i\hat{\omega}_2, \sqrt{\frac{B_2A_1}{B_1A_2}}\right)}.$$

Here $\hat{\omega}_2$ is half the imaginary period of ns and p, q, A_1, A_2, B_1, B_2 are constants.

The shape of the solitary wave depends on the parameters in the equation; in particular, the shape depends on γ. By varying the parameters we can in fact construct a one-parameter family of stall solutions. The one-parameter family for the one-pulse solitary waves is shown in Figure 2.5.

It can be shown that for a large parameter range (see [7]), the stall solutions in the one-parameter family are asymptotically stable and hence belong to the basic union attractor.

3. Basic controllability. Let us consider now the issue of controllability. In finite dimensional control theory, a system is said to be controllable if for every two points $x_0, x_1 \in X$ and every two real numbers $t_0 < t_1$, there exists a control function u such that the unique solution of the equation

$$\dot{x} = F(x,u), \qquad x(t_0) = x_0 \tag{3.1}$$

satisfies $x(t_1) = x_1$.

In infinite dimensional spaces this notion of controllability is too restrictive. For practical control applications one can never have more than finitely many control parameters, if for no other reason than the fundamentals of computing require computer outputs to be finite. There is therefore no hope that nonlinear evolution equations in infinite dimensional spaces will be controllable in this strict sense in practical applications.

If an evolution equation has an attractor and a basic attractor, its solutions will converge asymptotically to the attractor for all initial conditions and to the basic attractor for almost all initial conditions. The simplest thing one could ask of the control is that it make all or almost all initial conditions give rise to solutions that converge to a given component in the basic attractor. A more stringent requirement on the control would be that it make all or a prevalent set of (almost all) initial conditions give rise to solutions that converge to a given component in the global attractor. This requires one to have enough control authority over the local unstable manifolds of the hyperbolic trajectories in the attractor to make them attractive. Consider the following definitions.

DEFINITION 3.1. (3.1) *is* basically controllable *if for all bounded sets* M, *all* $x_o \in M$, *every minimal component of the basic attractor* $\mathcal{B}_j$, *and all* $\epsilon > 0$, *there exists a finite time* $T(M)$ *and a control function* $u(t)$, *such that the solution* $x(t)$ *with initial data* $x_o \in M$ *satisfies*

$$\|x(t) - \mathcal{B}_j\| < \epsilon$$

for all $t > T(M)$.

This definition says that given an initial point one can steer to any component of the basic attractor in finite time. It is hopeless to get a finite T for x_o lying in a prevalent (full measure) set in the infinite dimensional space, for the reason discussed above. It is not wise to attempt to control every solution in the $\mathcal{A}$-attractor, because in general it ($\mathcal{C}$) contains many hyperbolic solutions and their heteroclinic connections.

DEFINITION 3.2. (3.1) *is* attractively controllable *if for all bounded sets* M, *all* $x_o \in M$, *every trajectory* z *in the attractor* $\mathcal{A}$, *and all* $\epsilon > 0$, *there exists a finite time* $T(M)$ *and a control function* $u(t)$, *such that the solution* $x(t)$ *with initial data* $x_o \in M$ *satisfies*

$$\|x(t) - \omega(z)\| < \epsilon$$

for all $t > T(M)$.

This definition says that given an initial point one can steer to the ω limit set of any trajectory in the $\mathcal{A}$-attractor in finite time.

We will in what follows, for the sake of brevity, also speak of basic controllability as b-controllability and attractive controllability as a-controllability. Clearly a-controllability implies b-controllability. Not surprisingly, the control construction relies heavily on the geometry of the basic attractor. Consequently it is referred to as *basic control*. The remainder $\mathcal{C} = \mathcal{A} \backslash \mathcal{B}$ from section 2.1 plays a large role in basic control. In general one would like to use its heteroclinic connections to move efficiently from one minimal basic attractor to another.

To get stronger results than a-controllability or b-controllability of the form that one could get from an arbitrary initial condition to a point in the attractor in a short time is, in general, hopeless. Since one's control only actuates finitely many dimensions, one would need to wait an arbitrarily long time (determined by the dissipation rate) for the part of the solution which, in some sense, is perpendicular to the control action to settle onto the desired point.

4. Basic control for design flow of the linearized equation. The most important component of the basic attractor of the vMG equation is the design flow component. The goal is to construct a basic control that makes all solutions converge to the design flow. The simplest approach one could take would be to linearize the system about design flow, $(\Phi_0, \Psi_0, 0)$, corresponding to a throttle parameter γ_0, and apply the classical optimal control theory. We define the control parameter $u = \gamma - \gamma_0$, and we make a change of coordinates $(t, \eta) = (t, \theta - \frac{1}{2}t)$ to simplify the equations. Furthermore, we define the variable $y = (y_1, y_2, y_3) = (\varphi, \Phi - \Phi_0, \Psi - \Psi_0)$. The linearized equations can now be written as

$$\dot{y} = Ay + Bu, \tag{4.1}$$

where

$$A = \begin{bmatrix} \nu\partial_\theta^2 + \psi_c'(\Phi_0) & 0 & 0 \\ 0 & \frac{1}{l_c}\psi'(\Phi_0) & -\frac{1}{l_c} \\ 0 & \frac{1}{4l_cB^2} & \frac{-\gamma_0}{4l_cB^2}(F_T^{-1})'(\Psi_0) \end{bmatrix} \tag{4.2}$$

and

$$B = \begin{bmatrix} 0 \\ 0 \\ \frac{1}{4l_cB^2}(F_T^{-1})'(\Psi_0) \end{bmatrix}. \tag{4.3}$$

Since the operator A is sectorial, it generates an analytic semigroup in X. We denote by $T(t)$ the semigroup operator on X, and the norm and inner product will be denoted by $\|\cdot\|$ and $\langle\cdot,\cdot\rangle$, respectively. In this form the equations can be tackled using the standard optimal control theory in Hilbert spaces (see Lions [19] and Banks [3]).

Observe first that this system is block diagonal. It can be split into two parts: a two dimensional part that describes the evolution of the average flow and the pressure rise, and a part of codimension 2 which describes the evolution of stall. This second part does not depend on the control parameter γ and can therefore be integrated separately. In other words, stall does not depend on the control parameter and is therefore uncontrollable. The problem is now reduced to a two dimensional problem.

We seek a feedback control that will minimize the cost functional

$$J(y) = \frac{1}{2}(y_1^2(t_f) + y_2^2(t_f)) + \frac{1}{2}\int_0^{t_f} S(y_1^2(t) + y_2^2(t)) + Ru^2 dt.$$

It is a well-known result [3] that the optimal feedback control is given by

$$u = -\frac{1}{R}B^T Q(t) y(t),$$

where the symmetric matrix $Q(t)$ satisfies the matrix Riccati equation

$$\dot{Q}(t) = -Q(t)A - AQ(t) + \frac{1}{R}Q(t)BB^TQ(t) - SI, \qquad Q(t_f) = I.$$

Here I is the identity matrix and S and R are positive constants. Solving the Riccati equation is equivalent to solving the three ODEs

$$\begin{aligned} \dot{q}_{11} &= (\tfrac{1}{4l_cB^2}(F_T^{-1})'(\Psi_0))^2 q_{12}^2 - S - 2(a_{11}q_{11} + a_{21}q_{12}), \\ \dot{q}_{12} &= (\tfrac{1}{4l_cB^2}(F_T^{-1})'(\Psi_0))^2 q_{12}q_{22} - a_{12}q_{11} - (a_{11} + a_{22})q_{12} - a_{21}q_{22}, \\ \dot{q}_{22} &= (\tfrac{1}{4l_cB^2}(F_T^{-1})'(\Psi_0))^2 q_{22}^2 - S - 2(a_{12}q_{12} + a_{22}q_{22}), \\ q_{ij}(t_f) &= \delta_{ij}, \end{aligned} \tag{4.4}$$

where δ_{ij} is the Kronecker delta. This system can be solved backwards in time numerically. The optimal feedback control is now given by

$$u = \frac{1}{4l_c B^2}(F_T^{-1})'(\Psi_0)((\Phi - \Phi_0)q_{12}(t) + (\Psi - \Psi_0)q_{22}(t)).$$

Design flow solutions that sit to the right of the peak on the compressor characteristic (see Figure 2.1) have the property that all the uncontrollable modes are stable, so in this case the design flow is b-controllable and, in fact, a-controllable. However, the fact that stall is uncontrollable in the linear system renders this control of limited use. It is well known that in the nonlinear system for design flow solutions close to the peak of the characteristic, small nonaxisymmetric disturbances can cause the system to go into stall. On the other hand, the result by Banaszuk, Hauksson, and Mezić [2] proves that design flow can be globally stabilized by using only throttle control. Their control is, however, not optimal in any sense of the word, and it reacts very violently to high frequency disturbances which are damped in the viscous model. The question that now arises is whether there is something in between these two control constructions, i.e., does there exist a control which globally stabilizes the design flow but does not require as much control effort as the control constructed by Banaszuk, Hauksson, and Mezić? We tackle this problem in the next section.

5. Basic control of design flow for the nonlinear equation. The basic attractor for constant throttle functions has been analyzed completely [7], [8], and one would like to use this knowledge of the asymptotic dynamics when constructing a control law. However, when γ is no longer constant but a function of the state variables, the components of the basic attractor may change, altering the asymptotic dynamics.

Let us assume for now that we only consider control strategies that move the throttle in an adiabatic fashion. Restricting the control to this class guarantees that the basic attractor is unchanged. The best one can hope to do here is to slide the solution along the basic attractor until it reaches the desired operation point. If stall occurs, then one slides the system along the branch of stable stall cells by increasing γ until the saddle-node bifurcation point is reached, stall ceases to exist, and the flow converges to design flow. This design flow is achieved at a very low pressure rise. In order to increase the pressure rise we decrease γ again until the desired operation point is reached. Figure 5.1 shows the bifurcation diagram for the first stall solution.

Within the class of adiabatic controls, this is the optimal control. When one considers the larger class of controls that are not necessarily adiabatic, this control strategy is no longer optimal. It has, however, been shown that a reduced order model, which reduces the flow to the basic attractor, captures the dynamics of the full model not only qualitatively but also quantitatively [8]. Transient behavior, while the flow is going from one component of the basic attractor to another, is also well captured. With this in mind it makes sense to modify the adiabatic control construction by tracking trajectories on the basic attractor. If the tracking is done in an optimal fashion, one would hope that the resulting control strategy is close to an optimal strategy with respect to some cost function. We claim that our basic control strategy is in this sense near optimal.

Let $\xi(t)$ be a parameterization of a family of stationary solutions in the basic union attractor $\cup_\gamma \mathcal{B}$. Note that since we are working in a rotating frame of reference, (η, t), stall solutions will be stationary solutions. Then, corresponding to $\xi(t)$, we can find $\bar{\gamma}(t)$ such that for $\gamma = \bar{\gamma}(t)$, $\xi(t)$ is a stationary solution of (1.1)–(1.3). We

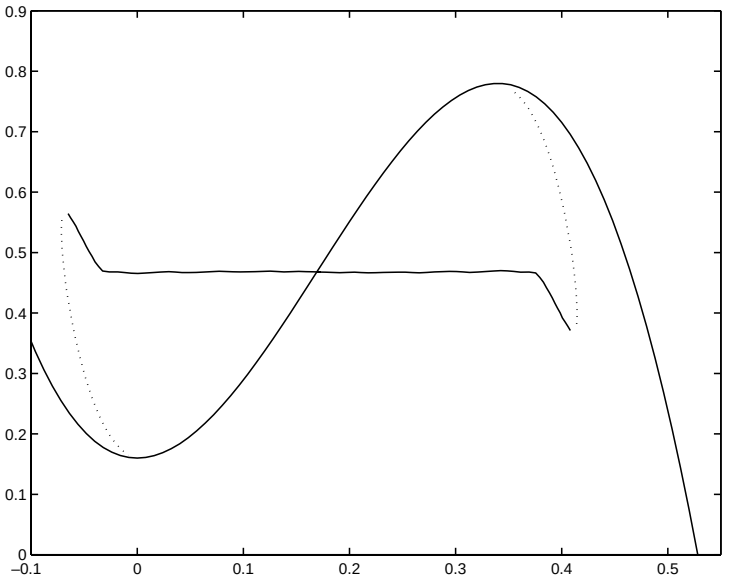


FIG. 5.1. *This figure shows the bifurcation diagram for the first stall solution in the* (Φ, Ψ)*-plane. The flat solid curve represents the stable branch, and the dotted curves represent the unstable branches of stall cells.*

denote the solution of the system (1.1)–(1.3) as $x(t) = (\Phi, \Psi, \varphi)(t) = \xi(t) + y(t)$ and the control parameter $\gamma = \bar{\gamma}(t) + u(t)$. Let us now linearize this system about the trajectory $\xi(t)$ and write it as

$$\dot{y} = A(t)y + B(t)u - \dot{\xi}(t). \tag{5.1}$$

Here $A(t)$ and $B(t)$ depend on time through the trajectory $\xi(t)$. Our goal is to make y as small as possible with very little control effort. In other words, we want to find a regulator for (5.1) which is optimal with respect to the cost function

$$J(u) = \frac{1}{2}\|S_f y(t_f)\|^2 + \frac{1}{2}\int_0^{t_f} \|Sy(t)\|^2 + Ru^2(t)dt,$$

where R is a constant and S_f and S are symmetric positive definite linear operators.

This is a well-known problem, and it can be solved exactly (see Sage and White [28] or Banks [3]). The optimal feedback control is given by

$$u(t) = -\frac{1}{R}B^T(t)(Q(t)y(t) - \zeta(t)),$$

where the symmetric operator $Q(t)$ satisfies the equation

$$\dot{Q}(t) = -Q(t)A(t) - A(t)Q(t) + \frac{1}{R}Q(t)B(t)B^T(t)Q(t) - S, \qquad Q(t_f) = S_f, \tag{5.2}$$

and the function $\zeta(t)$ satisfies

$$\dot{\zeta}(t) = -\left(A(t) - \frac{1}{R}B(t)B^T(t)Q(t)\right)^T \zeta(t) - Q(t)\dot{\xi}(t), \qquad \zeta(t_f) = 0. \tag{5.3}$$

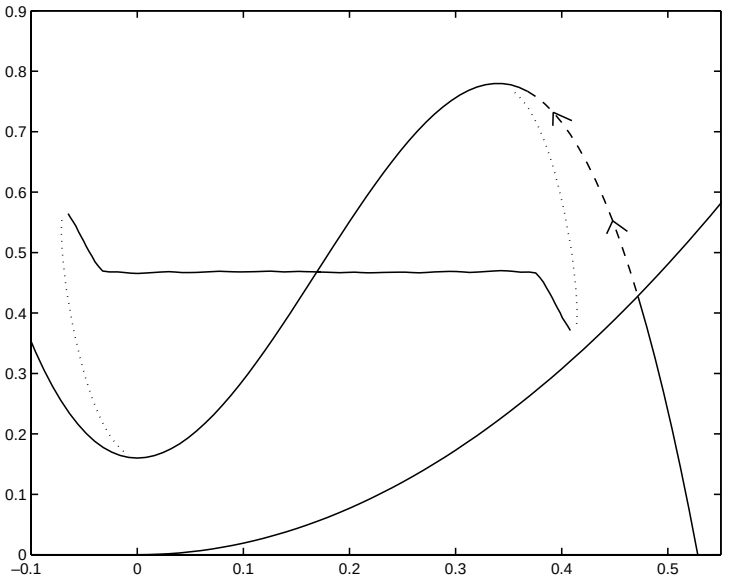


FIG. 5.2. *This figure shows the throttle setting $\gamma = \gamma_1$ that defines the start of the trajectory ξ_1, which is shown as a dashed line.*

These equations can be solved backwards in time to yield the optimal control u.

It turns out that when one linearizes the system about a trajectory on the basic attractor, then all but finitely many directions in state space will be uncontrollable. The uncontrolled dynamics are stable, and, as a result, (5.2) and (5.3) will reduce to finitely many ODEs.

5.1. Construction of the controller. To make the construction of the controller as simple and intuitive as possible, we proceed in the following way. When a disturbance occurs in the system that is large enough so that the system cannot recover without intervention, we change the control parameter to a setting where the only component in the basic attractor is the design flow. This consists of increasing γ to a level γ_1 so that the throttle characteristic no longer intersects the branch of stall cells. We then wait until the flow is in a small-enough neighborhood U of the design flow. This design flow setting is, however, at a low pressure rise level, so to increase the pressure rise we now track a trajectory ξ_1 to the desired design flow setting (see Figure 5.2). As we will prove later, if the state is close enough to the starting point of ξ_1 and the cost on the control small enough, this strategy will work for all initial disturbances.

This control construction will still be very close to the original one as the system will settle into stall or surge very fast and then traverse near the basic attractor towards the design flow corresponding to the throttle setting γ_1.

The linearization of the system about the trajectory ξ_1 is exactly that given by (4.2) and (4.3), except for that now these operators are time-dependent, i.e., instead of (Φ_0, Ψ_0) we have $(\Phi, \Psi)_{\bar{\gamma}(t)}$. Just like before, there is an uncontrollable subspace of codimension 2, but since we are on the right side of the peak of the characteristic, this space is stable and small disturbances will decay in time. We therefore only need to consider the first two modes which describe the flow in the (Φ, Ψ)-plane, and it

suffices to know q_{11}, q_{12}, q_{22}, ζ_1, and ζ_2. The equations for these coefficients are

$$\begin{aligned}
\dot{q}_{11} &= (\tfrac{1}{4l_cB^2}(F_T^{-1})'(\Psi_{\bar{\gamma}(t)}))^2\tfrac{q_{12}^2}{R} - s_{11} - 2(a_{11}(t)q_{11} + a_{21}(t)q_{12}),\\
\dot{q}_{12} &= (\tfrac{1}{4l_cB^2}(F_T^{-1})'(\Psi_{\bar{\gamma}(t)}))^2\tfrac{q_{12}q_{22}}{R} - s_{12} - a_{12}(t)q_{11}\\
&\quad -(a_{11}(t) + a_{22}(t))q_{12} - a_{21}(t)q_{22},\\
\dot{q}_{22} &= (\tfrac{1}{4l_cB^2}(F_T^{-1})'(\Psi_{\bar{\gamma}(t)}))^2\tfrac{q_{22}^2}{R} - s_{22} - 2(a_{12}(t)q_{12} + a_{22}(t)q_{22}),\\
q(t_f)_{ij} &= s_{fij}
\end{aligned} \tag{5.4}$$

and

$$\begin{aligned}
\dot{\zeta}_1 &= -a_{11}(t)\zeta_1 - (a_{21}(t) - (\tfrac{1}{4l_cB^2}(F_T^{-1})'(\Psi_{\bar{\gamma}(t)}))^2\tfrac{q_{21}}{R})\zeta_2 - q_{11}\dot{\xi}_1 - q_{12}\dot{\xi}_1,\\
\dot{\zeta}_2 &= -a_{12}(t)\zeta_1 - (a_{22}(t) - (\tfrac{1}{4l_cB^2}(F_T^{-1})'(\Psi_{\bar{\gamma}(t)}))^2\tfrac{q_{22}}{R})\zeta_2 - q_{12}\dot{\xi}_1 - q_{22}\dot{\xi}_1,\\
\zeta(t_f) &= 0.
\end{aligned} \tag{5.5}$$

The basic feedback control is now given by

$$u(t) = \gamma_1 \tag{5.6}$$

if $x(t)$ has not reached the neighborhood U and

$$u(t) = \frac{1}{4l_cB^2R}(F_T^{-1})'(\Psi_{\bar{\gamma}(t)})[(\Phi - \Phi_{\bar{\gamma}(t)})q_{12}(t) + (\Psi - \Psi_{\bar{\gamma}(t)})q_{22}(t) + \zeta_2] \tag{5.7}$$

otherwise.

THEOREM 5.1. *There exists an open set U around $\xi_1(0) = (\varphi_1, \Phi_1, \Psi_1)$, a constant R, and a prevalent set $Y \subset X$ such that the control strategy given by* (5.6) *and* (5.7) *make all solutions of the vMG equation, with initial conditions in Y, converge to the desired stable design flow.*

Proof. For $u = \gamma_1$, the only component in the basic attractor is the design flow at the starting point of $\xi_1 = \xi_1(0)$. Hence there exists a prevalent set $Y \subset X$ such that for all initial conditions in Y the solution will converge to $\xi_1(0)$. Let $U \subset X$ be an open neighborhood of $\xi_1(0)$. By making U smaller, we can make sure that the stall disturbance is as small as we please and that (φ, Φ, Ψ) is as close to the starting point of the trajectory ξ_1 as we please before we start tracking the trajectory.

On the other hand, if there is no stall disturbance, i.e., if $\varphi = 0$, then by decreasing R we can track the trajectory ξ_1 as closely as we want. The feedback control is robust with respect to small disturbances, so as long as the stall disturbance stays small we can guarantee that we stay within a small neighborhood of the trajectory. It therefore remains only to show that small stall disturbances will remain small.

The energy method for (1.1) gives us that

$$\begin{aligned}
\tfrac{1}{2}\tfrac{d}{dt}\|\varphi\|_1^2 &\le -\nu\|\varphi_\eta\|_1^2 + \tfrac{1}{2\pi}\int_0^{2\pi}\psi_c'(\Phi + \varphi)\varphi_\eta^2 d\theta\\
&\le (-\nu + \tfrac{1}{2\pi}\max_\theta \psi_c'(\Phi + \varphi))\|\varphi\|_1^2.
\end{aligned}$$

Here $\|\cdot\|_1$ is the Sobolev norm in $\bar{H}^1$. The trajectory ξ_1 lies to the right of the peak of the compressor characteristic, and on that side the derivative of the characteristic is negative. By making Φ stay close enough to the trajectory and the L^∞ norm of φ small enough, we can guarantee that the second term will be less than or equal to

zero. The L^∞ norm is bounded by the $\bar{H}^1$ norm, and as a result if (φ, Φ, Ψ) stays close to the trajectory, then small stall disturbances will decay. □

One can clearly construct the trajectory so that the point being tracked by the control will be moving in a smooth fashion. Furthermore, parameterizing the trajectory in a very slow manner approaches the adiabatic control construction which will guarantee that γ will be positive and bounded away from zero. This then guarantees that the control given will be a smooth function, and the existence and regularity results mentioned in the introduction still hold.

5.2. Modifications of control construction. The control construction given here above is but one of many possible constructions. It could be beneficial to construct a regulator for the starting point of ξ_1 instead of just setting $u = \gamma_1$, or one could track a trajectory along the stall branch. The first extension is very simple, but the other one is not as trivial and we outline an approach to it here. The hard part lies in linearizing the system in the correct basis so that the problem reduces to a finite dimensional problem.

In [7] we proved that for specific parameter values when one linearizes the system about a stall solution there is a subspace of codimension 3 which is uncontrollable but stable for a large parameter range. The three dimensional subspace which can be influenced by the throttle parameter is spanned by the average flow, the pressure rise, and one more vector. This vector corresponds to the internal degree of freedom that widens and shrinks the stall cell, and is given by $(\varphi^2 - \overline{\varphi^2}, 0, 0)$, where φ is the stall solution and $\overline{\varphi^2}$ represents the average of φ^2. One of the key elements of that proof was that a part of the linearized operator, the Lamé operator, given by

$$L = \nu\partial_{\eta\eta} + \psi_c'(\Phi + \varphi),$$

had two eigenvectors given by $\varphi^2 + a_1$ and $\varphi^2 + a_2$, and those two combined to give the vector that interacted with pressure rise and average flow.

Now let (φ, Φ, Ψ) be a general stall solution. Then the corresponding operator L will have one zero eigenvalue corresponding to translations of the solution.

THEOREM 5.2. *Assume that the self-adjoint operator L has two eigenvectors of the form $\chi^+ = \varphi^2 + k_1^+\varphi + k_2^+$ and $\chi^- = \varphi^2 + k_1^-\varphi + k_2^-$ corresponding to distinct eigenvalues, each with one dimensional eigenspace. Assume also that the spectrum has a simple zero eigenvalue and the rest of the spectrum is negative.*

Then the vMG equation linearized about the stall solution has an uncontrollable subspace $\mathcal{Q}$ of codimension 3 which is stable, and $\mathcal{Q}^\perp$ is the span of $\{(\chi^+ - \overline{\chi^+}, 0, 0), (0, 1, 0), (0, 0, 1)\}$.

Here the overline represents the angular average as before.

Proof. We know from [7] that the spectrum of the Lamé potential consists of five single eigenvalues and the rest of the spectrum is negative and double and converges to infinity. Varying the Fréchet derivative ψ_c' by a small amount will be a relatively compact perturbation which will perturb the spectrum only slightly [18]. In particular, we can assume that the spectrum of L converges to $-\infty$.

Because (φ, Φ, Ψ) is a traveling wave solution of (1.1)–(1.3) and since the characteristic is a cubic polynomial, we know that

$$\varphi_{\eta\eta} = a\varphi^3 + b\varphi^2 + c\varphi + d$$

for some constants a, b, c, d which depend on Φ. By completing the quadratures, one can also show that φ satisfies the equation

$$\frac{1}{2}(\varphi_\eta)^2 = \frac{1}{4}a\varphi^4 + \frac{1}{3}b\varphi^3 + \frac{1}{2}c\varphi^2 + d\varphi + e$$

for some constant e. Using these two equations, one can show that the two eigenvalues λ^+ and λ^- corresponding to the two eigenvectors χ^+ and χ^- are the real roots of

$$(c - \nu\lambda)\left(\frac{\nu\lambda}{3a} + \frac{c}{a} - \frac{2b^2}{9a^2}\right) - \frac{2bd}{3a} - 4e = 0,$$

and we have

$$\begin{aligned} k_1^+ &= \tfrac{2b}{3a} = k_1^- \\ k_2^+ &= \tfrac{1}{c-\nu\lambda^+}\left(\tfrac{2bd}{3a} + 4e\right) \\ k_2^- &= \tfrac{1}{c-\nu\lambda^-}\left(\tfrac{2bd}{3a} + 4e\right). \end{aligned}$$

Notice that

$$\psi_c'(\Phi + \varphi) = \chi^+ + C = \chi^- + C'$$

for some constants C and C'.

Let $\phi \in H^1 \backslash \mathrm{span}\{\chi^+, \chi^-\}$, and define $\chi = \chi^+ - \overline{\chi^+} = \chi^- - \overline{\chi^-}$. Since eigenspaces corresponding to different eigenvalues are orthogonal and $\mathrm{span}\{\chi^+, \chi^-\} = \mathrm{span}\{\chi, 1\}$, we know that $\langle \phi, 1\rangle = 0$ and thus $\phi \in \bar{H}^1$. Furthermore, since $\psi_c' \in \mathrm{span}\{\chi^+, \chi^-\}$, we have $\overline{\psi_c'}\phi = \langle \psi_c', \phi\rangle = 0$. We can now conclude that if ϕ lies in an eigenspace of L corresponding to the eigenvalue λ, then $(\phi, 0, 0)$ lies in an eigenspace of A corresponding to the same eigenvalue. Let $\mathcal{Q}$ be the subspace spanned by all such functions, and notice that we have $A\mathcal{Q} \subset \mathcal{Q}$.

For any $\phi \in H^1 \backslash \mathrm{span}\{\chi^+, \chi^-\} \subset \bar{H}^1$, we have $(\phi, 0, 0) \in \mathcal{Q}$, and we can conclude that $\mathrm{codim}(\mathcal{Q})$ in $H^1 \times \mathbf{R}^2$ is 4. Since $\mathcal{Q} \subset \bar{H}^1 \times \mathbf{R}^2$, we can also conclude that $\mathrm{codim}(\mathcal{Q})$ in $\bar{H}^1 \times \mathbf{R}^2$ is 3. The orthogonal complement of $\mathcal{Q}$ is $\mathcal{Q}^\perp = \mathrm{span}\{(\chi^+ - \overline{\chi^+}, 0, 0), (0, 1, 0), (0, 0, 1)\}$ and one quickly verifies that $A\mathcal{Q}^\perp \subset \mathcal{Q}^\perp$. From all of the above we conclude that the spectrum of A is

$$\sigma(A) = \sigma(A|_{\mathcal{Q}}) \cup \sigma(A|_{\mathcal{Q}^\perp})$$

and that A is block diagonal with two blocks. The block corresponding to $\mathcal{Q}$ is orthogonal to the control action and is therefore uncontrollable. We know that $\mathcal{Q}$ contains a one dimensional center subspace corresponding to direction of propagation of the stall solution, and the rest of $\mathcal{Q}$ belongs to the stable subspace of A. ∎

As a result of the above theorem, when one wants to track a trajectory along stall solutions, one only has to consider the projection of the linearized system onto the three dimensional space $\mathcal{Q}^\perp$. One therefore only needs to consider a 3×3 projection of $Q(t)$.

Remark 5.3. Theorem 5.2 has some implications for nonaxisymmetric control actuation. If we assume control actuation that does not destroy this spectral structure of the equation, then the theorem would indicate that the nonaxisymmetric component of the control action should be shaped like the vector $\chi = \chi^+ - \overline{\chi^+}$. This would move the state along the one-parameter family of stall cells as opposed to trying to push off this strongly attracting surface.

Figure 5.3 shows the stall cell and the vector χ for two different parameter settings: on the left for the specific parameter values chosen in [7] and on the right for a general parameter setting. This figure shows that the general characteristics of the vector χ are the same.

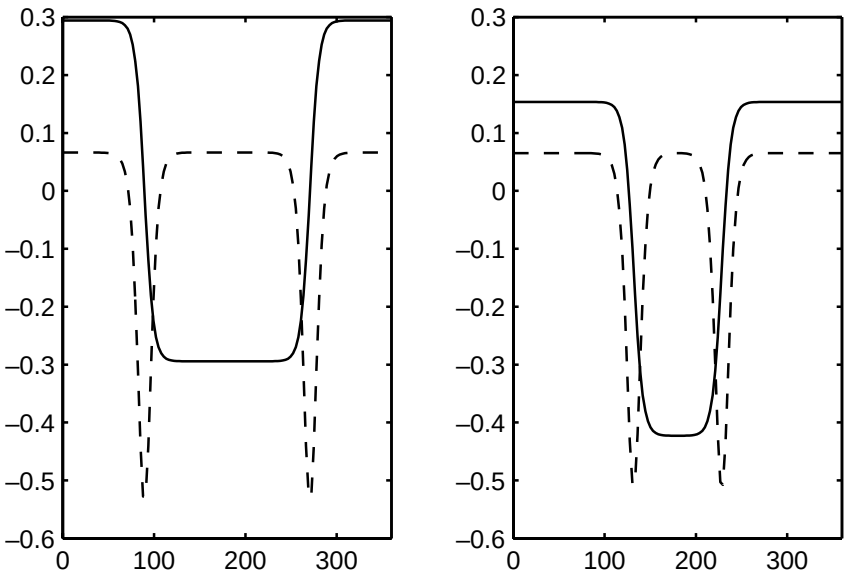


FIG. 5.3. *This figure shows stall cells and the vector χ for the specific parameter settings given in* [7] *on the left and a general parameter setting on the right. The stall cell is given by a solid line and χ by a dashed line.*

6. Numerical simulations. Here we present some numerical simultions that display how our control (5.6)–(5.7) performs, and its performance is compared with that of the backstepping control given by Banaszuk, Hauksson, and Mezić [2]. For all of the simulations the initial condition is a small disturbance in the average flow and pressure rise but a large disturbance in the stall direction. The two ODEs (1.2) and (1.3) are solved by a Runge–Kutta routine, which is coupled together with a Lax–Wendroff scheme, which solves (1.1).

The backstepping control is a much more forceful control that uses more control effort, as can be seen in Figure 6.1. It does, however, kill the disturbance faster, as seen in Figures 6.2, 6.3, and 6.4. The state has a much smaller excursion in the (Φ, Ψ)-plane with the basic control, and in particular the pressure rise never drops completely (see Figures 6.5, 6.6, and 6.4).

It should be clear from these simulations that the backstepping control reacts too strongly to stall disturbances. One could argue that the gain chosen for the stall component in the control construction should be decreased, but here we have set it at its smallest allowed level, or equal to $\sqrt{\pi/6}$ [2].

Surge is in general harder to control than stall. It requires more control effort and is a more violent instability. We present here some simulations which show how the two controls handle a surging compressor. As can be seen in Figure 6.7, both controls are saturated. The control strategy with the least effort that could recover design flow from surge would probably just involve increasing γ slightly and then waiting for the system to complete a single surge oscillation. Figures 6.8 and 6.9 show the (Φ, Ψ) phase planes during the transient, and Figures 6.10 and 6.11 show how the stall transient behaves. The backstepping control does a better job at recovering design flow quickly, but at a greater cost, as can be seen in Figure 6.7. The pressure rise in the compressor is shown as a function of time in Figure 6.12.

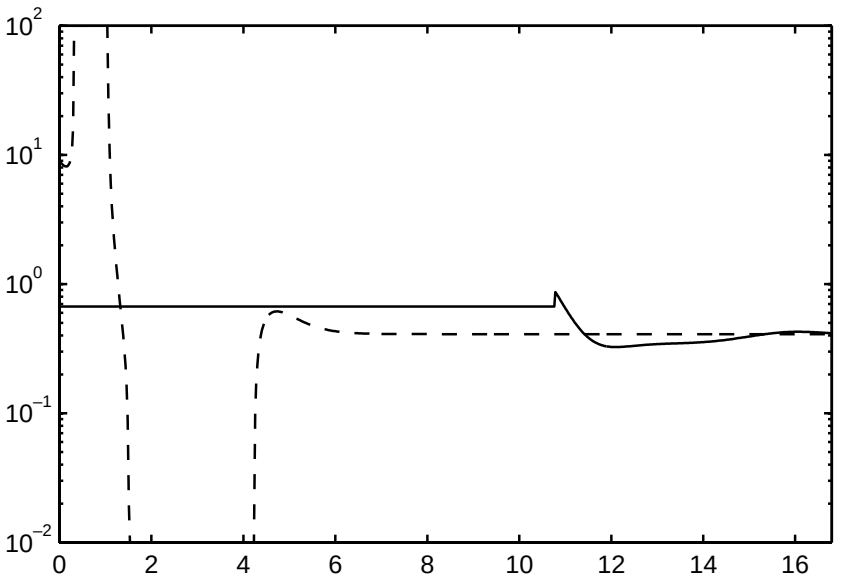


FIG. 6.1. *Here we compare on a logarithmic scale the control effort by the two strategies. The backstepping control result is shown in a dashed line and the basic control in a solid line. The backstepping control kills the disturbance in half the time it takes our control to do so, but at the cost of using extreme control effort.*

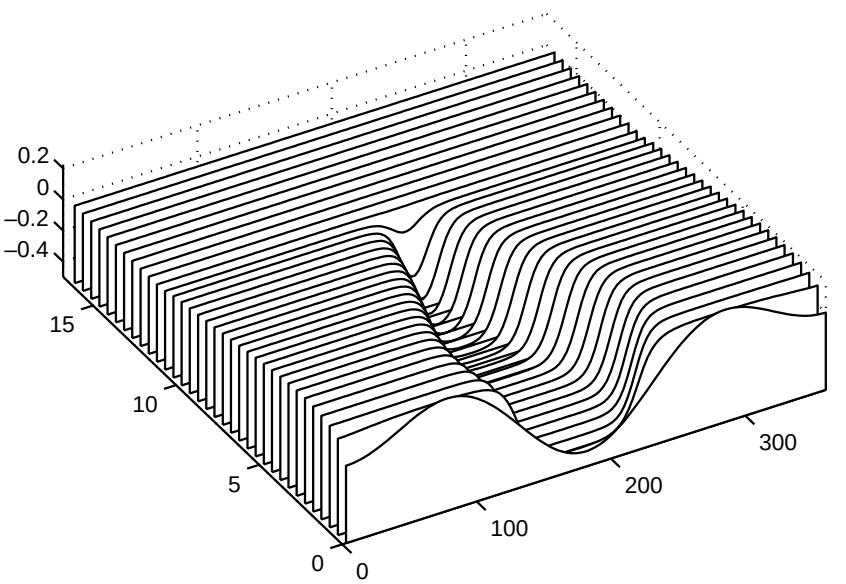


FIG. 6.2. *The evolution of the stall disturbance under the basic control. The system goes into a fully blown stall which then slowly narrows until it vanishes.*

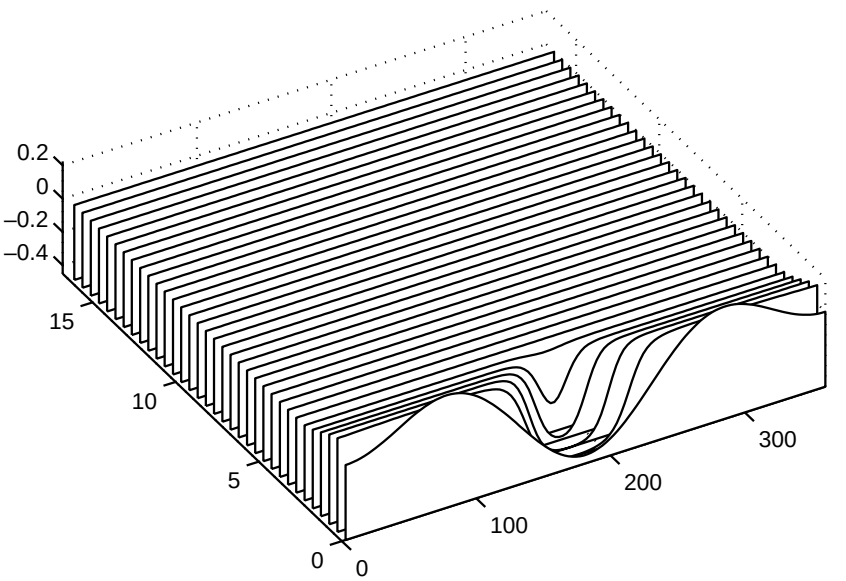


FIG. 6.3. *The evolution of the stall disturbance under the backstepping control. Here the system never reaches a fully blown stall because the control reacts very strongly to stall disturbances.*

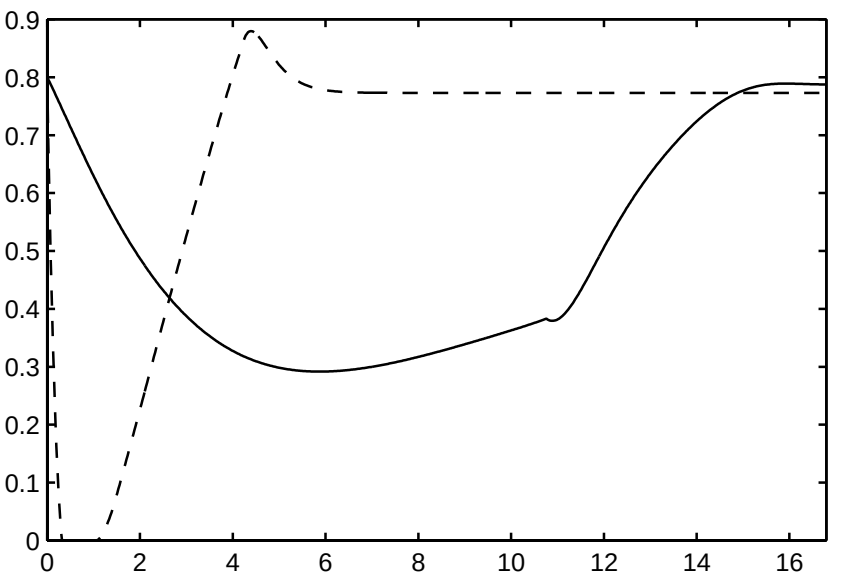


FIG. 6.4. *Here we compare the pressure rise delivered by the compressor under the two strategies. The backstepping control result is shown in a dashed line and the basic control in a solid line. The backstepping control kills the disturbance in half the time it takes our control to do so. However, the pressure rise drops considerably more for the backstepping control.*

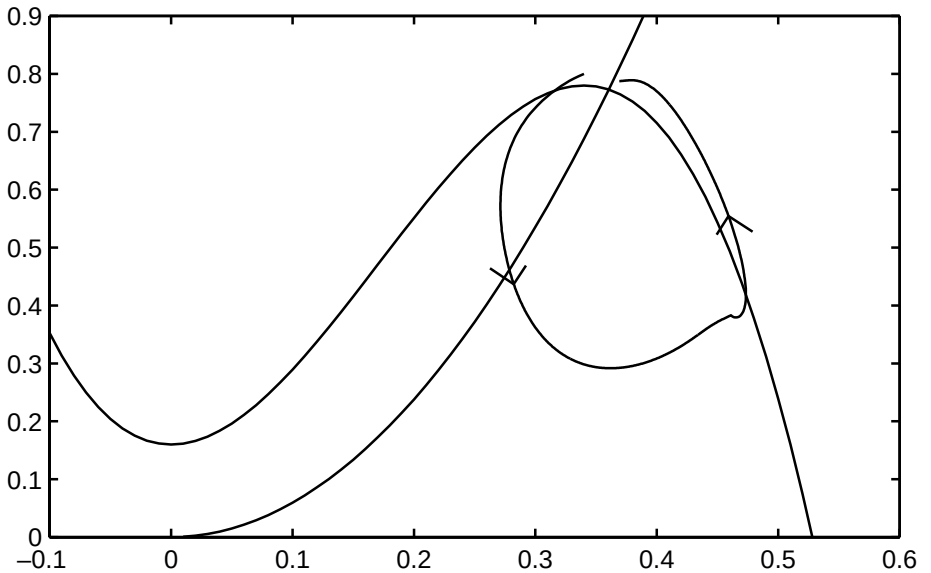


FIG. 6.5. *The* (Φ, Ψ)*-phase plane for the basic control. The large disturbance grows into stall, but design flow is then recovered.*

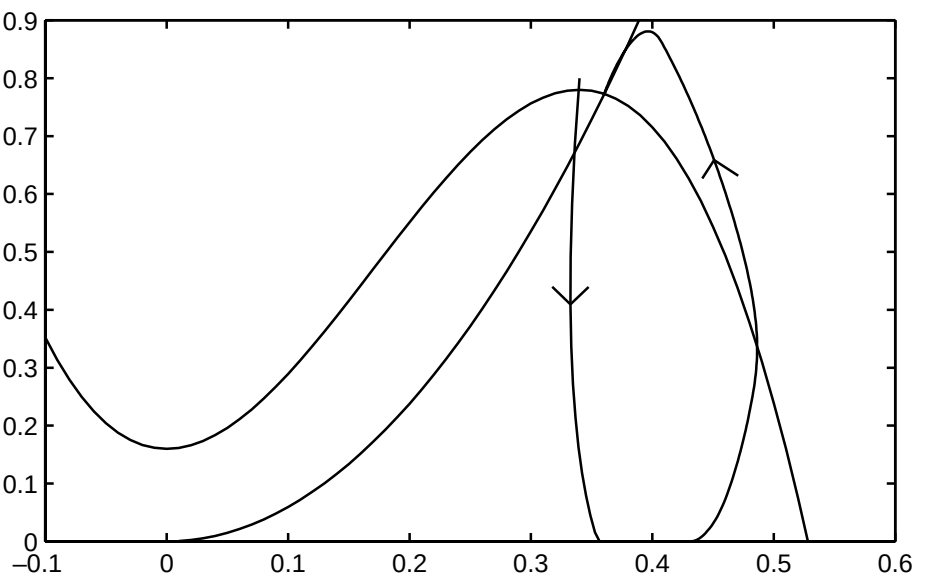


FIG. 6.6. *The* (Φ, Ψ)*-phase plane for the backstepping control. The large disturbance grows into stall, and here the excursion in this phase plane is much larger than before. There is a complete loss of pressure rise over a certain time interval. Design flow, however, is recovered.*

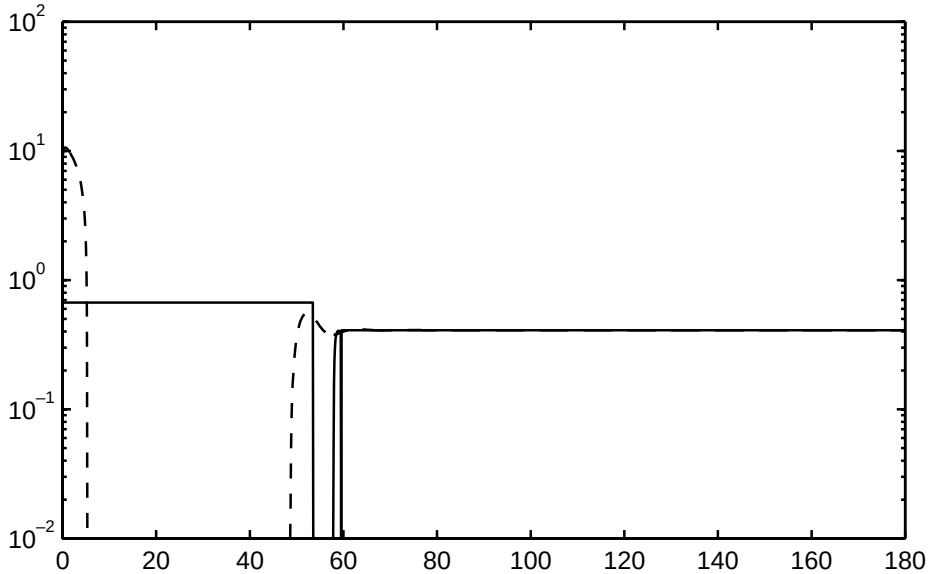


FIG. 6.7. *Here we compare on a logarithmic scale the control effort by the two strategies. The backstepping control result is shown in a dashed line and our control in a solid line. The backstepping control recovers design flow considerably faster than the basic control. Both control strategies saturate which indicates how hard it is to control surge.*

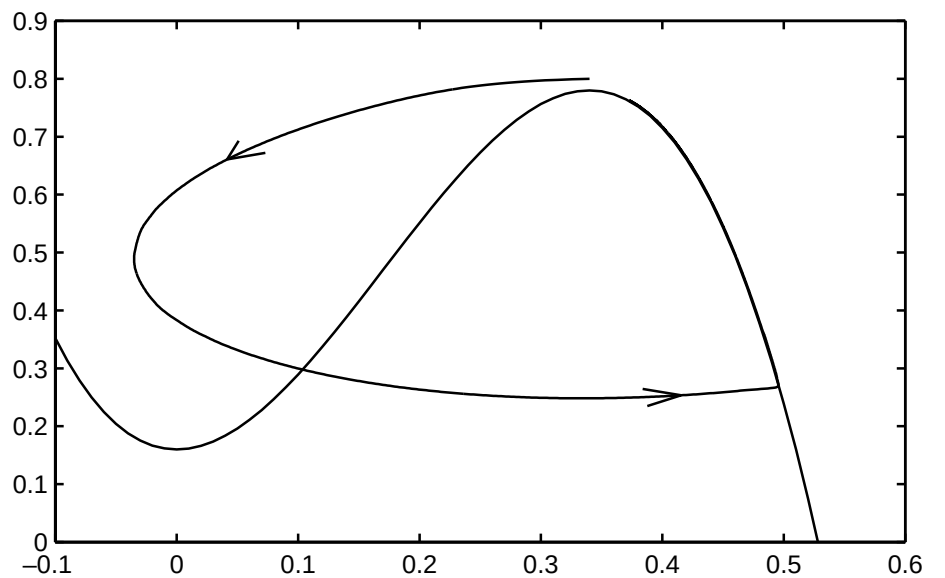


FIG. 6.8. *The* (Φ, Ψ)*-plane when controlling a surging compressor with the basic control.*

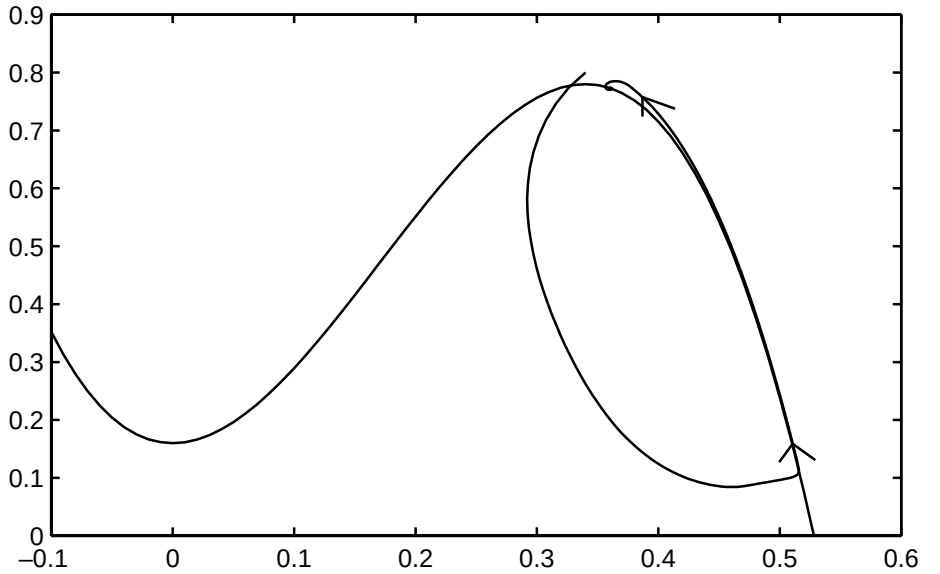


FIG. 6.9. *The* (Φ, Ψ)*-plane when controlling a surging compressor with the backstepping control.*

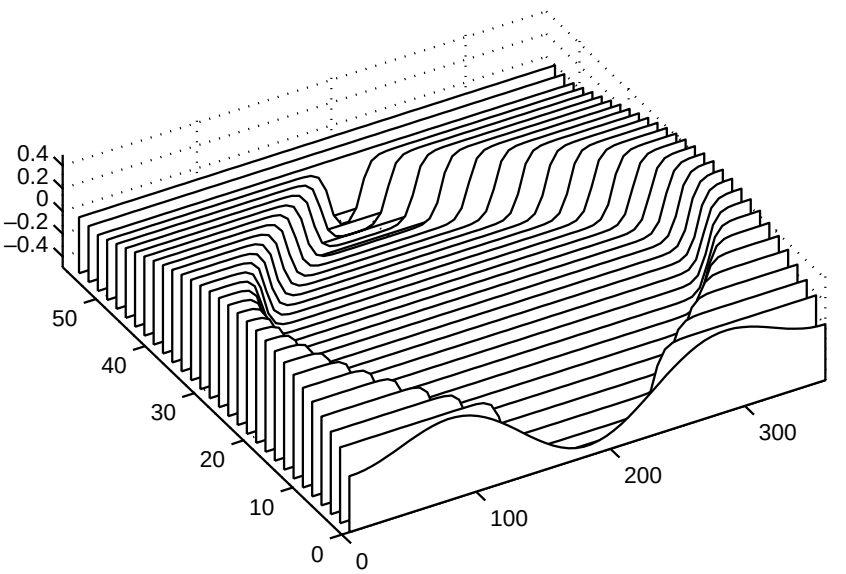


FIG. 6.10. *The transient behavior in* φ *as the basic controller recovers design flow.*

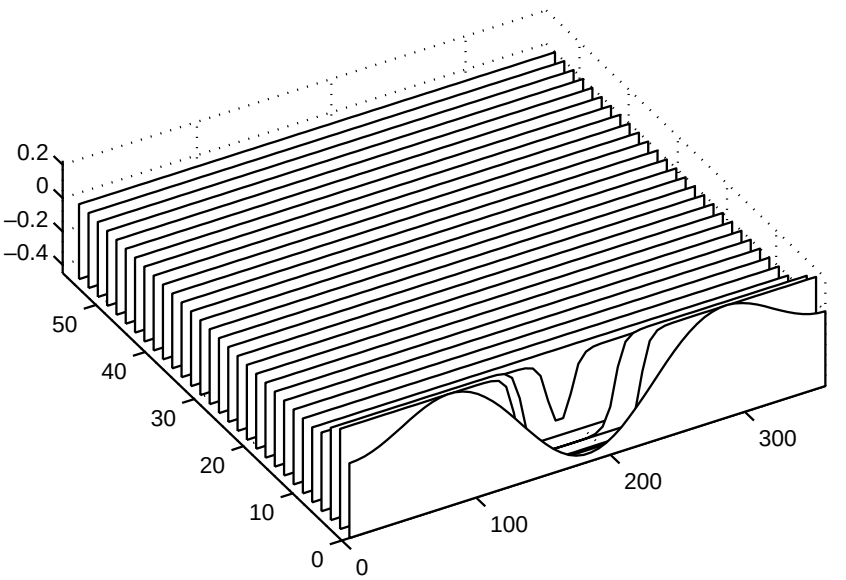


FIG. 6.11. *The transient behavior in φ as the backstepping controller recovers design flow.*

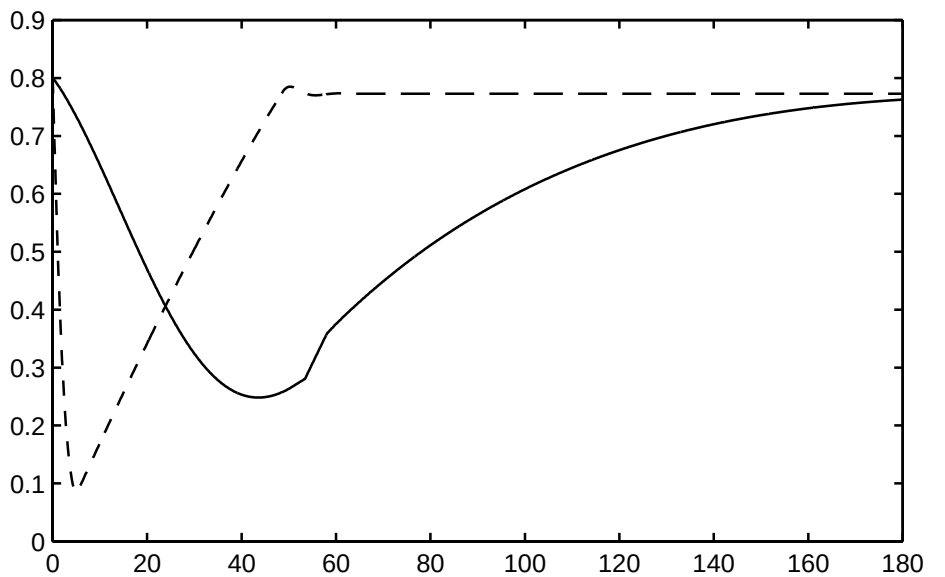


FIG. 6.12. *Here we compare the pressure rise delivered by the compressor under the two strategies during surge instability. The backstepping control result is shown in a dashed line and the basic control in a solid line. The backstepping control kills the disturbance in a shorter time than it takes the basic control to do so.*

7. Basic controllability of the vMG equation. We consider here the problem of when the vMG equation is b-controllable. Although this problem is not of great practical importance, it is more important theoretically to show that there exists at least one example of a system which is b-controllable.

Here above we have constructed a basic feedback control for design flow. In order to have b-controllability, it suffices to show that we can get from any trajectory of the attractor $\mathcal{A}$ to any component in the basic attractor. However, the next theorem says that this is impossible with only throttle control.

THEOREM 7.1. *If the global attractor of the vMG equation contains more than one stall component, then the vMG equation with throttle control only is not b-controllable.*

Proof. The only way we can influence the stall part of the equation, with only throttle control at our disposal, is through Φ (see (1.1)). Consider now a solution of the vMG equation such that $\varphi(0,\theta) = k\sin\theta$ for some nonzero constant k. If we consider only (1.1), then this is a time dependent parabolic equation for which it is well known that the number of local maxima and local minima is nonincreasing in time. It is therefore not possible to start with this initial condition and use throttle control to get to a stall component which has more than one local maximum and one local minimum. In fact, one can show that this is true for all initial conditions close enough to the one chosen here, so the vMG equation is not b-controllable with throttle control. □

Assume now that the parameter B is small enough that there is no surge component in the basic attractor [7], and further assume that we can influence the flow in a nonaxisymmetric way by air injection or bleeding in addition to throttle control. We will avoid the difficult issue of exactly how to model the air injection but will instead make the following simple hypothesis.

Hypothesis 1. If the flow at time t is axisymmetric, i.e., $\varphi(t,\cdot) = 0$, then by use of air injectors or bleeding we can modify the flow so that for $j = 1,\dots,n$ and some small $\epsilon > 0$, $\varphi(t+\epsilon,\cdot)$ will be small but nonzero, $2\pi/j$ periodic, and will have exactly j local maxima and minima.

We have already proven in Theorem 5.1 that there exists a basic feedback control, which uses only throttle, that makes all initial conditions converge asymptotically to the stable design flow component $\mathcal{B}_0$. To prove b-controllability, it therefore suffices to construct a set of controls that take us from the design flow to any one of the stall components $\mathcal{B}_1,\dots,\mathcal{B}_n$. Recall that if the basic attractor has n stall components, then as one decreases the throttle γ the design flow undergoes n supercritical Hopf bifurcations which give rise to the stall solutions [7], [8], [23].

THEOREM 7.2. *Let the basic attractor of the Moore–Greitzer equation consist of design flow and n stall solutions $\mathcal{B}_1,\dots,\mathcal{B}_n$. Assume* Hypothesis 1 *and that we have throttle control. Then the vMG equation is b-controllable.*

Proof. We start with the stable design flow $\mathcal{B}_0$. By Theorem 5.1 there exists a throttle control strategy that makes all solutions in a bounded set $M \subset X$ converge to $\mathcal{B}_0$. Now we slowly decrease the throttle setting to $\gamma = \gamma_*$. By decreasing the throttle γ beyond the point where the design flow undergoes the jth Hopf bifurcation, the jth harmonic becomes unstable. For this throttle setting, design flow is unstable and hence not in the basic attractor, and the system must go into stall. In other words, the component $\mathcal{B}_0$ of the basic attractor has bifurcated to the unstable component $\mathcal{C}_0$ of the reminder (see Theorem 2.3). The basic attractor itself now consists of the stable stall cells $\mathcal{B}_i$ (see [8]). The same theorem says that all the minimal basic attractors

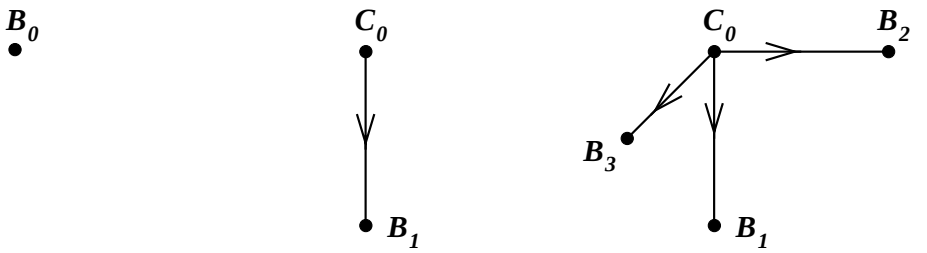


FIG. 7.1. *The global attractor for small B and $\gamma = \gamma_*$. ν decreases from left to right.*

$\mathcal{B}_i$ are connected to the unstable manifolds W_i^u of the unstable design flow $\mathcal{C}_0$ (see Figure 7.1). Now start with design flow $\mathcal{C}_0$, and consider the initial conditions for the deviation $\varphi(0, x)$ from design flow. By using the air injectors or bleeding we can, by Hypothesis 1, make the solution at time ϵ be $2\pi/j$ periodic with exactly j local maxima and j local minima. The rest of the argument concludes that this initial data lie in the jth unstable manifold W_j^u of design flow $\mathcal{C}_0$ and must follow it to the jth stable stall component B_j. Namely, the unstable manifold to the ith stall solution is $2\pi/i$ periodic. Our solution, which is $2\pi/j$ periodic, must settle onto the global attractor, and it can only settle onto the design flow or a stall solution which is $2\pi/k$ periodic, where $k = jm$ for some integer m. However, the number of maxima and minima of $\varphi(t, \cdot)$ is nonincreasing in time (see the proof of Theorem 7.1) and thus m cannot be greater than one. Furthermore, the design flow is unstable with respect to $2\pi/j$ periodic disturbances so the solution cannot settle onto design flow. The only possibility remaining is that the solution settle onto the jth stall solution $\mathcal{B}_j$.

We have now constructed a control which takes the stable design flow $\mathcal{B}_0$ to $\mathcal{B}_j$; in particular, it takes the stable design flow to an open set U in the basin of attraction of $\mathcal{B}_j$. By continuity this control will take a small neighborhood V of the stable design flow into U. We have already constructed in Theorem 5.1 a throttle control which guarantees that for any initial condition we can get from any bounded set $M \subset X$ into V in finite time. We can thus get into U in finite time, and the proof is complete. □

8. Conclusion. We defined b-controllability and a-controllability and presented arguments why these would be meaningful definitions of controllability for infinite dimensional nonlinear dynamical systems. We proved that the vMG equation with throttle control is not b-controllable, and we showed how the control can be modified, by including air injection or bleeding, to make the vMG equation b-controllable.

The backstepping control presented by Banaszuk, Hauksson, and Mezić was the first attempt at constructing a control strategy for the Moore–Greitzer PDE. The vMG equation, which is a better physical model for the airflow through the compression system, has different asymptotic dynamics than the Moore–Greitzer equation, and these asymptotic dynamics have been analyzed by the authors in [7], [6], and [8]. Here we go one step further and use the knowledge of the asymptotic dynamics to construct a control strategy that utilizes the dynamics and hence needs considerably less control effort.

Given a good knowledge of the asymptotic dynamics, one could create a host of different control strategies that would recover design flow operation from stall and surge. By using the dynamics one can considerably reduce the control effort, and it is in this sense that we say that our control construction is near optimal. In fact,

one could use this knowledge to modify the backstepping controller by Banaszuk, Hauksson, and Mezić in that it does not need to react to stall in such a forceful manner.

We have shown here that the basic control does a good job of recovering design flow from stall disturbances. It does not do it as quickly as the backstepping control but with a much smaller control effort, and in this sense the basic control is a much more realistic strategy. Neither of the controls does a good job of recovering design flow from surge disturbances. This is due to the fact that surge disturbances are of a more violent nature and are harder to control. Both controls saturate in this case.

Recall that the control constructed here is mainly intended for recovering design flow operation after large disturbances have occurred. Ideally it would be coupled together with a regulator that would keep the state near design flow during normal operation, and then when large disturbances occur the near optimal control would take over. It is also important to mention that we do not consider the control constructed here as the best possible control. We would however argue that a good knowledge of the basic attractor of the equation is essential to construct a good control strategy. Furthermore, we believe that this approach of analyzing the basic attractor and using the knowledge of the asymptotic dynamics to construct basic control strategies for b-controllable systems can offer a viable alternative to linearizing high dimensional systems and applying linear optimal control theory to the linearized system.

Acknowledgments. We would like to thank Petar Kokotović, Bassam Bamieh, Dan Fontain, Igor Mezić, and Andrzej Banaszuk for fruitful discussions.

REFERENCES

[1] R. A. Adomaitis and E. H. Abed, *Local nonlinear control of stall inception in axial flow compressors*, in AIAA paper 93–2230, 29th Joint Propulsion Conference and Exhibit, American Institute of Astronautics and Aeronautics, Washington, DC, 1993.

[2] A. Banaszuk, H. A. Hauksson, and I. Mezić, *A backstepping controller for a nonlinear partial differential equation model of compression system instabilities*, SIAM J. Control Optim., 37 (1999), pp. 1503–1537.

[3] S. P. Banks, *State-Space and Frequency-Domain Methods in the Control of Distributed Parameter Systems*, Peter Peregrinus, Ltd., London, UK, 1983.

[4] B. Birnir, *Global attractors and basic turbulence*, in Nonlinear Coherent Structures in Physics and Biology, K. M. Spatschek and F. G. Mertens, eds., NATO ASI Series 329, New York, 1994, pp. 321–334.

[5] B. Birnir and R. Grauer, *The global attractor of the damped and driven sine-Gordon equation*, Comm. Math. Phys., 162 (1994), pp. 539–590.

[6] B. Birnir and H. A. Hauksson, *A finite dimensional attractor of the Moore–Greitzer PDE model*, SIAM J. Appl. Math., 59 (1998), pp. 636–650.

[7] B. Birnir and H. A. Hauksson, *The basic attractor of the viscous Moore-Greitzer equation*, J. Nonlinear Sci., submitted.

[8] B. Birnir and H. A. Hauksson, *The low-dimensional flow on the basic attractor of the viscous Moore-Greitzer equation*, SIAM J. Appl. Math., submitted.

[9] B. Birnir and K. Nelson, *The existence of smooth attractors of damped and driven nonlinear wave equations*, J. Differential Equations, submitted, 1998.

[10] B. D. Coller, *Hopf/Hopf Interaction in Compressor Models*, preprint, California Institute of Technology, Pasadena, CA, 1996.

[11] P. Constantin, C. Foias, B. Nicolaenko, and R. Temam, *Integral Manifolds and Inertial Manifolds for Dissipative Partial Differential Equations*, Springer, New York, Berlin, 1989.

[12] I. J. Day, *Axial compressor performance during surge*, J. Propulsion and Power, 10 (1994), pp. 329–336.

[13] A. Debussche and R. Temam, *Inertial manifolds and the slow manifolds in meteorology*, Differential Integral Equations, 4 (1991), pp. 897–931.

[14] J. F. Escuret and V. Garnier, *Numerical simulations of aerodynamic flow instabilities in*

axial compressors, in Von Karman Institute for Fluid Dynamics 1995–1996 Lecture Series 1, Rhode-St.-Genese, Belgium.

[15] E. M. Greitzer, *Surge and rotating stall in axial flow compressors, Part* II, in Trans. ASME J. Engineering for Power, 98 (1976), pp. 190–217.

[16] H. A. Hauksson, *The Basic Attractor of the Viscous Moore-Greitzer Equations*, Ph.D. thesis, University of Michigan, Ann Arbor, MI, 1998.

[17] B. R. Hunt, T. Sauer, and J. A. Yorke, *Prevalence: A translation-invariant "almost every" on infinite-dimensional spaces*, Bull. Amer. Math. Soc., 27 (1992), pp. 217–238.

[18] T. Kato, *Perturbation Theory for Linear Operators*, Springer, Berlin, 1980.

[19] J. L. Lions, *Optimal Control of Systems Governed by Partial Differential Equations*, Springer, New York, Berlin, 1971.

[20] C. A. Mansoux, D. L. Gysling, J. D. Setiawan, and J. D. Paduano, *Distributed nonlinear modeling and stability analysis of axial compressor stall and surge*, in Proceedings of the American Control Conference, Baltimore, MD, American Automatic Control Council, Evanston, IL, 1994, pp. 2305–2316.

[21] M. Marion and R. Temam, *Nonlinear Galerkin methods*, J. Numer. Anal., 5 (1989), pp. 1139–1157.

[22] F. E. McCaughan, *Numerical results for axial flow compressor instability*, ASME J. Turbomachinery, 111 (1989), pp. 434–441.

[23] F. E. McCaughan, *Bifurcation analysis of axial flow compressor stability*, SIAM J. Appl. Math., 50 (1990), pp. 1232–1253.

[24] I. Mezić, *A Large-Scale Theory of Axial Compression System Dynamics*, preprint, University of California, Santa Barbara, CA, 1998.

[25] J. Milnor, *On the concept of attractor*, Comm. Math. Phys., 99 (1985), pp. 177–195.

[26] F. K. Moore and E. M. Greitzer, *A theory of post-stall transients in axial compression systems: Part* I - *Development of equations*, ASME J. Engineering for Gas Turbines and Power, 108 (1986), pp. 68–76.

[27] V. A. Pliss and G. R. Sell, *Approximation dynamics and the stability of invariant sets*, IMA preprint 1393, Institute of Mathematics and its Applications, Minneapolis, MN, 1996.

[28] A. P. Sage and C. C. White, *Optimum Systems Control*, 2nd ed., Prentice–Hall, Englewood Cliffs, NJ, 1977.

[29] P. Du Val, *Elliptic Functions and Elliptic Curves*, Cambridge University Press, London, New York, 1973.

[30] A. G. Wilson and C. Freeman, *Stall inception and development in an axial flow aeroengine*, Trans. ASME, J. Turbomachinery, 116 (1994), pp. 216–225.

SIAM J. CONTROL OPTIM.
Vol. 38, No. 5, pp. 1581–1602

ASYMPTOTIC STABILITY AND ENERGY DECAY RATES FOR SOLUTIONS OF THE WAVE EQUATION WITH MEMORY IN A STAR-SHAPED DOMAIN*

M. AASSILA†, M. M. CAVALCANTI‡, AND J. A. SORIANO‡

Abstract. We study the asymptotic stability and give the energy decay rates for solutions of the wave equation with boundary dissipation of the memory type.

Key words. wave equation, boundary condition of memory type

AMS subject classifications. 35L05, 35L20, 35B37

PII. S0363012998344981

1. Introduction.

A basic linear model for the evolution of sound in a compressible fluid is the system of partial differential equations

$$\rho v_t(t,x) + \operatorname{grad} p(x,t) = 0, \tag{1.1}$$

$$\alpha p_t(t,x) + \operatorname{div} v(t,x) = 0, \quad t > 0, \ \ x \in \mathbf{R}^n, \tag{1.2}$$

where p denotes acoustic pressure and v the velocity field; cf., e.g., Leis [9]. In what follows the equilibrium density ρ and the compressibility α will be assumed to be constant and then without loss of generality to be equal to 1. Eliminating v from this system one obtains a wave equation for the pressure p:

$$p_{tt}(t,x) = \Delta p(t,x), \quad t > 0, \ \ x \in \mathbf{R}^n. \tag{1.3}$$

When the fluid is enclosed in a region $\Omega \subset \mathbf{R}^n$, (1.3) has to be supplemented by conditions at $\partial\Omega$, the boundary of Ω. The following three dissipative boundary conditions are discussed in the mathematical literature on time domain models for acoustics.

First, equating the acoustic impedance $\xi(x) \in \mathbf{C}$ of the boundary surface at x with the ratio between the fluid's pressure and its velocity normal to the surface results in

$$\frac{\partial p}{\partial \nu}(t,x) + \xi(x) p_t(t,x) = 0, \quad t > 0, \ \ x \in \partial\Omega, \tag{1.4}$$

where $\frac{\partial}{\partial \nu}$ denotes the derivative in the direction of the outer normal of $\partial\Omega$. The well-posedness of the equation with boundary conditions (1.4) and the asymptotic behavior of its solutions has been investigated in [3], [4], [15].

Second, adding a friction term $\beta(x)p_t(t,x)$, $\beta > 0$, to the classic Robin condition yields

$$\frac{\partial p}{\partial \nu}(t,x) + \beta(x) p_t(t,x) + \alpha(x) p(x,t) = 0, \quad t > 0, \ \ x \in \partial\Omega. \tag{1.5}$$

*Received by the editors September 21, 1998; accepted for publication (in revised form) November 15, 1999; published electronically May 26, 2000.
http://www.siam.org/journals/sicon/38-5/34498.html

†Institut de Recherche Mathématique Avancée, 7 Rue René Descartes, 67084 Strasbourg cédex, France.

‡Departamento de Matemática, Universidade Estadual de Maringá, 87020-900 Maringá, PR, Brasil (marcelo@gauss.dma.uem.br, soriano@gauss.dma.uem.br).

This condition has been studied, e.g., in [8].

Third, modeling the boundary surface as independent oscillations and equating the velocity δ_t of the impenetrable surface with the normal velocity of the fluid at boundary points leads to

$$n(x)\delta_{tt}(t,x) + d(x)\delta_t(t,x) + k(x)\delta(t,x) = -p(t,x),$$

$$\frac{\partial p}{\partial \nu}(t,x) + \delta_{tt}(t,x) = 0, \quad t > 0, \ \ x \in \partial\Omega. \tag{1.6}$$

In [2], where this boundary model is formulated for the velocity potential, spectral properties of the generator of the solution semigroup are given.

Looking for more general results, we find in [11, Equation (6.3.11)] that the pressure of the combination of a wave $F(T_i)$, $T_i = t - (x_1 \sin\theta - x_2 \cos\theta)$, that is incident at angle θ onto the surface $x_2 = 0$, with the reflected wave in direction $T_r - t = -(x_1 \cos\theta + x_2 \cos\theta)$, is of the form

$$p(t,x) = F(T_i) + F(T_r) + \int_{-\infty}^{+\infty} F(\tau) W(T_r - \tau)\, d\tau. \tag{1.7}$$

Here W represents the modification of the reflected wave that is caused by the motion of the surface. This means that a general linear reflection process is to be modeled by convolution of the acoustic wave with a function that characterizes the boundary material. In order to cast (1.4)–(1.6) into a common form, we write

$$\frac{\partial p}{\partial \nu}(t,x) + dk * p_t(t,x) = 0, \quad t > 0, \ \ x \in \partial\Omega, \tag{1.8}$$

where $dk * p_t(t,x) = \int_0^t dk(\tau,x) p_t(t-\tau,x)$.

Another approach that also leads to convolution boundary conditions of the form (1.8) is the modeling of the boundary as the surface of a viscoelastic material.

In this paper, we are concerned more precisely with the asymptotic behavior of solutions to the following problem:

$$u'' - \Delta u = 0 \quad \text{in } \Omega \times \mathbf{R}_+, \tag{1.9}$$

$$u = 0 \quad \text{on } \Gamma_0 \times \mathbf{R}_+, \tag{1.10}$$

$$\frac{\partial u}{\partial \nu} + \int_0^t k(t-s,x) u'(s)\, ds + a(x) g(u') = 0 \quad \text{on } \Gamma_1 \times \mathbf{R}_+, \tag{1.11}$$

$$u(x,0) = u_0(x), \quad u'(x,0) = u_1(x), \quad x \in \Omega, \tag{1.12}$$

where $\Omega \subset \mathbf{R}^n$ is an open bounded domain with boundary $\Gamma = \Gamma_0 \cup \Gamma_1$ of class C^2, $a : \Gamma_1 \to \mathbf{R}_+ \in L^\infty(\Gamma_1)$ is such that $a(x) \geq a_0 > 0$, $k : \mathbf{R}_+ \times \Gamma_1 \to \mathbf{R}_+ \in C^2(\mathbf{R}_+, L^\infty(\Gamma_1))$, and $g : \mathbf{R} \to \mathbf{R}$ is a continuous nondecreasing function such that $g(0) = 0$ and $|g(x)| \leq 1 + C|x|$, $C > 0$.

Problems related to (1.9)–(1.12) were studied by many authors, e.g., Muñoz Rivera [12], Tadayuki and Rinko [17], Prüss [14], Dix and Torrejon [5], Guesmia [6], Propst and Prüss [13], Renardy, Hrusa, and Nohel [16], and the references therein.

Our paper is organized as follows. In section 2, we state our main results. In section 3 we give the proofs, and in section 4 we study the strong asymptotic stability under weak assumptions on Ω and g.

2. Main results. The following hypotheses are made on Ω and on the functions k and g:

$$\Gamma_0 \neq \emptyset \quad \text{or} \quad \inf_{\Gamma_1 \times \mathbf{R}_+} k \neq 0, \tag{2.1}$$

$$m \cdot \nu \geq \delta > 0 \quad \text{on } \Gamma_1, \quad m \cdot \nu \leq 0 \quad \text{on } \Gamma_0, \quad m(x) = x - x^0 (x^0 \in \mathbf{R}^n), \tag{2.2}$$

$$k \geq 0 \quad \text{on } \Gamma_1 \times \mathbf{R}_+, \tag{2.3}$$

$$k' \leq 0 \quad \text{on } \Gamma_1 \times \mathbf{R}_+, \tag{2.4}$$

$$\exists \alpha > 0, \quad k'' \geq -\alpha k' \quad \text{on } \Gamma_1 \times \mathbf{R}_+, \tag{2.5}$$

$$C_1 |x|^p \leq |g(x)| \leq C_2 |x|^{1/p} \quad \text{if } |x| \leq 1, \tag{2.6}$$

$$C_3 |x| \leq |g(x)| \leq C_4 |x| \quad \text{if } |x| \geq 1, \tag{2.7}$$

where $p \geq 1$ and $C_i \, (1 \leq i \leq 4)$ are four positive constants.

Remark 2.1. (1) An example of function k satisfying (2.3)–(2.5) is

$$k(t,x) = f(x)\, e^{-\alpha t} + g(x) \quad \text{on } \Gamma_1 \times \mathbf{R}_+,$$

where $f, g \in L^\infty(\Gamma_1, \mathbf{R}_+)$.

(2) The condition (2.1) implies that the formula $\int_\Omega |\nabla u|^2 \, dx + \int_{\Gamma_1} k \, |u|^2 \, d\Gamma$ defines a norm on $H^1(\Omega)$ equivalent to the usual one.

(3) The condition (2.2) implies that $\overline{\Gamma}_0 \cap \overline{\Gamma}_1 = \emptyset$, and Ω is star-shaped with respect to x^0. It could be weakened, as was done in [8].

For the sake of completeness, we give a brief outline of the well-posedness of problem (1.9)–(1.12). This problem could be rewritten in the following form:

$$\langle u'', v \rangle + \int_0^t d\beta \, (t - \tau, u'(\tau), v) + \alpha(u, v) + \int_{\Gamma_1} a(x) g(u') v \, d\Gamma = 0, \tag{2.8}$$

$$u(0) = u_0, \quad u'(0) = u_1, \tag{2.9}$$

where

$$\beta(t, u, v) = \int_\Gamma k(t,x) u(x) v(x) \, d\Gamma, \quad \beta(0, u, v) = 0, \quad t > 0, \quad u, v \in H^1_{\Gamma_0}(\Omega),$$

$$H^1_{\Gamma_0}(\Omega) = \left\{ u \in H^1(\Omega); \ \ u = 0 \quad \text{on } \Gamma_0 \right\},$$

$$\alpha(u, v) = \int_\Omega \nabla u \cdot \nabla v \, dx, \quad u, v \in H^1_{\Gamma_0}(\Omega).$$

The problem (2.8)–(2.9) could be reformulated as an evolutionary integral equation of variational type, and thanks to the monotonicity and to the growth condition $|g(x)| \leq 1 + C|x|$ assumed on the function g, we can have access to the results and methods developed in section 6 of Prüss [14] to obtain the following theorem.

THEOREM 2.1. *For all given initial data* $(u_0, u_1) \in H^1_{\Gamma_0}(\Omega) \times L^2(\Omega)$, *problem* (1.9)–(1.12) *admits a unique global weak solution*

$$u \in C\left(\mathbf{R}_+, H^1_{\Gamma_0}(\Omega)\right) \cap C^1\left(\mathbf{R}_+, L^2(\Omega)\right). \tag{2.10}$$

Furthermore, if $(u_0, u_1) \in \left(H^2(\Omega) \cap H^1_{\Gamma_0}(\Omega)\right) \times H^1_{\Gamma_0}(\Omega)$ *and* g *is globally Lipschitz continuous, then the solution has the following regularity:*

$$u \in C\left(\mathbf{R}_+, H^2(\Omega)\right) \cap C^1\left(\mathbf{R}_+, H^1_{\Gamma_0}(\Omega)\right) \cap C^2\left(\mathbf{R}_+, L^2(\Omega)\right). \tag{2.11}$$

We define the energy of the solution given by (2.10) by the following formula:

$$E(t) := \frac{1}{2}\int_\Omega \left(|u'(t,x)|^2 + |\nabla u(t,x)|^2\right) dx + \frac{1}{2}\int_{\Gamma_1} k(t,x)\,|u(t,x) - u_0(x)|^2\, d\Gamma \tag{2.12}$$

$$-\frac{1}{2}\int_{\Gamma_1}\int_0^t k'(t-s,x)\,|u(t,x) - u(s,x)|^2\, ds\, d\Gamma.$$

Our main results are the following.

THEOREM 2.2. *Assume that hypotheses* (2.1)–(2.7) *hold and that*

$$\alpha \inf_{\Gamma_1} k(0) > -2 \inf_{\Gamma_1} k'(0). \tag{2.13}$$

Then we have

$$E(t) \leq CE(0)e^{-\omega t} \quad if \;\; p = 1, \;\; C > 0, \;\; \omega > 0, \;\; t \geq 0, \tag{2.14}$$

$$E(t) \leq \frac{CE(0)}{(1+t)^{2/(p-1)}} \quad if \;\; p > 1, \;\; C > 0, \;\; t \geq 0 \tag{2.15}$$

for every weak solution to (1.9)–(1.12) *and initial data* $(u_0, u_1) \in H^1_0(\Omega) \times L^2(\Omega)$.

Remark 2.2. Theorem 2.2 has a serious drawback: it never can be applied for bounded functions g (because of $C_3 > 0$ in (2.7)). The purpose of the following theorem is to obtain a variant of Theorem 2.2 for bounded feedback functions.

THEOREM 2.3. *Assume* (2.1)–(2.5) *and assume that the function* g *is bounded and globally Lipschitz continuous and that the inequalities* (2.6) *are satisfied with some positive constants* C_1, C_2 *and with a number* p *satisfying*

$$p \geq n - 1. \tag{2.16}$$

Then for every strong solution u *to* (1.9)–(1.12) *we have*

$$E(t) \leq \frac{CE(0)}{(1+t)^{2/(p-1)}}, \quad C > 0, \;\; t \geq 0. \tag{2.17}$$

To end this section, we recall the following useful lemma.

LEMMA 2.4 (see [7, Lemma 9.1]). *Let $E : \mathbf{R}_+ \to \mathbf{R}_+$ be a nonincreasing function and assume that there exist two constants $\alpha > 0$ and $T > 0$ such that*

$$\int_t^{+\infty} E^{\alpha+1}(s)\,ds \le TE(0)^{\alpha}E(t) \quad for\ all\ \ t \in \mathbf{R}_+. \tag{2.18}$$

Then we have

$$E(t) \le E(0)\left(\frac{T+\alpha t}{T+\alpha T}\right)^{-1/\alpha} \quad for\ all\ \ t \ge T. \tag{2.19}$$

3. Proofs. To justify all the computations that follow, we will assume that u is a strong solution and that by a classical density argument the results of Theorem 2.2 still hold for the weak solution.

To simplify the computations and without lost of generality, we will transform the boundary condition (1.11) in another more practical one considering $u_0 = 0$ on Γ_1.

A simple integration by parts yields

$$\int_0^t k(t-s,x)u'(s,x)\,ds$$

$$= [k(t-s,x)u(s,x)]_0^t + \int_0^t k'(t-s,x)u(s,x)\,ds$$

$$= \int_0^t k'(t-s,x)u(s,x)\,ds + k(0,x)u(t,x).$$

Hence, the problem (1.9)–(1.12) is now transformed into

$$u'' - \Delta u = 0 \quad \text{in } \Omega \times \mathbf{R}_+, \tag{3.1}$$

$$u = 0 \quad \text{on } \Gamma_0 \times \mathbf{R}_+, \tag{3.2}$$

$$\frac{\partial u}{\partial \nu} + \int_0^t k'(t-s,x)u(s,x)\,ds + k(0)u + a(x)g(u') = 0 \quad \text{on } \Gamma_1 \times \mathbf{R}_+, \tag{3.3}$$

$$u(x,0) = u_0(x), \quad u'(x,0) = u_1(x) \quad \text{in } \Omega. \tag{3.4}$$

LEMMA 3.1. *The energy defined by* (2.12) *is nonincreasing and it holds that*

$$\frac{1}{2}\int_S^T \int_{\Gamma_1} \int_0^t k''(t-s)\left(u(t)-u(s)\right)^2 ds\,d\Gamma\,dt + \int_S^T \int_{\Gamma_1} a(x)g(u')u'\,d\Gamma\,dt$$

$$-\frac{1}{2}\int_S^T \int_{\Gamma_1} k'\,|u|^2\,d\Gamma\,dt = E(S) - E(T) \le E(S)$$

for every $0 \le S < T < +\infty$. In particular, $\int g(u')u' \le c|E'(0)|$ for a suitable constant c.

Proof.

$$E(T) - E(S) = \int_S^T E'(t)\,dt$$

$$= \int_S^T \int_\Omega (u'u'' + \nabla u \cdot \nabla u')\,dx\,dt + \int_S^T \int_{\Gamma_1} kuu'\,d\Gamma\,dt + \frac{1}{2}\int_S^T \int_{\Gamma_1} k' |u|^2\,d\Gamma\,dt$$

$$- \int_S^T \int_{\Gamma_1} \int_0^t k'(t-s)\,(u(t) - u(s))\,u'(t)\,ds\,d\Gamma\,dt$$

$$-\frac{1}{2}\int_S^T \int_{\Gamma_1} \int_0^t k''(t-s)\,(u(t) - u(s))^2\,ds\,d\Gamma\,dt.$$

Using (3.1)–(3.3) and the Green formula, we obtain

$$E(T) - E(S) = -\int_S^T \int_{\Gamma_1} a(x) g(u') u'\,d\Gamma\,dt + \frac{1}{2}\int_S^T \int_{\Gamma_1} k' |u|^2\,d\Gamma\,dt$$

$$-\frac{1}{2}\int_S^T \int_{\Gamma_1} \int_0^t k''(t-s)\,(u(t) - u(s))^2\,ds\,d\Gamma\,dt$$

$$+ \int_S^T \int_{\Gamma_1} uu' \left(k - k(0) - \int_0^t k'(t-s)ds\right) d\Gamma\,dt.$$

Hence

$$E(S) - E(T)$$

$$= \int_S^T \int_{\Gamma_1} a(x) g(u') u'\,d\Gamma\,dt + \frac{1}{2}\int_S^T \int_{\Gamma_1} \int_0^t k''(t-s)\,(u(t) - u(s))^2\,ds\,d\Gamma\,dt$$

$$-\frac{1}{2}\int_S^T \int_{\Gamma_1} k' |u|^2\,d\Gamma\,dt. \qquad \square$$

LEMMA 3.2. *Setting*

$$Mu := 2(m \cdot \nabla u) + (n-1)u,$$

it holds that

$$\int_S^T E^{\frac{p-1}{2}} \int_\Omega \left(|u'|^2 + |\nabla u|^2\right) dx\,dt$$

$$= -\left[E^{\frac{p-1}{2}} \int_\Omega u'(Mu)\,dx\right]_S^T + \frac{p-1}{2}\int_S^T E^{\frac{p-3}{2}} E' \int_\Omega u'(Mu)\,dx\,dt$$

$$+ \int_S^T E^{\frac{p-1}{2}} \int_{\Gamma_0} (m \cdot \nu) |\nabla u|^2\,d\Gamma\,dt + \int_S^T E^{\frac{p-1}{2}} \int_{\Gamma_1} (m \cdot \nu) \left(|u'|^2 - |\nabla u|^2\right) d\Gamma\,dt$$

$$+ \int_S^T E^{\frac{p-1}{2}} \int_{\Gamma_1} (Mu) \frac{\partial u}{\partial \nu}\,d\Gamma\,dt.$$

Proof. We have

$$0 = \int_S^T E^{\frac{p-1}{2}} \int_\Omega (Mu)\,(u'' - \Delta u)\; dx\, dt$$

$$= \left[E^{\frac{p-1}{2}} \int_\Omega u'(Mu)\, dx\right]_S^T - \frac{p-1}{2} \int_S^T E^{\frac{p-3}{2}} E' \int_\Omega u'(Mu) dx\, dt$$

$$- \int_S^T E^{\frac{p-1}{2}} \int_\Omega (u'(Mu') + (Mu)\Delta u)\, dx\, dt,$$

and

$$\int_\Omega (u'(Mu') + (Mu)\Delta u)\, dx$$

$$= \int_\Omega \left(m \cdot \nabla (u')^2 + (n-1)\,|u'|^2 - \nabla u \cdot \nabla (Mu)\right) dx + \int_\Gamma (Mu) \frac{\partial u}{\partial \nu} d\Gamma$$

$$= \int_\Omega \left(m \cdot \nabla (u')^2 + (n-1)\,|u'|^2 - 2\,|\nabla u|^2 - m \cdot \nabla\,|\nabla u|^2 - (n-1)\,|\nabla u|^2\right) dx$$

$$+ \int_\Gamma (Mu) \frac{\partial u}{\partial \nu} d\Gamma$$

$$= - \int_\Omega \left(|u'|^2 + |\nabla u|^2\right) dx + \int_\Gamma \left((m \cdot \nu)\left(|u'|^2 - |\nabla u|^2\right) + (Mu)\frac{\partial u}{\partial \nu}\right) d\Gamma$$

$$= - \int_\Omega \left(|u'|^2 + |\nabla u|^2\right) dx + \int_{\Gamma_0} \left(-(m \cdot \nu)\,|\nabla u|^2 + (2m \cdot \nabla u)\frac{\partial u}{\partial \nu}\right) d\Gamma$$

$$+ \int_{\Gamma_1} (m \cdot \nu) \left((|u'|^2 - |\nabla u|^2) + (Mu)\frac{\partial u}{\partial \nu}\right) d\Gamma\,.$$

Since (3.2) implies that $\nabla u = \frac{\partial u}{\partial \nu}\nu$ on Γ_0, we obtain the desired result. ∎

LEMMA 3.3. *It holds that*

$$\int_S^T E^{\frac{p-1}{2}} \int_\Omega \left(|u'|^2 + |\nabla u|^2\right) dx\, dt \le C(\varepsilon)E(S) + \varepsilon \int_S^T E^{\frac{p+1}{2}}(t)\, dt + C(\varepsilon)E(S)$$

$$+(n-1) \int_S^T E^{\frac{p-1}{2}} \int_{\Gamma_1} u \frac{\partial u}{\partial \nu} d\Gamma\, dt + \frac{R^2}{\delta} \int_S^T E^{\frac{p-1}{2}} \int_{\Gamma_1} \left(\frac{\partial u}{\partial \nu}\right)^2 d\Gamma\, dt,$$

where $0 \le S < T < +\infty$, $R = ||m||_{L^\infty(\Omega)}$, C *is a positive constant,* ε *is an arbitrary small real number, and* δ *comes from* (2.2).

Proof. Here and in what follows C will denote various positive constants which may be different at different occurences.

From the nonincreasingness of the energy, we deduce that

$$\left| E^{\frac{p-1}{2}} \int_\Omega u' M(u)\, dx \right| \le CE,$$

$$\left| E^{\frac{p-3}{2}} E' \int_\Omega u'(Mu)\, dx \right| \le -CE^{\frac{p-1}{2}} E' \le -C\left(E^{\frac{p+1}{2}}\right)',$$

and

$$\int_S^T E^{\frac{p-3}{2}} E' \int_\Omega u'(Mu) dx\, dt \le CE(S).$$

Thanks to the Young inequality, we have

$$2\int_S^T E^{\frac{p-1}{2}} \int_{\Gamma_1} \frac{\partial u}{\partial \nu}\,(m\cdot\nabla u)\, d\Gamma\, dt$$

$$\le \delta \int_S^T E^{\frac{p-1}{2}} \int_{\Gamma_1} |\nabla u|^2\, d\Gamma\, dt + \frac{R^2}{\delta} \int_S^T E^{\frac{p-1}{2}} \int_{\Gamma_1} \left(\frac{\partial u}{\partial \nu}\right)^2 d\Gamma\, dt,$$

and by the relations (2.6)–(2.7), we obtain

$$\int_{|u'|\le 1} (m\cdot\nu)\,|u'|^2\, d\Gamma \le C \int_{|u'|\le 1} (m\cdot\nu)\,(u'g(u'))^{\frac{2}{p+1}}\, d\Gamma$$

$$\le C\left(\int_{\Gamma_1} (m\cdot\nu)\, u' g(u')\, d\Gamma\right)^{\frac{2}{p+1}} \le C\,(-E')^{\frac{2}{p+1}}$$

and

$$\int_{|u'|\ge 1} (m\cdot\nu)\,|u'|^2\, d\Gamma \le \int_{|u'|\ge 1} (m\cdot\nu)\, u' g(u') d\Gamma \le -CE'.$$

Hence

$$\int_S^T E^{\frac{p-1}{2}} \int_{|u'|\le 1} (m\cdot\nu)\,|u'|^2\, d\Gamma\, dt \le C\int_S^T E^{\frac{p-1}{2}}\,(-E')^{\frac{2}{p+1}}\, dt$$

$$\le \int_S^T \left(\varepsilon E^{\frac{p+1}{2}} - C(\varepsilon)E'\right) dt$$

$$\le \varepsilon \int_S^T E^{\frac{p+1}{2}}(t)\, dt + C(\varepsilon)E(S), \tag{3.5}$$

and

$$\int_S^T E^{\frac{p-1}{2}} \int_{|u'|\ge 1} (m\cdot\nu)\,|u'|^2\, d\Gamma dt \le CE(S). \tag{3.6}$$

From this, it follows that

$$\int_S^T E^{\frac{p-1}{2}} \int_\Omega \left(|u'|^2 + |\nabla u|^2\right) dx\, dt \le CE(S) + \varepsilon \int_S^T E^{\frac{p+1}{2}}(t)\, dt$$

$$+C(\varepsilon)E(S) + (n-1)\int_S^T E^{\frac{p-1}{2}} \int_{\Gamma_1} u\frac{\partial u}{\partial \nu} d\Gamma\, dt$$

$$+\delta \int_S^T E^{\frac{p-1}{2}} \int_{\Gamma_1} |\nabla u|^2\, d\Gamma\, dt + \frac{R^2}{\delta}\int_S^T E^{\frac{p-1}{2}} \int_{\Gamma_1} \left(\frac{\partial u}{\partial \nu}\right)^2 d\Gamma\, dt$$

$$-\int_S^T E^{\frac{p-1}{2}} \int_{\Gamma_1} (m\cdot\nu)\, |\nabla u|^2\, d\Gamma\, dt,$$

and then we deduce the result by $m \cdot \nu \ge \delta > 0$ on Γ_1.

LEMMA 3.4. *It holds that*

$$-\int_S^T E^{\frac{p-1}{2}} \int_{\Gamma_1} u\frac{\partial u}{\partial \nu} d\Gamma\, dt$$

$$\le CE(S) + \varepsilon \int_S^T E^{\frac{p-1}{2}} \int_\Omega |u'|^2\, dx\, dt + \varepsilon \int_S^T E^{\frac{p+1}{2}}(t) dt + C(\varepsilon)E(S)$$

for every $\varepsilon > 0$, $0 \le S < T < \infty$.

Proof. We will use the idea introduced by Conrad and Rao [4]. Let φ be a solution of

$$-\Delta\varphi = 0 \quad \text{in } \Omega, \qquad \varphi = u \quad \text{on } \Gamma.$$

By the classical results of elliptic partial differential equations theory we have

$$\int_\Omega \nabla\varphi \cdot \nabla u\, dx = \int_\Omega |\nabla\varphi|^2\, dx,$$

$$\int_\Omega |\varphi|^2\, dx \le C\int_\Gamma |u|^2\, d\Gamma,$$

$$\int_\Omega |\varphi'|^2\, dx \le C\int_\Gamma |u'|^2\, d\Gamma.$$

Multiplying (3.1) with $E^{\frac{p-1}{2}}\varphi$ we obtain

$$-\int_S^T E^{\frac{p-1}{2}} \int_\Gamma \varphi\frac{\partial u}{\partial \nu} d\Gamma\, dt = -\left[E^{\frac{p-1}{2}} \int_\Omega u'\varphi\, dx\right]_S^T + \frac{p-1}{2}\int_S^T E^{\frac{p-3}{2}}E' \int_\Omega u'\varphi dx\, dt$$

$$+\int_S^T E^{\frac{p-1}{2}} \int_\Omega u'\varphi' dx\, dt - \int_S^T E^{\frac{p-1}{2}} \int_\Omega \nabla u\cdot\nabla\varphi dx\, dt.$$

It follows that

$$\left[E^{\frac{p-1}{2}}\int_\Omega u'\varphi\,dx\right]_S^T \le CE(S),$$

$$\int_S^T E^{\frac{p-3}{2}}E'\int_\Omega u'\varphi dx\,dt \le CE(S),$$

$$\int_S^T E^{\frac{p-1}{2}}\int_\Omega u'\varphi' dx\,dt$$

$$\le \varepsilon\int_S^T E^{\frac{p-1}{2}}\int_\Omega |u'|^2\,dx\,dt + C(\varepsilon)\int_S^T E^{\frac{p-1}{2}}\int_\Omega |\varphi'|^2\,dx\,dt$$

$$\le \varepsilon\int_S^T E^{\frac{p-1}{2}}\int_\Omega |u'|^2\,dx\,dt + C(\varepsilon)\int_S^T E^{\frac{p-1}{2}}\int_{\Gamma_1} |u'|^2\,d\Gamma\,dt$$

$$\le \varepsilon\int_S^T E^{\frac{p-1}{2}}\int_\Omega |u'|^2\,dx\,dt + C(\varepsilon)E(S) + \varepsilon\int_S^T E^{\frac{p+1}{2}}(t)dt. \qquad \square$$

LEMMA 3.5. *Define*

$$\gamma := \gamma(x) = \frac{\lambda}{k(0)} \quad with \quad \lambda > \max\left\{\frac{n-1}{2}, \frac{R^2}{\delta}\,||k(0)||_{L^\infty(\Gamma_1)}\right\}.$$

Then it holds that

$$\int_S^T E^{\frac{p+1}{2}}(t)\,dt \le CE(S) + (1-\lambda)\int_S^T E^{\frac{p-1}{2}}\int_{\Gamma_1} k(0)\,|u|^2\,d\Gamma\,dt$$

$$+2\int_S^T E^{\frac{p-1}{2}}\int_{\Gamma_1}\gamma\left(\int_0^t k'(t-s)u(s)\,ds\right)^2 d\Gamma\,dt.$$

Proof. Thanks to (3.3) and to the Young inequality we have

$$\frac{R^2}{\delta}\left(\frac{\partial u}{\partial\nu}\right)^2 + (n-1)u\frac{\partial u}{\partial\nu}$$

$$\le \gamma\left(\left(\frac{\partial u}{\partial\nu} + k(0)u\right)^2 - k^2(0)\,|u|^2\right) + (n-1-2\gamma k(0))\,u\frac{\partial u}{\partial\nu}$$

$$\le \gamma\left(a(x)g(u') + \int_0^t k'(t-s)u(s)\,ds\right)^2 - \lambda k(0)\,|u|^2$$

$$+(n-1-2\lambda)u\frac{\partial u}{\partial\nu}$$

$$\le 2\gamma a^2(x)g(u')^2 + 2\gamma\left(\int_0^t k'(t-s)u(s)\,ds\right)^2 - \lambda k(0)\,|u|^2$$

$$+(n-1-2\lambda)u\frac{\partial u}{\partial\nu}.$$

By (2.1)–(2.7) we obtain

$$2\int_{|u'|\le 1}\gamma a^2(x)g(u')^2d\Gamma \le C\int_{|u'|\le 1}(u'g(u'))^{\frac{2}{p+1}}\,d\Gamma$$

$$\le C\left(\int_{\Gamma_1}u'g(u')d\Gamma\right)^{\frac{2}{p+1}} \le C\,(-E')^{\frac{2}{p+1}}$$

and

$$2\int_{|u'|\ge 1}\gamma a^2(x)g(u')^2d\Gamma \le C\int_{|u'|\ge 1}u'g(u')\,d\Gamma \le CE'. \tag{3.7}$$

Hence

$$2\int_S^T E^{\frac{p-1}{2}}\int_{|u'|\le 1}\gamma a^2(x)g(u')^2d\Gamma\,dt \le \varepsilon\int_S^T E^{\frac{p+1}{2}}(t)dt + C(\varepsilon)E(S)$$

and

$$2\int_S^T E^{\frac{p-1}{2}}\int_{|u'|\ge 1}\gamma a^2(x)g(u')^2d\Gamma\,dt \le CE(S).$$

Consequently

$$\int_S^T E^{\frac{p-1}{2}}\int_\Omega\left(|u'|^2+|\nabla u|^2\right)dx\,dt$$

$$\le CE(S)+\varepsilon\int_S^T E^{\frac{p+1}{2}}(t)\,dt + C(\varepsilon)E(S) - \lambda\int_S^T E^{\frac{p-1}{2}}\int_{\Gamma_1}k(0)\,|u|^2\,d\Gamma\,dt$$

$$+2\int_S^T E^{\frac{p-1}{2}}\int_{\Gamma_1}\gamma\left(\int_0^t k'(t-s)u(s)\,ds\right)^2 d\Gamma\,dt$$

$$+(n-1-2\lambda)\int_S^T E^{\frac{p-1}{2}}\int_{\Gamma_1}u\frac{\partial u}{\partial\nu}d\Gamma\,dt.$$

On the other hand, from (2.5) it holds that

$$-\frac{1}{2}\int_S^T E^{\frac{p-1}{2}}\int_{\Gamma_1}\int_0^t k'(t-s)\,(u(t)-u(s))^2\,ds\,d\Gamma\,dt$$

$$\le \frac{1}{2\alpha}\int_S^T E^{\frac{p-1}{2}}\int_{\Gamma_1}\int_0^t k''(t-s)\,(u(t)-u(s))^2\,ds\,d\Gamma\,dt$$

$$\le C\int_S^T E^{\frac{p-1}{2}}E'dt \le CE(S)$$

and

$$\frac{1}{2}\int_S^T E^{\frac{p-1}{2}}\int_{\Gamma_1} k\,|u|^2\,d\Gamma\,dt \le \frac{1}{2}\int_S^T E^{\frac{p-1}{2}}\int_{\Gamma_1} k(0)\,|u|^2\,d\Gamma\,dt.$$

Thus from (2.12) and taking the inequalities above into account we deduce

$$2\int_S^T E^{\frac{p+1}{2}}(t)\,dt \le (1-\lambda)\int_S^T E^{\frac{p-1}{2}}\int_{\Gamma_1} k(0)\,|u|^2\,d\Gamma\,dt$$

$$+CE(S)+C(\varepsilon)E(S)+\varepsilon\int_S^T E^{\frac{p+1}{2}}(t)\,dt$$

$$+2\int_S^T E^{\frac{p-1}{2}}\int_{\Gamma_1}\gamma\left(\int_0^t k'(t-s)u(s)\,ds\right)^2 d\Gamma\,dt$$

$$-(2\lambda+1-n)\int_S^T E^{\frac{p-1}{2}}\int_{\Gamma_1} u\frac{\partial u}{\partial\nu}d\Gamma\,dt.$$

From Lemma 3.4 , the last inequality yields

$$\int_S^T E^{\frac{p+1}{2}}(t)\,dt \le C(\varepsilon)E(S)+\frac{(1-\lambda)}{1-\varepsilon(2\lambda+2-n)}\int_S^T E^{\frac{p-1}{2}}\int_{\Gamma_1} k(0)\,|u|^2\,d\Gamma\,dt$$

$$+\frac{2}{1-\varepsilon(2\lambda+2-n)}\int_S^T E^{\frac{p-1}{2}}\int_{\Gamma_1}\gamma\left(\int_0^t k'(t-s)u(s)\,ds\right)^2 d\Gamma\,dt. \qquad \square$$

LEMMA 3.6. *Let $\varepsilon>0$ be such that*

$$\varepsilon\inf_{\Gamma_1} k'(0)+1>0. \tag{3.8}$$

Then, for any $0\le S<T<\infty$ we have

$$\int_S^T E^{\frac{p-1}{2}}\int_{\Gamma_1}\gamma\left(\int_0^t k'(t-s,x)u(s)\,ds\right)^2 d\Gamma dt$$

$$\le CE(S)+\frac{\lambda}{\varepsilon\alpha}\int_S^T E^{\frac{p-1}{2}}\int_{\Gamma_1}|u|^2\,d\Gamma dt.$$

Proof. Let $\varepsilon>0$ be such that (3.8) holds and we define

$$h:=h(x)=\frac{k(0)}{\alpha\,(1+\varepsilon k'(0))}\quad \text{on } \Gamma_1, \tag{3.9}$$

$$I:=\left(\int_0^t k'(t-s,x)u(s)\,ds\right)^2-h\int_0^t k''(s)\,(u(t)-u(s))^2\,ds+hk'u^2.$$

By (3.8) it holds that $h \geq 0$ and $h \in L^\infty(\Gamma_1)$. On the other hand we have

$$I \leq \left(\int_0^t -k'(t-s,x)\,ds\right)\left(\int_0^t -k'(t-s,x)u^2(s)\,ds\right) - h\int_0^t k''(t-s,x)u^2(s)\,ds$$

$$+2hu\int_0^t k''(t-s,x)u(s)\,ds + hk'(0)u^2 - hk'u^2 + hk'u^2$$

$$\leq (k-k(0))\int_0^t k'(t-s,x)u^2(s)\,ds - h\int_0^t k''(t-s,x)u^2(s)\,ds$$

$$+h\left(k'(0)+\frac{1}{\varepsilon}\right)u^2 + \varepsilon h\left(\int_0^t k''(t-s,x)u(s)\,ds\right)^2.$$

Applying the Hölder inequality to the last integral in the above inequality and using the fact that

$$\int_0^t k'(t-s,x)u^2(s)\,ds \leq 0 \quad \text{and} \quad \varepsilon hk'\int_0^t k''(t-s,x)u^2(s)\,ds \leq 0,$$

we obtain

$$I \leq \left[\frac{1}{\alpha}k(0) - h\left(1+\varepsilon k'(0)\right)\right]\int_0^t k''(t-s,x)u^2(s)\,ds + h\left(k'(0)+\frac{1}{\varepsilon}\right)|u|^2,$$

and hence by (3.9) we have

$$I \leq \frac{1}{\varepsilon\alpha}k(0)\,|u|^2.$$

As $h \in L^\infty(\Gamma_1)$, it holds that

$$\int_S^T E^{\frac{p-1}{2}}\int_{\Gamma_1}\gamma I\,d\Gamma dt \leq \int_S^T E^{\frac{p-1}{2}}\int_{\Gamma_1}\frac{\lambda}{\varepsilon\alpha}k(0)\,|u|^2\,d\Gamma\,dt,$$

$$\int_S^T E^{\frac{p-1}{2}}\int_{\Gamma_1}\left\{\gamma\left(\int_0^t k'(t-s,x)u(s)\,ds\right)^2\right.$$

$$\left.-h\int_0^t k''(s)\left(u(t)-u(s)\right)^2 ds + hk'\,|u|^2\right\}d\Gamma dt$$

$$\leq \int_S^T E^{\frac{p-1}{2}}\int_{\Gamma_1}\frac{1}{\varepsilon\alpha}k(0)\,|u|^2\,d\Gamma\,dt,$$

and by Lemma 3.1 it follows that

$$\int_S^T E^{\frac{p-1}{2}}\int_{\Gamma_1}\gamma\left(\int_0^t k'(t-s,x)u(s)\,ds\right)^2 d\Gamma dt$$

$$\leq CE(S) + \frac{\lambda}{\varepsilon\alpha}\int_S^T E^{\frac{p-1}{2}}\int_{\Gamma_1}|u|^2\,d\Gamma dt. \qquad \square$$

Hence we deduce from Lemma 3.5 that

$$\int_S^T E^{\frac{p+1}{2}}(t)dt \leq CE(S) + \int_S^T E^{\frac{p-1}{2}} \int_{\Gamma_1} \left((1-\lambda)\, k(0) + \frac{2\lambda}{\varepsilon\alpha} \right) |u|^2 \, d\Gamma dt.$$

We would like to make a choice of λ such that

$$(1-\lambda)\, k(0) + \frac{2\lambda}{\varepsilon\alpha} \leq 0,$$

and then by Lemma 2.4 we deduce the desired decay rates.

The condition

$$\alpha \inf_{\Gamma_1} k(0) > -2 \inf_{\Gamma_1} k'(0)$$

implies that

$$\exists \varepsilon' > 0 \ \text{ such that } \quad \alpha \inf_{\Gamma_1} k(0) > -(2+\varepsilon') \inf_{\Gamma_1} k'(0).$$

We choose $\varepsilon > 0$ such that

$$-\varepsilon \inf_{\Gamma_1} k'(0) = \frac{\varepsilon' + 4}{2(\varepsilon' + 2)} \quad (\text{we always have} \quad \varepsilon \inf_{\Gamma_1} k'(0) + 1 > 0).$$

Then, we have

$$(1-\lambda)k(0) + \frac{2\lambda}{\varepsilon\alpha} = \frac{\lambda}{\varepsilon\alpha}\left(2 - \varepsilon\alpha k(0)\right) + k(0)$$

$$\leq \frac{\lambda}{\varepsilon\alpha}\Big(2 + \varepsilon(2+\varepsilon') \inf_{\Gamma_1} k'(0)\Big) + k(0)$$

$$= \frac{-\varepsilon'}{2\varepsilon\alpha}\lambda + k(0).$$

Hence, if we choose

$$\lambda = \max\left\{ n-1, \left(\frac{R^2}{\delta} + \frac{2\varepsilon\alpha}{\varepsilon'} \right) ||k(0)||_{L^\infty(\Gamma_1)} \right\}$$

(note that $\lambda > \max\{\frac{n-1}{2}, \frac{R^2}{\delta}\, ||k(0)||_{L^\infty(\Gamma_1)}\}$ still holds), we obtain

$$(1-\lambda)k(0) + \frac{2\lambda}{\varepsilon} \leq 0. \qquad \square$$

Proof of Theorem 2.3. Repeating the same arguments as in Theorem 2.2, except the part where the first inequality of (2.7) involving C_3 is applied (cf. (3.5)–(3.7)), it remains to establish the estimate

$$E^{\frac{p-1}{2}} \int_{\Gamma_1} |u'|^2 \, d\Gamma \leq \varepsilon E^{\frac{p+1}{2}} - C(\varepsilon)E' \tag{3.10}$$

for every $\varepsilon > 0$. Then the theorem will follow by applying Lemma 2.4.

For brevity we shall denote the norm of $L^\beta(\Gamma)$ by $||\cdot||_\beta$. Set

$$s := \frac{2}{p+1} \quad \text{and} \quad \alpha := \frac{2-s}{1-s}.$$

We have $0 < s < 1$ and $\alpha := \frac{2p}{p-1} > 2$. We establish for every $\varepsilon > 0$ the inequality

$$E^{\frac{p-1}{2}} \int_{\Gamma_1, |u'| \geq 1} |u'|^2 \, d\Gamma \leq \varepsilon E^{\frac{p+1}{2}} ||u'||_\alpha^\alpha - C(\varepsilon) E'. \tag{3.11}$$

Indeed, we deduce

$$E^{\frac{p-1}{2}} \int_{\Gamma_1, |u'| \geq 1} |u'|^2 \, d\Gamma \leq C E^{\frac{p-1}{2}} \int_{\Gamma_1, |u'| \geq 1} |u'|^{2-s} \left(u' g(u')\right)^s d\Gamma$$

$$\leq C E^{\frac{p-1}{2}} \left|\left||u'|^{2-s}\right|\right|_{\frac{1}{1-s}} \left|\left|\left(u' g(u')\right)^s\right|\right|_{\frac{1}{s}}$$

$$\leq C E^{\frac{p-1}{2}} ||u'||_\alpha^{(1-s)\alpha} ||u' g(u')||_1^s$$

$$\leq C E^{\frac{p-1}{2}} ||u'||_\alpha^{(1-s)\alpha} (-E')^s$$

$$\leq \varepsilon E^{\frac{p-1}{2(1-s)}} ||u'||_\alpha^\alpha - C(\varepsilon) E'$$

$$= \varepsilon E^{\frac{p+1}{2}} ||u'||_\alpha^\alpha - C(\varepsilon) E'.$$

Using the trace theorem

$$H^1(\Omega) \hookrightarrow L^{\frac{2p}{p-1}}(\Gamma) = L^\alpha(\Gamma),$$

which follows from (2.16), we have from (3.8) that

$$E^{\frac{p-1}{2}} \int_{\Gamma_1} |u'|^2 \, d\Gamma \leq C \varepsilon E^{\frac{p+1}{2}} - C(\varepsilon) E'.$$

Hence, (3.10) follows with another ε. □

4. Strong asymptotic stability. In this section we are interested in the following nonlinear damped wave equation with a memory term:

$$u'' - \Delta u + g(u') + \int_0^t h(t-\tau) \Delta u(\tau) \, d\tau = 0 \quad \text{in} \quad \Omega \times \mathbf{R}_+, \tag{4.1}$$

$$u = 0 \quad \text{on} \quad \Gamma \times \mathbf{R}_+, \tag{4.2}$$

$$u(x,0) = u_0(x), \;\; u'(x,0) = u_1(x) \quad \text{in} \quad \Omega. \tag{4.3}$$

It is easy to verify that all the results of section 2 still hold for (4.1)–(4.3). However, Theorems 2.2 and 2.3 are marked by the following features:

(a) the domain Ω is bounded;

(b) the dissipative term g is of a preassigned polynomial growth at the origin.

These assumptions are critically invoked in the proofs in the following ways:

(a) the boundedness of Ω allows the use of some compact imbedding theorems (and LaSalle's invariance principle can then be used to prove the strong asymptotic stability);

(b) the polynomial growth at the origin of the dissipative term g allows the construction of a standard Lyapunov function or the use of some specific integral inequalities, which are then used to yield the desired decay rates.

Our goal in this section is to weaken considerably the above assumptions (a)–(b). Indeed, in our formulation Ω is not necessarily bounded, and no growth assumption at the origin is imposed on g. This results in major difficulties, which require the development of a new approach in successfully solving the problem of strong asymptotic stability. This approach was introduced by the first author in [1]. More precisely the following assumptions are made on Ω, g and h:

(A1) Assumptions on Ω:

$$\Omega \text{ is of finite measure.}$$

(A2) Assumptions on g:

$$g : \mathbf{R} \to \mathbf{R} \quad \text{is locally Lipschitz continuous;}$$

$$g(x)x > 0 \quad \text{for all} \quad x \neq 0, \quad g(0) = 0;$$

there exists $q \geq 2$ satisfying $(n-2)q \leq 2n$ and $c_1, c_2 >$ such that

$$c_1|x| \leq |g(x)| \leq c_2|x|^q \quad \text{for all} \quad |x| \geq 1.$$

(A3) Assumptions on the kernel h:

$$h : \mathbf{R}_+ \to \mathbf{R}_+ \text{ is a bounded } C^2 \text{ function;}$$

there exists $l > 0$ such that

$$1 - \int_0^\infty h(x)\,dx = l > 0,$$

and there exists a positive constant η such that

$$h'(t) \leq -\eta h(t) \quad \text{for all} \quad t \geq 0.$$

We define the energy of a solution

$$u \in C(\mathbf{R}_+, H_0^1(\Omega)) \cap C^1(\mathbf{R}_+, L^2(\Omega)) \tag{4.4}$$

by the following formula:

$$E(t) := \frac{1}{2}|u'(t)|^2 + \frac{1}{2}|\nabla u(t)|^2.$$

Our main result in this section is the following theorem.

THEOREM 4.1. *It holds that*

$$E(t) \to 0 \quad as \quad t \to +\infty$$

for every weak solution to (4.1)–(4.3) *given by* (4.4).

In a first step we will assume that we have a strong solution to justify all the computations that follow, and by a classical density argument, the result of Theorem 4.1 still holds for weak solutions. The well-posedness can be derived from the same arguments as in [14].

A simple computation shows that

$$E'(t) = -(g(u'), u') + \int_0^t h(t-\tau)(\nabla u(\tau), \nabla u'(t))\, d\tau.$$

Let

$$(h\square\nabla u)(t) := \int_0^t h(t-\tau)|\nabla u(t) - \nabla u(\tau)|^2\, d\tau; \tag{4.5}$$

hence it is not difficult to see, after integrating by parts, that

$$\int_0^t h(t-\tau)(\nabla u(\tau), \nabla u'(t))\, d\tau = -\frac{1}{2}(h\square\nabla u)'(t) + \frac{1}{2}(h'\square\nabla u)(t)$$

$$+ \frac{1}{2}\frac{d}{dt}\left\{\left(\int_0^t h(s)\, ds\right)|\nabla u(t)|^2\right\} - \frac{1}{2}h(t)|\nabla u(t)|^2. \tag{4.6}$$

From (4.6) and the value of $E'(t)$ we deduce that

$$\frac{d}{dt}\left\{\frac{1}{2}|u'(t)|^2 + \frac{1}{2}|\nabla u(t)|^2 + \frac{1}{2}(h\square\nabla u)(t) - \frac{1}{2}\left(\int_0^t h(s)\, ds\right)|\nabla u(t)|^2\right\}$$

$$= -(g(u'), u') + \frac{1}{2}(h'\square\nabla u)(t) - \frac{1}{2}h(t)|\nabla u(t)|^2. \tag{4.7}$$

Then, defining the modified energy $e(t)$ by

$$e(t) := \frac{1}{2}|u'(t)|^2 + \frac{1}{2}\left(1 - \int_0^t h(s)\, ds\right)|\nabla u(t)|^2 + \frac{1}{2}(h\square\nabla u)(t) \tag{4.8}$$

from (4.7) one has

$$e'(t) = -(g(u'), u') + \frac{1}{2}(h'\square\nabla u)(t) - \frac{1}{2}h(t)|\nabla u(t)|^2. \tag{4.9}$$

Taking (4.8) into account we infer that

$$E(t) \le Me(t) \quad \text{for all} \quad t \ge 0, \tag{4.10}$$

where $M = \max\{l^{-1}, 1\}$.

Our aim is to show that

$$E(t) \to 0 \quad \text{as} \quad t \to +\infty,$$

but in view of (4.10) it is enough to prove that

$$e(t) \to 0 \quad \text{as} \quad t \to +\infty. \tag{4.11}$$

Let us note that by the assumptions assumed on h we have

$$e'(t) \leq 0 \quad \text{for all} \quad t \geq 0. \tag{4.12}$$

In order to prove (4.11), we need the three following lemmas.

LEMMA 4.2. *It holds that*

$$\int_0^t \int_\Omega |ug(u')|\, dxds = o(t), \quad t \to +\infty.$$

LEMMA 4.3. *It holds that*

$$\int_0^t \int_\Omega |u'|^2\, dxds = o(t), \quad t \to +\infty.$$

LEMMA 4.4. *It holds that*

$$\int_0^t J(s)\, ds = o(t), \quad t \to +\infty,$$

where

$$J(t) = -\left(\int_0^t h(s)\, ds\right) |\nabla u(t)|^2 + (h\square\nabla u)(t)$$

$$+ \int_0^t h(t-\tau)(\nabla u(\tau), \nabla u(t))\, d\tau. \tag{4.13}$$

Proof of Lemma 4.2. As g is locally Lipschitz continuous we have

$$\int_{|u'|\leq 1} |ug(u')|\, dx \leq c \int_\Omega (|u'|\, |g(u')|)^{\frac{1}{2}}\, |u|\, dx \leq c \left(\int_\Omega u' g(u')\, dx\right)^{\frac{1}{2}} \|u\|_{L^2(\Omega)}.$$

Similarly, by (A2) we have

$$\int_{|u'|>1} |ug(u')|\, dx \leq c \left(\int_\Omega u' g(u')\, dx\right)^{\frac{q}{q+1}} \|u\|_{L^{q+1}(\Omega)}.$$

Then from Hölder's inequality we obtain

$$\int_0^t \int_\Omega |ug(u')|\, dxds \leq c \left(\int_0^t \int_\Omega u' g(u')\, dxds\right)^{\frac{1}{2}} \sqrt{t} \sup_{[0,t]} \|u(s)\|_{L^2(\Omega)}$$

$$+ ct^{\frac{1}{q+1}} \left(\int_0^t \int_\Omega u' g(u')\, dxds\right)^{\frac{q}{q+1}} \sup_{[0,t]} \|u(s)\|_{L^{q+1}(\Omega)}.$$

Using the Hölder, Sobolev, and Poincaré inequalities we have

$$\|u(s)\|_{L^2(\Omega)} \leq c\, \|u(s)\|_{L^{q+1}(\Omega)} \leq c\, E(s)^{\frac{1}{2}} \leq c\, e(0)^{\frac{1}{2}} \quad \text{for all} \quad s \geq 0.$$

From these estimates it follows that

$$\int_0^t \int_\Omega |ug(u')|\,dxds \le c\sqrt{t} + ct^{\frac{1}{q+1}} = o(t), \quad t \to +\infty.$$

Proof of Lemma 4.3. Let $\varepsilon > 0$ be an arbitrarily small real and set

$$M(\varepsilon) = \sup\left\{ \frac{x}{g(x)}; \quad |x| \ge \sqrt{\frac{\varepsilon}{|\Omega|}} \right\};$$

by hypotheses (A2), we have $M(\varepsilon) < +\infty$. Clearly,

$$\int_{|u'|<\sqrt{\frac{\varepsilon}{|\Omega|}}} |u'|^2\,dx \le \varepsilon.$$

On the other hand

$$\int_{|u'|\ge\sqrt{\frac{\varepsilon}{|\Omega|}}} |u'|^2\,dx = \int_{|u'|\ge\sqrt{\frac{\varepsilon}{|\Omega|}}} \frac{|u'|}{g(u')}\,u'g(u')\,dx \le M(\varepsilon)\int_\Omega u'g(u')\,dx.$$

As

$$\int_{|u'|\ge\sqrt{\frac{\varepsilon}{|\Omega|}}} |u'|^2\,dx \le c\sqrt{2e(0)}\,\left(\int_{|u'|\ge\sqrt{\frac{\varepsilon}{|\Omega|}}} |u'|^2\,dx\right)^{\frac{1}{2}},$$

we deduce that

$$\int_\Omega |u'|^2\,dx \le \varepsilon + c\sqrt{2e(0)M(\varepsilon)}\,\left(\int_\Omega u'g(u')\,dx\right)^{\frac{1}{2}},$$

and then by the Hölder inequality

$$\int_0^t \int_\Omega |u'|^2\,dxds \le \varepsilon t + c\sqrt{2e(0)M(\varepsilon)}\,\sqrt{t}\,\left(\int_0^t \int_\Omega u'g(u')\,dxds\right)^{\frac{1}{2}}$$

$$\le \varepsilon t + ce(0)\sqrt{2M(\varepsilon)}\,\sqrt{t} = o(t), \quad t \to +\infty,$$

by Lemma 4.2.

Proof of Lemma 4.4. Define

$$I_1 := \int_0^t h(t-\tau)(\nabla u(\tau), \nabla u(t))\,d\tau.$$

We have

$$|I_1| = \left| \int_0^t h(t-\tau)(\nabla u(\tau) - \nabla u(t), \nabla u(t))\, d\tau + \int_0^t h(t-\tau)|\nabla u(t)|^2\, d\tau \right|$$

$$\leq |\nabla u(t)| \int_0^t h(t-\tau)|\nabla u(\tau) - \nabla u(t)|\, d\tau + \left(\int_0^t h(s)\, ds \right) |\nabla u(t)|^2$$

$$\leq \left(\operatorname{esssup}_{t\geq 0} |\nabla u(t)| \right) \left(\int_0^t h(s)\, ds \right)^{1/2} \left(\int_0^t h(t-\tau)|\nabla u(t) - \nabla u(\tau)|^2\, d\tau \right)^{1/2}$$

$$+ \left(\int_0^t h(s)\, ds \right) |\nabla u(t)|^2;$$

that is,

$$|I_1| \leq \left(\operatorname{esssup}_{t\geq 0} |\nabla u(t)| \right) \|h\|_{L^1(0,\infty)}^{1/2} \left[(h \square \nabla u)(t) \right]^{1/2} + \left(\int_0^t h(s)\, ds \right) |\nabla u(t)|^2. \tag{4.14}$$

Define

$$I_2 := (h \square \nabla u)(t);$$

then we have

$$|I_2| = \int_0^t h(t-\tau)|\nabla u(t) - \nabla u(\tau)|^2\, d\tau$$

$$\leq \int_0^t h(t-\tau) \left(|\nabla u(t)| + |\nabla u(\tau)| \right) |\nabla u(t) - \nabla u(\tau)|\, d\tau$$

$$\leq \left(2 \operatorname{esssup}_{\zeta\geq 0} |\nabla u(\zeta)| \right) \int_0^t h(t-\tau)|\nabla u(t) - \nabla u(\tau)|\, d\tau$$

$$\leq \left(2 \operatorname{esssup}_{\zeta\geq 0} |\nabla u(\zeta)| \right) \left(\int_0^t h(s)\, ds \right)^{1/2} \left(\int_0^t h(t-\tau)|\nabla u(t) - \nabla u(\tau)|^2\, d\tau \right)^{1/2};$$

that is,

$$|I_2| \leq \left(2 \operatorname{esssup}_{\zeta\geq 0} |\nabla u(\zeta)| \right) \|h\|_{L^1(0,\infty)}^{1/2} (h \square \nabla u)^{1/2}(t). \tag{4.15}$$

Combining (4.13), (4.14), and (4.15) we deduce that

$$J(t) \leq \left(3 \operatorname{esssup}_{\zeta\geq 0} |\nabla u(\zeta)| \right) \|h\|_{L^1(0,\infty)}^{1/2} \left[(h \square \nabla u)(t) \right]^{1/2}. \tag{4.16}$$

Thanks to the above inequality and to the hypotheses on h one has

$$J(t) \le \big(3\mathrm{esssup}_{\zeta\ge 0}|\nabla u(\zeta)|\big)\,\|h\|^{1/2}_{L^1(0,\infty)}\left(-\frac{1}{\eta}(h'\Box\nabla u)(t)\right)^{1/2}$$

$$\le 3\sqrt{2}\big(E(0)\big)^{1/2}\|h\|^{1/2}_{L^1(0,\infty)}\left(-\frac{1}{\eta}(h'\Box\nabla u)(t)\right)^{1/2} \tag{4.17}$$

$$= c\left(-\frac{1}{\eta}(h'\Box\nabla u)(t)\right)^{1/2}.$$

Integrating (4.17) over $(0,t)$, we obtain from (4.9) that

$$\int_0^t J(s)\,ds \le c\int_0^t 1\cdot\left(-\frac{1}{\eta}(h'\Box\nabla u)(s)\right)^{1/2} ds$$

$$\le c\left(\int_0^t 1\,ds\right)^{1/2}\left(\int_0^t -\frac{1}{\eta}(h'\Box\nabla u)(s)\,ds\right)^{1/2}$$

$$= c\sqrt{t}\int_0^t\left(-\frac{1}{\eta}(h'\Box\nabla u)(s)\right)ds \le \frac{2c}{\eta}e(0)\sqrt{t} = o(t),\quad t\to+\infty.$$

Proof of Theorem 4.1. Put

$$\phi(t) := (u(t), u'(t));$$

then we have

$$\phi'(t) = -|\nabla u(t)|^2 - (g(u'),u) + \int_0^t h(t-\tau)(\nabla u(\tau),\nabla u(t))\,d\tau + |u'(t)|^2. \tag{4.18}$$

Adding and subtracting appropriate terms in (4.18) it follows that

$$\phi'(t) = -2e(t) + 2|u'(t)|^2 - (g(u'),u)$$

$$-\left(\int_0^t h(s)\,ds\right)|\nabla u(t)|^2 + (h\Box\nabla u)(t) + \int_0^t h(t-\tau)(\nabla u(\tau),\nabla u(t))\,d\tau. \tag{4.19}$$

That is,

$$\phi'(t) = -2e(t) + 2|u'(t)|^2 - (g(u'),u) + J(t).$$

Then, it follows that

$$\phi(t) - \phi(0) = \int_0^t \big\{2|u'(t)|^2 - 2e(t) - (g(u'),u) + J(s)\big\}\,ds.$$

Assume that, contrary to our claim, $l := \lim_{t\to\infty} e(t) > 0$. Then by Lemmas 4.2–4.4, we have

$$\phi(t) - \phi(0) \le -2lt + o(t),\quad t\to+\infty.$$

It follows that $\phi(t) \to -\infty$ as $t \to +\infty$. But this is impossible because

$$\left| \int_\Omega uu' \, dx \right| \leq c \int_\Omega (|\nabla u|^2 + u'^2) \, dx \leq ce(0).$$

We conclude that $\lim_{t\to+\infty} E(t) = 0$.

Remark 4.1. If g was linear or superlinear near zero, then it is sufficient to assume that Ω is an open set in which the Poincaré inequality holds. If we, furthermore, add the term u to (4.1), then we can assume that $\Omega = \mathbf{R}^n$.

REFERENCES

[1] M. Aassila, *A new approach to the strong stabilization of distributed systems*, Differential Integral Equations, 11 (1998), pp. 369–376.

[2] J. T. Beale, *Spectral properties of an acoustic boundary condition*, Indiana Univ. Math. J., 25 (1976), pp. 895–917.

[3] G. Chen, *A note on the boundary stabilization of the wave equation*, SIAM J. Control Optim., 19 (1981), pp. 106–113.

[4] F. Conrad and B. Rao, *Decay of solutions of the wave equation in a star-shaped domain with nonlinear boundary feedback*, Asymptot. Anal., 7 (1993), pp. 159–177.

[5] J. G. Dix and R. M. Torrejon, *A quasilinear integrodifferential equation of hyperbolic type*, Differential Integral Equations, 6 (1993), pp. 431–447.

[6] A. Guesmia, *Stabilisation de l'équation des ondes avec conditions aux limites de type mémoire*, Afrika Math. (3), 10 (1999), pp. 14–25.

[7] V. Komornik, *Exact Controllability and Stabilization. The Multiplier Method*, John Wiley, Paris, 1994.

[8] V. Komornik and E. Zuazua, *A direct method for the boundary stabilization of the wave equation*, J. Math. Pures Appl., 69 (1990), pp. 33–54.

[9] R. Leis, *Initial Boundary Value Problems in Mathematical Physics*, John Wiley, New York, 1986.

[10] A. Majda, *The location of the spectrum for the dissipative acoustic operator*, Indiana Univ. Math. J., 25 (1976), pp. 973–987.

[11] P. M. Morse and K. U. Ingrad, *Theorical Acoustics*, McGraw-Hill, New York, 1968.

[12] J. E. Munoz Rivera, *Global solution of a quasilinear wave equation with memory*, Boll. Un. Mat. Ital. B, 8 (1994), pp. 289–303.

[13] G. Propst and J. Prüss, *On wave equations with boundary dissipation of memory type*, J. Integral Equations Appl., 8 (1996), pp. 99–123.

[14] J. Prüss, *Evolutionary Integral Equations and Applications*, Birkhäuser-Verlag, Basel, 1993.

[15] J. P. Quinn and D. L. Russell, *Asymptotic stability and energy decay rates for solutions of hyperbolic equations with boundary damping*, Proc. Roy. Soc. Edinburgh Sect. A, 77 (1977), pp. 97–127.

[16] M. Renardy, W. Hrusa, and J. Nohel, *Mathematical Problems in Viscoelasticity*, John Wiley, New York, 1987.

[17] H. Tadayuki and M. Rinko, *Equivalent condition for stability of a Volterra integro-differential equation*, J. Math. Anal. Appl., 174 (1993), pp. 298–316.

SIAM J. CONTROL OPTIM.
Vol. 38, No. 5, pp. 1603–1622

OPTIMAL POLICIES FOR n-DIMENSIONAL SINGULAR STOCHASTIC CONTROL PROBLEMS PART I: THE SKOROKHOD PROBLEM*

LUKASZ KRUK†

Abstract. An n-dimensional Brownian motion is controlled by adding a process of locally bounded variation to it so as to minimize an expected infinite-horizon discounted cost. We show, by direct probabilistic techniques, that the optimal control is a solution of a (generalized) Skorokhod problem.

Key words. singular stochastic control, optimal policy, the Skorokhod problem

AMS subject classification. 93E20

PII. S0363012998347535

1. Introduction.

Singular stochastic control of Markov processes is a class of problems in which we are allowed to alter the drift of a process (usually the Brownian motion) at a price proportional to the variation of the control used. Admissible controls may not be absolutely continuous and may have jumps. The corresponding Bellman equation for such a problem is a pair of differential inequalities, one of second, the other of first order. The latter one determines the nonaction region $\mathcal{C}$. It was shown that in many cases the optimal control makes the underlying Markov process a diffusion reflected at the boundary of $\mathcal{C}$ with an initial jump to its boundary if it starts outside $\mathcal{C}$. The first papers in which a problem of this type was considered are, to our knowledge, [1, 2].

One of the problems that attracted the attention of researchers in this field was to minimize

$$u_M(x) = E^x \int_0^\infty e^{-t}[h(X_t)dt + d\xi_t], \tag{1}$$

where $x \in R^n$ is fixed, h is a nonnegative, convex function (usually smooth),

$$X_t = x + \sqrt{2}W_t + M_t \tag{2}$$

with W_t an n-dimensional Brownian motion, M_t any adapted process of bounded variation over finite time-intervals, and ξ_t the variation of M up to time t. Variations of this problem in one dimension were considered in [3, 4, 11, 12, 13, 16, 17, 19, 20, 24] and other papers (see, e.g., [25] or [9] for more complete references). It was shown that the corresponding value function u is twice continuously differentiable (smooth fit), and the optimal policy is the Brownian motion reflected at the endpoints of the interval in which the absolute value of the derivative of u is less than 1. Simplified treatment of the one-dimensional case was later given in [6, 7].

*Received by the editors November 12, 1998; accepted for publication (in revised form) November 24, 1999; published electronically May 26, 2000. This work is a part of the author's Ph.D. thesis at the Courant Institute of Mathematical Sciences, New York University.
http://www.siam.org/journals/sicon/38-5/34753.html
†Courant Institute of Mathematical Sciences, New York University, 251 Mercer Street, New York, NY 10012 (kruk@cims.nyu.edu). Current address: Department of Mathematical Sciences, Carnegie Mellon University, Pittsburgh, PA 15213 (kruk@asdf4.math.cmu.edu).

Generalization of these results to higher dimensions, however, is not straightforward. Some regularity of both the domain and the vector field determining the reflection direction is necessary to define the Brownian motion reflected at the boundary of such a domain (see, e.g., [14, 23, 30, 33]). In the two-dimensional case, $C^{2,\alpha}$ regularity was obtained for both the value function and the boundary of the nonaction region by PDE methods: elliptic regularity and regularity for free-boundary problems [29]. In dimensions greater than two only partial results are known: the value function is $W^{2,\infty}_{loc}$; the optimal policy exists and is unique [26, 31], keeps the controlled process in some compact set $\tilde{D}$ containing $\overline{\mathcal{C}}$, and does not act in $\mathcal{C}$ [25]. Regularity of the free boundary away from some corner points for a similar n-dimensional problem was shown in [34].

Let us also mention that this class of stochastic control problems is closely related to the target zone models of the foreign exchange rates (see, e.g., [21]). In fact, our initial motivation for considering it was to use it in multilateral target zone modeling.

In this article we prove, by a direct probabilistic argument, that in any dimension the optimal policy minimizing (1) is a solution to the Skorokhod problem (in a slightly generalized sense, to be defined later). This approach, however, leaves the question of smooth fit in higher dimensions unanswered. The second part examines the case of a radially symmetric running cost h and considers a similar problem with ergodic (instead of exponential) discounting in which we limit the admissible controlled processes to Brownian motions reflected normally at the boundary of some domain.

2. The optimal policy as a solution to the Skorokhod problem.

2.1. Definitions and assumptions. Let $(W_t, \mathcal{F}_t, t \geq 0)$ be a standard n-dimensional Brownian motion defined on a complete probability space $(\Omega, \mathcal{F}, P)$. Let $\{\mathcal{F}_t\}$ be the augmentation of the filtration generated by W (see [18, p. 89]). Let, for a given $x \in R^n$, a process X_t be defined by (2), where M_t is a left-continuous process adapted to $\mathcal{F}_t$ such that (s.t.), for all $T > 0$, P almost surely (a.s.), the variation of $M_\cdot(\omega)$ on the interval $[0, T]$ is finite. As is customary in singular stochastic control theory, we write

$$M_t = \int_0^t N_s d\xi_s, \tag{3}$$

where $|N_t| = 1$ for every $t \geq 0$ a.s. and ξ is nondecreasing and left-continuous. In other words, $\xi_t(\omega)$ is the total variation of $M_\cdot(\omega)$ on the interval $[0, t]$, $N_t(\omega)$ is the Radon–Nikodym derivative of the measure induced on $[0, \infty)$ by $M_\cdot(\omega)$ with respect to its total variation $\xi_\cdot(\omega)$. In what follows, we shall always describe M_t as (N_t, ξ_t).

Let $h : R^n \to R$ be a strictly convex function satisfying, for appropriate positive constants C_0, c_0, and $q > 1$, the following conditions:

$$h \in C^{2,1}(R^n), \tag{4}$$

$$0 \leq h(x) \leq C_0(1 + |x|^q), \tag{5}$$

$$|h(x) - h(x + x')| \leq C_0(1 + h(x) + h(x + x'))^{1-1/q}|x'|, \tag{6}$$

$$h(x + \lambda x') + h(x - \lambda x') - 2h(x) \leq C_0\lambda^2(1 + h(x))^r, \qquad r = \left(1 - \frac{2}{q}\right)^+, \tag{7}$$

$$c_0|y|^2 \leq D^2h(x)y \cdot y \tag{8}$$

for all $x, x' \in R^n$, $|x'| \le 1$, $0 \le \lambda \le 1$. These conditions are similar to those in [26]. (As the authors of that paper expressed it, (5), (6), (7) say that "h is roughly speaking of polynomial growth.") In particular, (6) yields

$$|\nabla h(x)| \le \tilde{C}_0(1 + h(x)) \tag{9}$$

and (7) yields

$$D^2h(x)y \cdot y \le C_0(1 + h(x))|y|^2, \tag{10}$$

so the assumptions of [29] are also satisfied. We need them because we want to use the results from both these papers. It may be a good idea to see how (5)–(8) work for a quadratic function (Ax, x) with a symmetric positive definite matrix A in order to get used to them.

For convenience, we assume that $\inf_{x \in R^n} h(x) = h(0) = 0$.

For a given control process (N, ξ), we define the corresponding cost by

$$V_{(N,\xi)}(x) = E^x \int_0^\infty e^{-t}[h(X_t)dt + d\xi_t]. \tag{11}$$

The task is to minimize $V_{(N,\xi)}(x)$ in the class of all admissible controls, i.e., to find

$$u(x) = \inf_{N,\xi} V_{(N,\xi)}(x), \tag{12}$$

where (N_t, ξ_t) are as described above. If this minimum is attained for some $(\tilde{N}, \tilde{\xi})$, we say that

$$\nu_t = \int_0^t \tilde{N}_s d\tilde{\xi}_s \tag{13}$$

is an optimal policy (for x).

DEFINITION 2.1. *If, for a sequence (N_n, ξ_n) of controls,*

$$u(x) = \lim_{n \to \infty} V_{(N_n,\xi_n)}(x), \tag{14}$$

we say that (N_n, ξ_n) is a sequence of ϵ-optimal policies (for x).

To avoid unnecessary complications in the exposition of the proofs, we will use this definition rather than (as is done usually) check that for a given control (N, ξ) $u(x) \ge V_{(N,\xi)}(x) - \epsilon$.

LEMMA 2.2. *The optimal policy (if it exists) is unique.*

LEMMA 2.3. *For every $x \in R^n$ there exists a unique optimal policy ν^*. Moreover, if (N_n, ξ_n) is a sequence of ϵ-optimal policies for x, then we can extract from it a subsequence $n_k \to \infty$ s.t.*

$$\nu_t^{n_k} = \int_0^t N_s^{n_k} d\xi_s^{n_k} \to \nu_t^* \tag{15}$$

for $L_{eb} \times P$ almost all (t, ω), where L_{eb} is the Lebesgue measure on $[0, \infty)$.

These lemmas are essentially proved in [26]. It can be shown, by a direct control-theoretic approach given in [26], that the value function $u \in W^{2,\infty}_{loc}$ and, moreover, it satisfies the Bellman equation

$$\max(u - \triangle u - h, |\nabla u|^2 - 1) = 0 \tag{16}$$

in R^n (see the appendix at the end of this paper (section 3) for further discussion).

It can also be proven by PDE methods that (16) admits a unique nonnegative, convex $W^{2,\infty}_{loc}$ solution (see [8, 15, 29]). Nonnegativity and convexity of the value function are easy consequences of its definition (see [9, Lemma VIII.3.2]), so the value function u is a unique nonnegative, convex $W^{2,\infty}_{loc}$ solution of the Bellman equation (16).

In what follows, let us denote the total variation of a function $f(s)$ on an interval $[0,t]$ by $\bigvee_{[0,t]} f$.

As we have mentioned before, for $n = 1$ or 2 it is known that $u \in C^{2,\alpha}$ for every $\alpha \in (0,1)$. In both cases, the optimal policy makes X_t a Brownian motion reflected at $\partial\mathcal{C}$ at the direction of $-\nabla u$, where

$$\mathcal{C} = \{x \in R^n : |\nabla u(x)| < 1\}. \tag{17}$$

More precisely, the optimally controlled process X_t solves the following problem (with $G := \mathcal{C}$, $v := -\nabla u$).

DEFINITION 2.4. *Let G be an open domain in R^n, $S = \partial G$. Let $x_0 \in G$ and let v be an unit vector field defined on S, i.e., for each $x \in S$, $|v(x)| = 1$, pointing inside G (in particular, nontangential to S). We say that a continuous process*

$$\nu_t = \int_0^t N_s d\xi_s, \tag{18}$$

where $\xi_t = \bigvee_{[0,t]} \nu$, is a solution to a Skorokhod problem for reflected Brownian motion in $\overline{G}$ starting at x_0 with reflection direction v along S if

(a) *$|N_t| \equiv 1$, ξ_t is continuous and nondecreasing;*

(b) *the process X_t defined by*

$$X_t = x_0 + \sqrt{2}W_t + \int_0^t N_s d\xi_s \tag{19}$$

satisfies

$$X_t \in \overline{G}, \quad 0 \le t < \infty, \ \text{a.s.}; \tag{20}$$

(c) *for every $0 \le t < \infty$*

$$\xi_t = \int_0^t I_{[X_s \in \partial G, N_s = v(X_s)]} d\xi_s. \tag{21}$$

This is what is customarily called a solution to the Skorokhod problem for W_t, G, v (see, e.g., [23]).

Now we shall give a slightly modified definition of a solution to the Skorokhod problem.

DEFINITION 2.5. *Let G be an open domain in R^n, $S = \partial G$. Let $x_0 \in G$ and let v be a continuous unit vector field defined on S, i.e., for each $x \in S$ $|v(x)| = 1$. We say that a left-continuous process*

$$\nu_t = \int_0^t N_s d\xi_s, \tag{22}$$

where $|N_t| = 1$, $\xi_t = \bigvee_{[0,t]} \nu$, is a solution to the modified Skorokhod problem for reflected Brownian motion in $\overline{G}$ starting at x_0 with reflection direction v along S if

(a) *with probability* 1*, every possible jump of the process*

$$X_t = x_0 + \sqrt{2}W_t + \nu_t \tag{23}$$

occurs on some interval I *contained in* S *and parallel to the vector field* v *on* I*, i.e., such that for all* $\tilde{x} \in I$*,* $v(\tilde{x})$ *is parallel to* I*. If* X_t *encounters such an interval* I*, it instantaneously jumps to its endpoint in the direction of* v *on* I*.*

Stating this condition more formally, (a.s. $\omega \in \Omega$ *for all* $t \geq 0$*), the following statement is true:* $\xi_{t^+}(\omega) > \xi_t(\omega)$ *(i.e., the trajectory* $X_\cdot(\omega)$ *has a jump at* t*) iff there exists a closed interval*

$$I = \{z_0 + tv(x) : t \in [0, c]\} \subseteq S, \tag{24}$$

where $z_0 \in R^n$*,* $c > 0$*,* $x = X_t(\omega)$*, s.t.* $x \in I$ *and* $v(\tilde{x})$ *is parallel to* I *for every* $\tilde{x} \in I$*. Also, if we assume that* I *is a maximal interval with such properties (i.e., there is no* $\tilde{I}$ *enjoying the same properties and s.t.* I *is properly contained in* $\tilde{I}$*), then* $X_{t+}(\omega) = z_0 + cv(x)$*, and* (b), (c) *are as in the previous definition.*

In particular, if intervals of the type described in (a) do not exist (e.g., v is never tangential to S or G is strictly convex), then ξ_t is continuous in $[0, \infty)$ a.s. ω, and this definition reduces to the previous one.

The goal of this paper is to prove the following theorem in any dimension n.

THEOREM 2.6. *For every* $x \in \overline{\mathcal{C}}$ *the optimal policy* $\nu_t = \int_0^t N_s d\xi_s$ *is a solution to the modified Skorokhod problem for* W_t*,* $\mathcal{C}$*,* $-\nabla u$*.*

Intuitively, such a theorem should be true because for the optimally controlled process X_t all the inequalities in the corresponding verification theorem (see, e.g., Theorem VIII.4.1 in [9]) must, in fact, be equalities. We actually use this idea in one of the stages of our proof.

The rest of this paper is entirely devoted to the proof of Theorem 2.6. The proof is long and proceeds in several steps. In subsection 2.2 we construct a sequence of ϵ-optimal policies which are used in subsection 2.3 to prove the condition (b) from Definition 2.5. Subsection 2.4 contains the proof of (c). To prove (a), we need one of the consequences of the Bellman principle and the strong Markov property of the optimally controlled process. We show them in subsection 2.5; the following subsection provides the proof of (a). The concluding subsection 2.7 contains further discussion of some details of the proof.

2.2. Some ϵ-optimal policies. Let, for every $\epsilon > 0$,

$$\mathcal{C}_\epsilon = \{|\nabla u|^2 < 1 - \epsilon\}, \qquad S_\epsilon = \partial \mathcal{C}_\epsilon. \tag{25}$$

In the set $\mathcal{C}$ (defined by (17)), $u \in C^{4,\alpha}$ for every $\alpha \in (0, 1)$.

Indeed, let B be any open ball such that $\overline{B} \subseteq \mathcal{C}$. By Theorem 6.13 of [10], the Dirichlet problem

$$\tilde{u} - \triangle \tilde{u} = h$$

in B,

$$\tilde{u} = u \tag{26}$$

on ∂B, has a solution $\tilde{u} \in C^0(\overline{B}) \cap C^{2,\alpha}(B)$. In particular, $\tilde{u} - u \in W^{2,\infty}(B)$, so, by (26), $\tilde{u} - u \in W_0^{1,2}(B)$ (see [10, p. 154]). By Theorem 8.9 of [10], $u \equiv \tilde{u}$ in B, so

$u \in C^{2,\alpha}(B)$. By Theorem 6.17 of [10], $u \in C^{4,\alpha}(B)$ (and thus $u \in C^{4,\alpha}(\mathcal{C})$) for all $\alpha \in (0,1)$.

Also, u is strictly convex in $\mathcal{C}$ (see (4.10) of [29]). Thus in $\overline{\mathcal{C}}_{\frac{\epsilon}{2}}$, by compactness,

$$\inf_{x \in \overline{\mathcal{C}}_{\frac{\epsilon}{2}}} \inf_{\nu \in R^n, |\nu|=1} (D^2 u)\nu \cdot \nu \geq c_\epsilon > 0. \tag{27}$$

In particular, if $w = |\nabla u|^2$,

$$\nabla w(x) \neq 0 \ \text{ for all } \ x \in S_\epsilon, \tag{28}$$

because for $\nu = \frac{\nabla u}{|\nabla u|}$, $x \in S_\epsilon$,

$$w(x) = 1 - \epsilon, \qquad w(x) = \left| \frac{\partial u}{\partial \nu} \right|^2, \tag{29}$$

$$\frac{\partial^2 u}{\partial \nu^2} \geq c_\epsilon \ \text{ near } \ S_\epsilon, \tag{30}$$

so

$$\sqrt{w(x+\lambda\nu)} \geq \frac{\partial u}{\partial \nu}(x+\lambda\nu) \tag{31}$$

$$\geq \frac{\partial u}{\partial \nu}(x) + \frac{c_\epsilon}{2}\lambda \tag{32}$$

$$= \sqrt{w(x)} + \frac{c_\epsilon}{2}\lambda \tag{33}$$

for λ small enough, so

$$\frac{\partial \sqrt{w(x)}}{\partial \nu} \geq \frac{c_\epsilon}{2} \tag{34}$$

on S_ϵ, which clearly implies $\frac{\partial w(x)}{\partial \nu} \neq 0$ on S_ϵ. Thus, by the implicit function theorem, S_ϵ is a $C^{3,\alpha}$ hypersurface for every $\alpha \in (0,1)$. Of course, the vector field

$$v(x) = -\frac{\nabla u(x)}{|\nabla u(x)|} \tag{35}$$

is also $C^{3,\alpha}$ in the neighborhood of S_ϵ.

LEMMA 2.7. *$v(x)$ is not tangential to S_ϵ.*

Proof. Suppose that, at $x \in S_\epsilon$, $\nabla u(x)$ is tangential to S_ϵ. Then, by (33) and $v(x) = -\nu(x)$,

$$\sqrt{w(x - \lambda v(x))} \geq \sqrt{w(x)} + \frac{c_\epsilon}{2}\lambda = \sqrt{1-\epsilon} + \frac{c_\epsilon}{2}\lambda \tag{36}$$

for small $\lambda > 0$. But, on the other hand, $v(x)$ is tangential to $S_\epsilon = \{w(x) = 1 - \epsilon\}$, so

$$w(x - \lambda v(x)) = w(x) + O(\lambda^2) = 1 - \epsilon + O(\lambda^2), \tag{37}$$

a contradiction.

One can also see that ∇u points outside (not inside) $\mathcal{C}_\epsilon$ (actually, it is proven in [29] that $|\nabla u|$ increases in the direction of ∇u), so it is possible to define an

instantaneous reflection at S_ϵ in the direction $v(x)$, i.e., the solution to the Skorokhod problem for W_t, $\mathcal{C}_\epsilon$, $v(x)$ (see [23]). We shall denote it by

$$\nu_t^\epsilon = \int_0^t N_s^\epsilon d\xi_s^\epsilon. \tag{38}$$

LEMMA 2.8. *ν_t^ϵ are ϵ-optimal policies for our control problem.*

Proof. The value function u satisfies

$$\triangle u - u + h = 0 \text{ in } \mathcal{C}_\epsilon \tag{39}$$

and

$$\frac{\partial u}{\partial v} = -\sqrt{w(x)} = -\sqrt{1-\epsilon} \text{ on } \partial\mathcal{C}_\epsilon. \tag{40}$$

(Recall that v is the direction of $-\nabla u$, so $\sqrt{w(x)} = |\frac{\partial u}{\partial v}| = -\frac{\partial u}{\partial v}$.)

Thus, by Itô's rule for continuous semimartingales (see, e.g., [18, Theorem 3.3.6]) applied to $e^{-t}u(X_t^\epsilon)$, where

$$X_t^\epsilon = x + \sqrt{2}W_t + \nu_t^\epsilon, \tag{41}$$

taking into account the fact that $\mathcal{C}$ is bounded [29, Lemma 4.4], we get

$$\begin{aligned} u(x) &= E^x \int_0^\infty e^{-t}[h(X_t^\epsilon)dt + \sqrt{w(X_t^\epsilon)}d\xi_t^\epsilon] \\ &= E^x \int_0^\infty e^{-t}[h(X_t^\epsilon)dt + \sqrt{1-\epsilon}d\xi_t^\epsilon] \\ &\geq \sqrt{1-\epsilon}\, E^x \int_0^\infty e^{-t}[h(X_t^\epsilon)dt + d\xi_t^\epsilon] \\ &= \sqrt{1-\epsilon}\, J_x(\nu^\epsilon), \end{aligned} \tag{42}$$

where $J_x(\tilde{\nu})$ denotes the cost associated with choosing the policy

$$\tilde{\nu_t} = \int_0^t \tilde{N}_s d\tilde{\xi}_s, \tag{43}$$

i.e.,

$$J_x(\tilde{\nu}) = E^x \int_0^\infty e^{-t}[h(\tilde{X}_t)dt + d\tilde{\xi}_t], \tag{44}$$

and

$$\tilde{X}_t = x + \sqrt{2}W_t + \tilde{\nu}_t. \tag{45}$$

(The inequality in (42) is true because $h \geq 0$.)

This proves ϵ-optimality of ν^ϵ.

2.3. The optimal policy keeps the process in $\overline{\mathcal{C}}$. By Lemmas 2.3 and 2.8 we see that, for any fixed $x \in \mathcal{C}$, there exists a sequence $\epsilon_k \downarrow 0$ s.t.

$$\nu^{\epsilon_k} \to \nu \text{ almost everywhere (a.e.) } L_{eb} \times P \text{ on } [0,\infty) \times \Omega, \tag{46}$$

where ν^{ϵ_k} are defined by (38) and

$$\nu_t = \int_0^t N_s d\xi_s \tag{47}$$

is the unique optimal policy. Let

$$X_t = x + \sqrt{2}W_t + \nu_t \tag{48}$$

be the optimally controlled process.

LEMMA 2.9. *$X_t \in \overline{\mathcal{C}}$ for all $0 \le t < \infty$ a.s. if $x \in \mathcal{C}$.*

Proof. Let

$$A = \{\omega : (X_t^{\epsilon_k})(\omega) \in \overline{\mathcal{C}}_{\epsilon_k} \text{ for all } 0 \le t < \infty \text{ and all } k \ge 0\}. \tag{49}$$

From the definition of ν_t^ϵ (see the paragraph preceding (38)) and (41) we know that $P(A) = 1$. $\overline{\mathcal{C}}_{\epsilon_k} \subseteq \mathcal{C}$ for all $\epsilon > 0$, so for $\omega \in A$ and every $t \ge 0$, $X_t^{\epsilon_k}(\omega) \in \overline{\mathcal{C}}$. Also let

$$B = \{\omega : X^{\epsilon_k}(\omega) \to X(\omega) \text{ a.e. } L_{eb} \text{ on } [0,\infty)\}. \tag{50}$$

By (46) $P(B) = 1$. For all $\omega \in A \cap B$, $X_t(\omega) \in \overline{\mathcal{C}}$ L_{eb} a.e. on $[0,\infty)$, because $\overline{\mathcal{C}}$ is closed. But ν is left-continuous, so $\nu_t(\omega) \in \overline{\mathcal{C}}$ for all $0 \le t < \infty$ and $\omega \in A \cap B$ (for $t = 0$ it is true because $X(0) = x \in \mathcal{C}$). The proof is complete.

Remark. The statement of the last lemma is also true for $x \in \partial\mathcal{C}$. Indeed, in this case let $x_n \in \mathcal{C}$ be s.t. $x_n \to x$ and suppose we want to control the process starting at x. Policies

$$\tilde{\nu}^{x^n} = \nu^{x^n} + x_n - x \tag{51}$$

jump at time 0 from the starting point x to x_n and then follow ν^{x_n}, the optimal policy for the process starting at x_n. Because $x_n \to x$ and u is continuous, it is easy to see that $\tilde{\nu}^{x^n}$ is a sequence of ϵ-optimal policies for x and, by the last lemma,

$$X_t^n = x + \sqrt{2}W_t + \tilde{\nu}^{x^n} = x_n + \sqrt{2}W_t + \nu^{x^n} \in \overline{\mathcal{C}} \tag{52}$$

for all $0 \le t < \infty$ a.s. Repeating the proof of the last lemma we see that also the optimally controlled process

$$X_t = x + \sqrt{2}W_t + \nu_t^x \in \overline{\mathcal{C}} \text{ for all } 0 \le t < \infty \text{ a.s.} \tag{53}$$

2.4. The optimal policy acts only on $\partial\mathcal{C}$, and its push direction is $-\nabla u$. To prove (c) from Definition 2.5 (with $G = \mathcal{C}$, $v = -\nabla u$), we would like to use Meyer's version of Itô's rule for semimartingales to $e^{-t}u(X_t)$ and proceed similarly as in the proof of Theorem 3.1 in [29]. The problem is that we cannot assume $u \in C^2$, only $u \in W^{2,\infty}_{loc}$. A standard way to overcome this difficulty is to use a regularization of u by convolutions (see, e.g., [9, proof of Theorem VII.4.1]). Another regularization, perhaps more natural in this context, will be discussed in remark (2) in section 2.7. (Alternatively, we can apply Itô's formula directly to our case, without going through the regularization, if we use the results from [32].)

Let $\phi \in C^\infty$, $\phi \geq 0$, supp $\phi \subseteq B_1 = \{x : |x| \leq 1\}$, $\int_{R^n} \phi = 1$. Let $\phi_\epsilon = \frac{\phi(\frac{x}{\epsilon})}{\epsilon^n}$ and let

$$\tilde{u}_\epsilon(x) = (u * \phi_\epsilon)(x) = \int_{R^n} u(y)\phi(x-y)dy, \tag{54}$$

$\tilde{u}_\epsilon \in C^\infty$. Let $h^\epsilon = h * \phi_\epsilon$.

From (16)

$$u - \triangle u \leq h, \qquad |\nabla u| \leq 1, \tag{55}$$

so

$$\tilde{u}_\epsilon - \triangle \tilde{u}_\epsilon \leq h^\epsilon, \qquad |\nabla \tilde{u}_\epsilon| \leq 1 \text{ in } R^n. \tag{56}$$

Let $T > 0$. Using Meyer's version of Itô's rule for semimartingales [27, pp. 278, 301], we get

$$\begin{aligned} E^x(e^{-T}\tilde{u}_\epsilon(X_T)) = \tilde{u}_\epsilon(x) &\\ + E^x \int_0^T e^{-t}(\triangle \tilde{u}_\epsilon - \tilde{u}_\epsilon)(X_t)dt &\\ + E^x \int_0^T e^{-t}\nabla \tilde{u}_\epsilon(X_t)N_t d\xi_t &\\ + E^x \left\{ \sum_{0 \leq t < T} e^{-t}(\tilde{u}_\epsilon(X_{t+}) - \tilde{u}_\epsilon(X_t) - \nabla \tilde{u}_\epsilon(X_t)N_t(\xi_{t+} - \xi_t)) \right\}. & \end{aligned}$$

(The last term keeps account of jumps of X_t.)

By (56), we get

$$\begin{aligned} E^x(e^{-T}\tilde{u}_\epsilon(X_T)) + E^x \int_0^T e^{-t}h^\epsilon(X_t)dt - E^x \int_0^T e^{-t}\nabla \tilde{u}_\epsilon(X_t)N_t d\xi_t &\\ + E^x \left\{ \sum_{0 \leq t < T} e^{-t}(\tilde{u}_\epsilon(X_t) - \tilde{u}_\epsilon(X_{t+}) + \nabla \tilde{u}_\epsilon(X_t)N_t(\xi_{t+} - \xi_t)) \right\} &\\ \geq \tilde{u}_\epsilon(x). & \end{aligned} \tag{57}$$

Now, $X_t \in \overline{\mathcal{C}}$ for all $t \geq 0$ a.s. and, by Lemma 4.4 from [29], $\mathcal{C}$ is bounded. Because u, ∇u, $D^2 u$ are bounded on $B_R = \{x : |x| \leq R\}$, where $\overline{\mathcal{C}} \subseteq B_{R-1}$, we infer that h^ϵ, $\tilde{u}_\epsilon$, $\epsilon < 1$ are bounded uniformly on $\overline{\mathcal{C}}$ and

$$\tilde{u}_\epsilon \to u, \quad \nabla \tilde{u}_\epsilon \to \nabla u, \quad h^\epsilon \to h \quad \text{uniformly in } \overline{\mathcal{C}}. \tag{58}$$

On the other hand,

$$u(x) = E^x \int_0^\infty e^{-t}[h(X_t)dt + d\xi_t] \tag{59}$$

because ν is the optimal policy. In particular,

$$E^x \int_0^\infty e^{-t}d\xi_t < \infty, \tag{60}$$

so

$$E^x \int_0^T e^{-t} d\xi_t < \infty. \tag{61}$$

Thus, using the bounded convergence theorem, we get from (57)

$$\begin{aligned} &E^x(e^{-T}u(X_T)) + E^x \int_0^T e^{-t} h(X_t)dt - E^x \int_0^T e^{-t} \nabla u(X_t) N_t d\xi_t \\ &+E^x \left\{ \sum_{0 \le t < T} e^{-t}(u(X_t) - u(X_{t+}) + \nabla u(X_t) N_t(\xi_{t+} - \xi_t)) \right\} \\ &\qquad \ge u(x). \end{aligned} \tag{62}$$

The last term on the left-hand side is nonpositive because of convexity of u (recall that, by (48) and (47), $N_t(\xi_{t+} - \xi_t) = X_{t+} - X_t$), so we arrive at

$$E^x(e^{-T}u(X_T)) + E^x \int_0^T e^{-t} h(X_t)dt - E^x \int_0^T e^{-t} \nabla u(X_t) N_t d\xi_t \ge u(x). \tag{63}$$

Letting T go to infinity we get, because of boundedness of X_t, $|\nabla u| \le 1$, $|N_t| = 1$, and (60),

$$E^x \int_0^\infty e^{-t} h(X_t)dt - E^x \int_0^\infty e^{-t} \nabla u(X_t) N_t d\xi_t \ge u(x). \tag{64}$$

Using this inequality and (59) we arrive at

$$E^x \int_0^\infty e^{-t}(-\nabla u(X_t)) N_t d\xi_t \ge E^x \int_0^\infty e^{-t} d\xi_t, \tag{65}$$

i.e.,

$$E^x \int_0^\infty e^{-t}(1 + \nabla u(X_t) N_t) d\xi_t \le 0. \tag{66}$$

But $|\nabla u(X_t) N_t| \le 1$ by $|N_t| = 1$, $|\nabla u| \le 1$, so (66) gives

$$-\nabla u(X_t) N_t = 1, \quad d\xi \text{ a.e. on } [0, \infty), \quad P \text{ a.s.} \tag{67}$$

In particular, because $|\nabla u| < 1$ in $\mathcal{C}$,

$$X_t \in R^n - \mathcal{C}, \quad d\xi \text{ a.e. on } [0, \infty), \quad P \text{ a.s.}, \tag{68}$$

so

$$X_t \in \partial\mathcal{C}, \quad d\xi \text{ a.e. on } [0, \infty), \quad P \text{ a.s.} \tag{69}$$

But if $X_t \in \partial\mathcal{C}$, $|\nabla u(X_t)| = 1$, so (67) can hold only if

$$N_t = -\nabla u(X_t), \quad d\xi \text{ a.e. on } [0, \infty), \quad P \text{ a.s.} \tag{70}$$

Statements (69) and (70) prove the condition (c) from Definition 2.5.

2.5. The optimally controlled Brownian motion is a Markov process. Let V be the set of admissible policies (described in subsection 2.1). It is known that for singular stochastic control problems the Bellman principle holds; i.e., for every stopping time T of the filtration $(\mathcal{F}_t)$ and every $x \in R^n$

$$u(x) = \inf_{\tilde{\nu}} \left\{ E^x \int_0^T e^{-t}[h(\tilde{X}_t)dt + d\tilde{\xi}_t] + E^x(e^{-T}u(\tilde{X}_T)) \right\}, \tag{71}$$

where $\tilde{X}_t$, $\tilde{\nu}_t$, $\tilde{\xi}_t$ are as in (45), (43) (see [9, section VIII.5]). We want to prove that this infimum is actually attained for the optimal policy.

LEMMA 2.10. *Let $\nu_t = \int_0^t N_s d\xi_s$ be the optimal policy for x and let the process X_t defined by* (48) *be the optimally controlled Brownian motion starting at x. Then*

$$u(x) = E^x \int_0^T e^{-t}[h(X_t)dt + d\xi_t] + E^x e^{-T}u(X_T). \tag{72}$$

Proof. Let

$$V_\epsilon = \left\{ \nu_t \in V : \ \nu_t = \int_0^t \dot{\nu}_s ds : |\dot{\nu}_s| \le \frac{1}{\epsilon} \ \text{ for all } \ s \ge 0 \right\} \tag{73}$$

be a subclass of controls allowed in our problem. The value function

$$u^\epsilon(x) = \inf_{\tilde{\nu} \in V_\epsilon} \left\{ E^x \int_0^\infty e^{-t}[h(\tilde{X}_t)dt + d\tilde{\xi}_t] \right\} \tag{74}$$

converges to u uniformly on compact sets as $\epsilon \to 0$ [26] (see also the appendix). By the classical control theory for diffusion processes, a unique optimal policy $\tilde{\nu}_t^\epsilon = \int_0^t \tilde{N}_s^\epsilon d\tilde{\xi}_s^\epsilon$ for the problem (74) exists and

$$u^\epsilon(x) = E^x \int_0^T e^{-t}[h(\tilde{X}_t^\epsilon)dt + d\tilde{\xi}_t^\epsilon] + E^x e^{-T}u(\tilde{X}_T^\epsilon), \tag{75}$$

where

$$\tilde{X}_t^\epsilon = x + \sqrt{2}W_t + \tilde{\nu}_t^\epsilon. \tag{76}$$

The proof of (75) can be obtained, for example, by a suitable modification of the argument proving Lemma IV.3.1 in [9].

$u^\epsilon(x) \to u(x)$, so $\tilde{\nu}_t^\epsilon$ are ϵ-optimal policies for the problem (11)–(12) and the point x. By Lemma 2.3, there exists a sequence $\epsilon_k \downarrow 0$ s.t.

$$\nu_t^{\epsilon_k} \to \nu_t \tag{77}$$

for $L_{eb} \times P$ almost all (t, ω).

Assume, for simplicity, that

$$X_T^{\epsilon_k} \to X_T \tag{78}$$

a.s. and that the stopping time T is finite a.s. (We shall remove these assumptions later.) Then, by (77) and (78), letting $k \to \infty$ in (75) for $\epsilon = \epsilon_k$, we get, by Fatou's lemma,

$$\begin{aligned} u(x) &= \liminf_{k\to\infty} u^{\epsilon_k}(x) \\ &= \liminf_{k\to\infty} \left(E^x \int_0^T e^{-t}[h(\tilde{X}_t^\epsilon)dt + d\tilde{\xi}_t^\epsilon] + E^x e^{-T} u(\tilde{X}_T^\epsilon) \right) \\ &\geq E^x \liminf_{k\to\infty} \left(\int_0^T e^{-t}[h(\tilde{X}_t^\epsilon)dt + d\tilde{\xi}_t^\epsilon] + e^{-T} u(\tilde{X}_T^\epsilon) \right) \\ &= E^x \int_0^T e^{-t}[h(X_t)dt + d\xi_t] + E^x e^{-T} u(X_T). \end{aligned} \tag{79}$$

(A similar argument was used in the proof of Theorem 8 in [26].)

This, together with (71), yields (72) for a finite T under the assumption (78). Now suppose that (78) does not hold. However, by (77), there exists $\epsilon_k \downarrow 0$ s.t. (78) holds for $T + \epsilon_k$ instead of T. Using (72) for $T + \epsilon_k$ and letting $k \to \infty$ we get, again by Fatou,

$$\begin{aligned} u(x) &\geq E^x \int_0^{T+} e^{-t}[h(X_t)dt + d\xi_t] + E^x e^{-T} u(X_{T+}) \\ &= E^x \int_0^T e^{-t}[h(X_t)dt + d\xi_t] \\ &\quad + E^x e^{-T}(|X_{T+} - X_T| + (u(X_{T+}) - u(X_T)) + u(X_T)) \\ &\geq E^x \int_0^T e^{-t}[h(X_t)dt + d\xi_t] + E^x(e^{-T} u(X_T)) \end{aligned} \tag{80}$$

because $|\nabla u| \leq 1$; so, by (71) and (80), (72) holds for a finite T. The latter assumption may now be removed by considering (72) for $T \wedge n$ and another limiting procedure. Let us remark that if $P[T = \infty] > 0$, we do not have a problem with the interpretation of (72), because u is bounded in $\overline{\mathcal{C}}$, in which X_t takes its values a.s. In this case we can either regard $e^{-T}u(X_T)$ on $[T = \infty]$ as 0 or replace the last term in (72) by $E^x[I_{[t<\infty]} e^{-T} u(X_T)]$. The proof of Lemma 2.10 is complete.

In fact, (63) (the argument used to prove it goes through for a stopping time T also) and the above considerations yield another proof of (71) (for the problem considered here).

To proceed further, we need the following lemma.

LEMMA 2.11. *Let $T > 0$ be a constant. For PX_T^{-1} a.e. $\tilde{x} \in \overline{\mathcal{C}}$ the following statement is true:*

$$\vartheta_t^{\tilde{x}} := \nu_{t-T} - \nu_T \tag{81}$$

is the optimal policy controlling $\tilde{x} + \sqrt{2}\tilde{W}_t$, where $\tilde{W}_t = W_{t-T} - W_T$ is a Brownian motion starting from $\tilde{x}$ under the measure $\tilde{P}^{\tilde{x}} = P(\cdot | X_T = \tilde{x})$ (by this we mean the value of the regular conditional probability distribution of $(W_t, \nu_t, t \geq 0)$ given the σ-field $\sigma(X_T)$ on the event $X_T = \tilde{x}$).

Proof. We have

$$u(x) = E^x \int_0^T e^{-t}[h(X_t)dt + d\xi_t] + E^x \left(e^{-T} E^{X_T} \int_T^\infty e^{-(t-T)}[h(X_t)dt + d\xi_t] \right), \tag{82}$$

where $E^{X_T} = E(.|X_T)$ is the conditional expectation operator. For simplicity, assume that ν_t (ξ_t) is left-continuous for *all* $\omega \in \Omega$. This can be achieved by modifying ν on a set of measure zero, which clearly does not lead to any difficulty. Let

$$Y = \int_0^\infty e^{-t}[h(X_{t+T})dt + d\xi_{t+T}]. \tag{83}$$

Y is a $\sigma(W_t, \nu_t, t \geq 0)$-measurable random variable. As is well known (see, e.g., [28, Theorem 3, p. 174], which is stated in the one-dimensional case, but whose proof goes through also in n dimensions), for some Borel function $f : \mathcal{R}^n \to \mathcal{R}$

$$E^{X_T} Y = f(X_T) \tag{84}$$

$(PX_T^{-1}$ a.e.), i.e. for PX_T^{-1} a.e. $\tilde{x} \in \overline{\mathcal{C}}$,

$$f(\tilde{x}) = \tilde{E}^{\tilde{x}} \int_0^\infty e^{-t}[h(X_{t+T})dt + d\xi_{t+T}], \tag{85}$$

where the last expectation is the value of $E^{X_T} Y$ on $[X_T = \tilde{x}]$. Equation (82) combined with (83) and (84) yields

$$u(x) = E^x \int_0^T e^{-t}[h(X_t)dt + d\xi_t] + E^x(e^{-T} f(X_T)). \tag{86}$$

We want to prove that $f(\tilde{x}) = u(\tilde{x})$, PX_T^{-1} a.e., on $\overline{\mathcal{C}}$. $\tilde{W}_t$ is a Brownian motion independent on $\mathcal{F}_t$ (in particular on X_T), so it may be seen that $f(\tilde{x})$ is a payoff of the form (11) for the Brownian motion $\tilde{W}_t$ on $(\Omega, \mathcal{F}_t, \tilde{P}^{\tilde{x}})$ (starting at $\tilde{x}$). Indeed, $\tilde{E}^{\tilde{x}} \int_0^\infty e^{-t}[h(X_{t+T})dt + d\xi_{t+T}]$ defined as above is $(PX_T^{-1}$ a.e.) the expectation of Y under $\tilde{P}^{\tilde{x}}$. This can be seen by approximating Y by suitable finite sums, as in the definition of an integral, evaluating their expectations under $\tilde{P}^{\tilde{x}}$, and then going to the limit, using the bounded and monotone convergence theorems. Also, $\xi_\cdot(\omega)$ is left-continuous for all $\omega \in \Omega$ by assumption. Moreover, the value function for the latter control problem is again u. (The same argument as that for W_t shows that it is a nonnegative, convex $W_{loc}^{2,\infty}$ solution to (16), which is unique by Theorem 4.5 of [29].)

The above considerations lead to

$$u \leq f \tag{87}$$

$(PX_T^{-1}$ a.e.) on $\overline{\mathcal{C}}$.

We want to show that, in fact, equality holds in (87). Suppose it is not true, i.e.,

$$PX_T^{-1}(A) > 0, \tag{88}$$

where

$$A = \{\tilde{x} \in \overline{\mathcal{C}} : u(\tilde{x}) < f(\tilde{x})\}. \tag{89}$$

Of course, (87), (88), and (89) yield

$$E^x e^{-T} u(X_T) < E^x e^{-T} f(X_T). \tag{90}$$

Then, by (72), (90), and (86),

$$u(x) < E^x \int_0^T e^{-t}[h(X_t)dt + d\xi_t] + E^x(e^{-T} f(X_T)) = u(x), \tag{91}$$

a contradiction.

Thus, $u(\tilde{x}) = f(\tilde{x})$ for PX_T^{-1} a.e. $\tilde{x} \in \overline{\mathcal{C}}$. But, by the definitions of u and f, this means exactly that, for PX_T^{-1} a.e. $\tilde{x} \in \overline{\mathcal{C}}$, $\vartheta_t^{\tilde{x}}$ defined by (81) is indeed the optimal policy for $\tilde{W}_t$ under $\tilde{P}^{\tilde{x}}$. The proof is complete.

One can see that in Lemma 2.11, T can be any stopping time of the filtration $\{\mathcal{F}_t\}$. We must modify the above argument (and the statement of Lemma 2.11) suitably, replacing $\sigma(X_T)$ by $\sigma(T, X_T)$, PX_T^{-1} by $P[T, X_T]^{-1}$, $\tilde{P}^{\tilde{x}}$ by $\tilde{P}^{\tilde{x},t}$ (the value of the regular conditional probability distribution of $(T, W_t, \nu_t, t \geq 0)$ given $\sigma(T, X_T)$ on $[X_T = \tilde{x}, T = t]$), and using the strong Markov property of a Brownian motion.

As a by-product of the above reasoning, in particular of Lemma 2.11 for a stopping time T, we get that the optimally controlled process X_t is a strong Markov process with respect to the filtration $\{\mathcal{F}_t,\ t \geq 0\}$. Indeed, for any $t > 0$

$$X_{T+t} - X_T = (W_{T+t} - W_T) + (\nu_{T+t} - \nu_T),$$

the Brownian increment is independent on $\mathcal{F}_T$, and all the relevant information about the increment of ν that can be found in $\mathcal{F}_T$ is, as we have just seen, the value of X_T. Thus, all the information about

$$X_{T+t} = (X_{T+t} - X_T) + X_T$$

that can be found in $\mathcal{F}_T$ is actually the value of X_T.

2.6. Possible jumps of the optimal policy. Suppose that $x \in \partial\mathcal{C}$ has the following property.

There exists an interval $I \subseteq \partial\mathcal{C}$ such that

$$I = \{a + t\eta : t \in [0, c]\} \tag{92}$$

for some $\eta \in R^n$, $|\eta| = 1$, $a \in R^n$, $c > 0$, $\nabla u \equiv \eta$ on I, and $x \in I - \{a\}$.

We shall denote the set of all such x by D. We assume that I in the above definition is maximal, i.e., is the sum of all the intervals with such property. Then

$$u(a + \eta t) = u(a) + t, \qquad t \in [0, c], \tag{93}$$

because

$$\frac{\partial u}{\partial \eta} = \nabla u \cdot \eta = |\nabla u|^2 = 1 \ \text{ on } \ I. \tag{94}$$

Analytically, D can be defined by

$$D = \bigcup_{i=1}^{\infty} \left\{ x \in \partial\mathcal{C} :\ u(x) - u\left(x - \frac{1}{i}\nabla u(x)\right) = \frac{1}{i} \right\}, \tag{95}$$

so D is a countable sum of closed sets, in particular a Borel measurable set.

Let ν^x be the optimal policy for the process starting from $x \in D$ and let $x = a + t\eta$, a, t, and η as above. Then

$$\nu_t^x = \begin{cases} 0 & \text{if } t = 0, \\ a - x + \nu_t^a & \text{if } t > 0; \end{cases} \tag{96}$$

i.e., the optimal policy first jumps from x to a and proceeds optimally thereafter.

Indeed, if we define ν^x by (96), then, by (93),

$$J_x(\nu^x) = |a - x| + u(a) = t + u(a) = u(x). \tag{97}$$

Now we shall analyze the jumps of the optimal policy ν^x for an arbitrary $x \in \overline{\mathcal{C}}$.

LEMMA 2.12. *For every starting state $x \in \overline{\mathcal{C}}$ a.s. $\omega \in \Omega$ the only discontinuities of $X_\cdot(\nu^x_\cdot)$ are possibly jumps of the type described in Definition* 2.5(a) *(for $G = \mathcal{C}$, $v = -\nabla u$) at $X_t \in D$.*

Proof. First we want to prove that, a.s., X_t jumps only when $X_t \in D$. Suppose it is not true. Because ν^x is left-continuous, the only possible discontinuities of ν^x are jumps. Suppose ν^x does have jumps and let

$$T^\epsilon(\omega) = \inf\{t \geq 0 : X_t(\omega) \notin D, |X_t(\omega) - X_{t+}(\omega)| \geq \epsilon\}; \tag{98}$$

i.e., T^ϵ is the first time the process X_t undergoes a jump of magnitude at least ϵ starting outside D. Suppose ϵ is so small that $P[T^\epsilon < \infty] > 0$. By a theorem in [5, p. 84] (see also [18, p. 10]), there is a sequence $T_1, T_2, \ldots$ of stopping times exhausting the jumps of X_t (ν_t), i.e.,

$$\{(t,\omega) \in [0,\infty) \times \Omega : X_t(\omega) \neq X_{t+}(\omega)\} \in \sum_{i=1}^{\infty} \{(t,\omega) \in [0,\infty) \times \Omega : T_i(\omega) = t\}. \tag{99}$$

(This is slightly different than the original theorem in [5], which dealt with right-continuous, instead of left-continuous, processes, but it is easy to see that if the filtration itself is right-continuous, our version follows from the original one in a straightforward way.) Thus,

$$T^\epsilon(\omega) = \inf\{T_i(\omega) : X_{T_i(\omega)} \notin D, |X_{T_i(\omega)}(\omega) - X_{T_i(\omega)+}(\omega)| \geq \epsilon\}. \tag{100}$$

ν^x_t is a process of bounded variation on each finite interval (a.s.), so, a.s., on each interval $[0,T]$ there can be only finitely many jumps of magnitude at least ϵ, so in (100) we can use min instead of inf. Thus, T^ϵ is a stopping time. (It was not entirely obvious from the very beginning, because D does not have to be open or closed.)

On $[T^\epsilon < \infty]$ $X_{T^\epsilon} \in \overline{\mathcal{C}} - D$, $X_{T^\epsilon +} \in \overline{\mathcal{C}}$, so, by $|\nabla u| \leq 1$,

$$u(X_{T^\epsilon(\omega)+}(\omega)) + |X_{T^\epsilon(\omega)+}(\omega) - X_{T^\epsilon(\omega)}(\omega)| > u(X_{T^\epsilon(\omega)}(\omega)) \tag{101}$$

because ∇u is not identically equal to -1 on the interval joining $X_{T^\epsilon(\omega)}(\omega)$ to $X_{T^\epsilon(\omega)+}(\omega)$.

Using (72) for $T^\epsilon + \frac{1}{n}$, then letting $n \to \infty$ and using the bounded convergence theorem as in (80), we get

$$\begin{aligned} u(x) &= E^x \int_0^{T^\epsilon(\omega)} e^{-t}[h(X_t)dt + d\xi_t] \\ &\quad + E^x(e^{-T^\epsilon(\omega)}|X_{T^\epsilon(\omega)+}(\omega) - X_{T^\epsilon(\omega)}(\omega)|I_{[T^\epsilon<\infty)}) \\ &\quad + E^x(I_{[T^\epsilon<\infty)}e^{-T^\epsilon(\omega)}u(X_{T^\epsilon(\omega)+})) \\ &> E^x \int_0^{T^\epsilon(\omega)} e^{-t}[h(X_t)dt + d\xi_t] \\ &\quad + E^x(I_{[T^\epsilon<\infty)}e^{-T^\epsilon(\omega)}u(X_{T^\epsilon(\omega)})) \\ &= u(x); \end{aligned} \tag{102}$$

the inequality follows from (101), and the last equality follows from (72). But (102) is a clear contradiction. Thus, a.s., X_t jumps only when $X_t \in D$.

The fact that for $X_t \in D$ the optimally controlled process X_t a.s. jumps to the endpoint of the interval I from the Definition 2.5(a), where $X_t \in I$, follows from (96) and Lemma 2.11 combined with the fact that there are at most countably many jumps of X_t (see (99)). The lemma is proved.

This ends the proof of Theorem 2.6.

COROLLARY 2.13. *If $D = \emptyset$, then X_t, ξ_t, ν_t are continuous.*

COROLLARY 2.14. *Let $x \in D$. Then there exists an interval I as in* (92) *such that $x = a + t\eta$, $t > 0$. Then ν^a is continuous at* 0, *because $a \notin D$. Thus, by* (96), *the optimal policy ν^x starting from x jumps immediately to a and $\lim_{t\downarrow 0} \nu_t^x = a$.*

2.7. Remarks and supplements. (1) Using Lemma 2.9, the remark following it, the boundedness of $\mathcal{C}$, (70), and letting $T \to \infty$ in (62) we get

$$\begin{aligned} E^x \int_0^\infty & e^{-t}[h(X_t)dt + d\xi_t] \\ & + E^x \left\{ \sum_{0 \le t < \infty} e^{-t}(u(X_t) - u(X_{t+}) + \nabla u(X_t) N_t(\xi_{t+} - \xi_t)) \right\} \\ & \ge u(x). \end{aligned}$$

But, by (59), it means that the second expectation is 0, so all the terms

$$u(X_t) - u(X_{t+}) + \nabla u(X_t) N_t(\xi_{t+} - \xi_t) = 0, \tag{103}$$

P a.s. (because, as we have explained in the justification of (63), these terms are all nonpositive). But $N_t = -\nabla u(X_t)$ and $|\nabla u| \le 1$, so, by $\xi_{t+} - \xi_t = |X_{t+} - X_t|$, we have

$$\frac{\partial u}{\partial v} \equiv 1 \ \text{ if } \ v = \frac{X_{t+} - X_t}{|X_{t+} - X_t|} \tag{104}$$

on the whole interval joining X_t to X_{t+}. Thus, because $X_t \in \overline{\mathcal{C}}$ for every $t \ge 0$ a.s., we conclude that, for a.e. $\omega \in \Omega$ and all $t \ge 0$ s.t. $X_t \ne X_{t+}$, $[X_t, X_{t+}]$ is contained in some interval I of the form given in (a) of the Definition 2.5. However, this method does not ensure that X_{t+} is actually the endpoint of such I, and neither does it ensure that X_t *must* jump at points of D. This is the reason why we need an argument like that given in sections 2.5–2.6.

(2) Another way of regularizing the value function u in the proof given in subsection 2.4 is to use, instead of $\tilde{u}_\epsilon$, the function u^ϵ defined by (74), which is the value function for an approximating control problem (see the appendix; compare also [35, proof of Theorem 4.3]). u^ϵ satisfies

$$u^\epsilon - \triangle u^\epsilon + \frac{1}{\epsilon} \max(|\nabla u^\epsilon| - 1, 0) = h \ \text{ in } \ R^n. \tag{105}$$

Now we cannot claim that $|\nabla u^\epsilon| \le 1$, but from the estimates in [26] we know that u^ϵ, ∇u^ϵ, $D^2 u^\epsilon$ are bounded uniformly on B_R ($\overline{\mathcal{C}} \subseteq B_{R-1}$) for $\epsilon > 0$. Moreover, for a subsequence (still denoted by u^ϵ) $u_\epsilon \to u$, $\nabla u_\epsilon \to \nabla u$ uniformly in B_R and, from (105), $u^\epsilon - \triangle u^\epsilon \le h$, so we can repeat the proof of subsection 2.4 with u^ϵ instead of $\tilde{u}_\epsilon$ and h instead of h^ϵ.

(3) Let the starting point $x \in R^n - \overline{\mathcal{C}}$. By a modification of the argument given in [29] we can show that in this case the optimal policy jumps immediately to some point $\hat{x} \in \partial\mathcal{C}$ and then follows the optimal policy $\nu^{\hat{x}}$. We provide the details for the sake of completeness.

Let us introduce, following [29], the following change of coordinates. Let y_0 be a point at which the minimum of u in R^n is attained. By Lemma 4.4 in [29], it is unique and belongs to $\mathcal{C}$. Choose $\delta > 0$, $\mu > 0$ s.t.

$$B_{2\delta}(y_0) \subseteq \mathcal{C}, \tag{106}$$

$$D^2u(\tilde{x})y \cdot y \geq \mu|y|^2 \tag{107}$$

for all $\tilde{x} \in B_{2\delta}(y_0)$ (recall that u is strictly convex in $\mathcal{C}$ by (4.10) of [29]),

$$\mu \leq |\nabla u(\tilde{x})|^2 \leq \frac{1}{2} \tag{108}$$

for all $\tilde{x} \in \partial B_\delta(y_0)$, and

$$\nabla u(y_0 + \delta\theta) \cdot \theta \geq \mu \tag{109}$$

for all $\theta \in S_1$, where S_1 is the unit sphere. For $\theta \in S_1$, we define the gradient flow $\psi(t, \theta)$ by

$$\frac{d}{dt}\psi(t, \theta) = \nabla u(\psi(t, \theta)) \tag{110}$$

for all $t \geq 0$ and

$$\psi(0, \theta) = y_0 + \delta\theta. \tag{111}$$

Let $x \in R^n - B_\delta(y_0)$, $x = \psi(t, \theta)$. We assign to x the coordinates (t, θ). It was shown in [29] that this change of variables is a homeomorphism between $R^n - B_\delta(y_0)$ and $[0, \infty) \times S_1$. (This part of the argument given there goes through in n dimensions; compare remarks on p. 332 of [9].) Now suppose that x, the starting point of the process (2), belongs to $R^n - \overline{\mathcal{C}}$. Let θ_0 be s.t. $x = \psi(t_0, \theta_0)$ for some $t_0 > 0$ and let

$$t_1 = \inf\{t \leq t_0 : \psi(t, \theta_0) \in \partial\mathcal{C}\}.$$

(This set is nonempty because of (106), (111), and the fact that, by assumption, x does not belong to $\overline{\mathcal{C}}$.) Let $\hat{x} = \psi(t_1, \theta_0)$. Then $\hat{x} \in \partial\mathcal{C}$, so the vector $v := \nabla u(\hat{x})$ has the norm 1. Let

$$L = \{\hat{x} + tv : t \geq 0\}.$$

Then, by $|v| = 1$, convexity of u, and the fact that $|\nabla u| \leq 1$, we have $u(\hat{x} + tv) = u(\hat{x}) + t$ for $t \geq 0$, in particular, $\nabla u(\hat{x} + tv) = v$, $t \geq 0$. But this means that, by the definition of the gradient flow ψ, for $t \geq t_1$,

$$\psi(t, \theta_0) = \hat{x} + (t - t_1)v; \tag{112}$$

so, in particular, $x = \hat{x} + (t_0 - t_1)v$ (put $t = t_0$ in (112)) and $\nabla u \equiv v$ on L. Thus,

$$u(x) = u(\hat{x}) + (t_0 - t_1)$$

and

$$|x - \hat{x}| = t_0 - t_1;$$

so, by an argument analogous to that given at the beginning of subsection 2.6,

$$\nu_t^x = \hat{x} - x + \nu_t^{\hat{x}}$$

for every $t > 0$ P-a.s.

(4) It is clear that any process that solves the Skorokhod problem for W_t, $\mathcal{C}$, $-\nabla u$ in the sense of Definition 2.5 is an optimal policy for our problem (compare, e.g., the verification Theorem VIII.4.1 from [9]). Thus, uniqueness of the optimal policy implies uniqueness of a solution to the modified Skorokhod problem for W_t, $\mathcal{C}$, $-\nabla u$.

Usually, some assumptions about regularity of the boundary of a region are necessary to prove existence and uniqueness of a solution to the Skorokhod problem. Here, all such assumptions are hidden in the very nature of the stochastic control problem and, if $n \geq 3$ we do not know what they are, i.e., how regular $\partial\mathcal{C}$ really is. The conjecture is that for higher dimensions "smooth fit" holds also, $\partial\mathcal{C}$ is smooth, say $C^{2,\alpha}$, and ∇u is nontangential to $\partial\mathcal{C}$, so the optimal policy is a solution to the Skorokhod problem in the usual sense, as in the two-dimensional case.

3. Appendix. In this paper, we have used some results from [26], where a problem slightly different from minimizing (11) was considered. However, it is easy to establish analogous results in our case by essentially the same techniques. The aim of this appendix is to provide a concise discussion of some minor adjustments which make the proofs from [26] work in the case of (11) and (12) and to sketch the main idea of the $W_{loc}^{2,\infty}$ regularity result contained in [26]. Compare also [31], where the results from [26] are used for the problem (11)–(12) with no additional comments.

We make all the assumptions of section 2.1. For the proof of Lemma 2.2 in section 2 of this paper, first let us remark that, by strict convexity of h,

$$\triangle_2 h(x,y) = \frac{h(x) + h(y)}{2} - h\left(\frac{x+y}{2}\right) > 0 \tag{113}$$

for all $x, y \in R^n$, $x \neq y$, so the proof of Theorem 7 in [26] also goes through in our case. Indeed, if we have two optimal policies

$$\nu_t^1 = \int_0^t N_s d\zeta_s, \tag{114}$$

$$\nu_t^2 = \int_0^t M_s d\eta_s, \tag{115}$$

take

$$\nu_t = \frac{\nu_t^1 + \nu_t^2}{2} = \int_0^t \tilde{N}_s d\tilde{\xi}_s. \tag{116}$$

We have

$$\xi_t = \bigvee_{[0,t]} \left[\frac{\nu_s^1 + \nu_s^2}{2}\right] \leq \frac{1}{2}\left[\bigvee_{[0,t]} \nu_s^1 + \bigvee_{[0,t]} \nu_s^2\right] = \frac{\zeta_t + \eta_t}{2}, \tag{117}$$

which allows us to use the proof in [26]. Also, we can mimic the proof of Theorem 8 from that paper and the corollary following it to get the proof of our Lemma 2.3.

The argument proving that the value function u defined by (12) is in $W^{2,\infty}_{loc}$ and (16) holds is essentially like that given in [26]. Its main idea is easy to understand: we begin with an approximating problem of controlling (11) in the class V_ϵ defined by (73), which is a subclass of controls allowed in our problem. Then a standard argument (see, e.g., [22, Theorem 5, p. 289]) shows that the value function u^ϵ defined by (74) belongs to $W^{2,\infty}_{loc}$ and satisfies the Bellman equation (105), so in fact, by elliptic regularity, $u^\epsilon \in C^{2,\alpha}$ for every $\alpha \in (0,1)$. One gets, by a direct probabilistic approach, a priori estimates for u^ϵ, ∇u^ϵ, and D^2u^ϵ independent of ϵ by estimating appropriate difference quotients. Then, using $u^\epsilon \to u$, we have the same estimates for u, and an easy additional argument using (105) convinces us that (16) holds. In fact, convergence of u^ϵ to u, although intuitively obvious (every function of bounded variation on finite intervals can be easily approximated by elements of V_ϵ as $\epsilon \to 0$), is surprisingly technical to prove (see [26]).

For another, PDE-based proof of existence of a $W^{2,\infty}_{loc}$ solution to the Bellman equation (16) see [8, 15, 29].

Acknowledgments. I want to express my deep gratitude to my advisor, Professor Marco Avellaneda, for many helpful discussions. I also thank the referees for their valuable suggestions.

REFERENCES

[1] J. A. Bather and H. Chernoff, *Sequential decisions in the control of a spaceship*, in Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability 3, University of California Press, Berkeley, CA, 1967, pp. 181–207.

[2] J. A. Bather and H. Chernoff, *Sequential decisions in the control of a spaceship (finite fuel)*, J. Appl. Probab., 49 (1967), pp. 584–604.

[3] V. E. Beneš, L.A. Shepp, and H. S. Witsenhausen, *Some solvable stochastic control problems*, Stochastics, 4 (1980), pp. 181–207.

[4] P. L. Chow, J.-L. Menaldi, and M. Robin, *Additive control of stochastic linear systems with finite horizon*, SIAM J. Control Optim., 23 (1985), pp. 858–899.

[5] C. Dellacherie, *Capacités et Processus Stochastiques*, Springer-Verlag, Berlin, 1972.

[6] A. Dixit, *A simplified treatment of the theory of optimal regulation of Brownian motion*, J. Econom. Dynam. Control, 15 (1991), pp. 675–685.

[7] B. Dumas, *Super contact and related optimality conditions*, J. Econom. Dynam. Control, 15 (1991), pp. 675–685.

[8] L. C. Evans, *A second order elliptic equation with a gradient constraint*, Comm. Partial Differential Equations, 4 (1979), pp. 555–572; erratum, Comm. Partial Differential Equations, 4 (1979), p. 1199.

[9] W. H. Fleming and H.M. Soner, *Controlled Markov Processes and Viscosity Solutions*, Springer-Verlag, New York, 1993.

[10] D. Gilbarg and N. Trudinger, *Elliptic Differential Equations of Second Order*, 2nd ed., Springer-Verlag, New York, 1985.

[11] J. M. Harrison, *Brownian Motion and Stochastic Flow Systems*, Wiley, New York, 1985.

[12] J. M. Harrison and M. I. Taksar, *Instantaneous control of Brownian motion*, Math. Oper. Res., 8 (1983), pp. 454–466.

[13] J. M. Harrison and A. J. Taylor, *Optimal control of a Brownian storage system*, Stochastic Process Appl., 6 (1978), pp. 179–194.

[14] N. Ikeda and S. Watanabe, *Stochastic Differential Equations and Diffusion Processes*, 2nd ed., North Holland-Kodansha, Amsterdam, Tokyo, 1989.

[15] H. Ishii and S. Koike, *Boundary regularity and uniqueness for an elliptic equation with gradient constraint*, Comm. Partial Differential Equations, 8 (1983), pp. 317–346.

[16] I. Karatzas, *The monotone follower problem in stochastic decision theory*, Appl. Math. Optim., 7 (1981), pp. 175–189.

[17] I. Karatzas, *A class of singular stochastic control problems*, Adv. in Appl. Probab., 15 (1983), pp. 225–254.
[18] I. Karatzas and S. E. Shreve, *Brownian Motion and Stochastic Calculus*, Springer-Verlag, New York, 1988.
[19] I. Karatzas and S. E. Shreve, *Connections between optimal stopping and singular stochastic control* I. *Monotone follower problems*, SIAM J. Control Optim., 22 (1984), pp. 856–877.
[20] I. Karatzas and S. E. Shreve, *Connections between optimal stopping and singular stochastic control* II. *Reflected follower problems*, SIAM J. Control Optim., 23 (1985), pp. 433–541.
[21] P. R. Krugman, *Target zones and exchange rate dynamics*, Quart. J. Econom., 106 (1991), pp. 669–682.
[22] N. V. Krylov, *Controlled Diffusion Processes*, Springer-Verlag, New York, 1980.
[23] P.-L. Lions and A. S. Sznitman, *Stochastic differential equations with reflecting boundary conditions*, Comm. Pure Appl. Math., 37 (1984), pp. 511–537.
[24] J. L. Menaldi and M. Robin, *On some cheap control problems for diffusion processes*, Trans. Amer. Math. Soc., 278 (1983), pp. 771–802.
[25] J. L. Menaldi, M. Robin, and M. I. Taksar, *Singular ergodic control for multidimensional Gaussian processes*, Math. Control Signals Systems, 5 (1992), pp. 93–114.
[26] J. L. Menaldi and M. I. Taksar, *Optimal correction problem of a multidimensional stochastic system*, Automatica J. IFAC, 25 (1989), pp. 223–232.
[27] P. A. Meyer, *Séminaire de Probabilités* X II: *Théorie des Intégrales Stochastiques (Université de Strassbourg)*, Lecture Notes in Math. 511, Springer-Verlag, Berlin, Heidelberg, New York, 1976.
[28] A. N. Shiriaev, *Probability*, 2nd ed., Springer-Verlag, New York, 1996.
[29] H. M. Soner and S. E. Shreve, *Regularity of the value function for a two-dimensional singular stochastic control problem*, SIAM J. Control Optim., 27 (1989), pp. 876–907.
[30] D. Stroock and S. R. S. Varadhan, *Diffusion processes with boundary conditions*, Comm. Pure Appl. Math., 24 (1971), pp. 147–225.
[31] M. Taksar, *Convex solutions to variational inequalities and multidimensional singular control*, in The Dynkin Festschrift, Markov Processes and Their Applications, Progr. Probab. 34, Birkhäuser, Boston, 1994, pp. 371–386.
[32] A. J. Veretennikov, *On strong solutions of Itô stochastic equations with jumps*, Theory Probab. Appl., 32 (1988), pp. 148–152.
[33] S. Watanabe, *On stochastic differential equations for multidimensional diffusion processes with boundary conditions*, J. Math. Kyôto Univ., 11 (1971), pp. 169–180.
[34] S. A. Williams, P. L. Chow, and J. L. Menaldi, *Regularity of the free boundary in singular stochastic control*, J. Differential Equations, 111 (1994), pp. 175–201.
[35] H. Zhu, *Generalized solution in singular stochastic control: The nondegenerate problem*, Appl. Math. Optim., 25 (1992), pp. 225–245.

SIAM J. CONTROL OPTIM.
Vol. 38, No. 5, pp. 1623–1638

A GENERALIZED MATHEMATICAL PROGRAM WITH EQUILIBRIUM CONSTRAINTS*

J. V. OUTRATA†

Abstract. The paper concerns an optimization problem with a generalized equation among the constraints. This model includes standard mathematical programs with parameter-dependent variational inequalities or complementarity problems as side constraints. Using Mordukhovich's generalized differential calculus, we derive necessary optimality conditions and apply them to problems, where the equilibria are governed by implicit complementarity problems and by hemivariational inequalities.

Key words. generalized equation, coderivative, subdifferential, constraint qualification

AMS subject classifications. 49J40, 49J52, 90C

PII. S0363012999352911

1. Introduction. According to [10], by *mathematical program with equilibrium constraints* (MPEC) one understands an optimization problem where, among the constraints, the following *variational inequality* (VI) arises:

$$\begin{aligned} &\text{for a given } x \in \mathbb{R}^n, \text{ find } y \in \Gamma(x) \text{ such that} \\ &\langle F(x,y), v-y\rangle \geq 0 \quad \text{for all } v \in \Gamma(x), \end{aligned} \tag{1.1}$$

where F maps $\mathbb{R}^n \times \mathbb{R}^m$ into $\mathbb{R}^m$ and Γ is a closed- and convex-valued multifunction mapping $\mathbb{R}^n$ into subsets of $\mathbb{R}^m$. If $\Gamma(x) = \mathbb{R}^m_+$ for all $x \in \mathbb{R}^n$, (1.1) reduces to a *nonlinear complementarity problem* (NCP). In what follows x is called the *control* or *decision parameter* and y the *state variable*. Both these variables arise generally in the objective and are possibly subject to other constraints.

Not all equilibria, however, can be modeled in the form (1.1). In continuum mechanics [3], for instance, we find the so-called VIs of second kind, which, after an appropriate discretization, attain the following form:

$$\begin{aligned} &\text{for a given } x \in \mathbb{R}^n, \text{ find } y \in \Xi \text{ such that} \\ &\langle F(x,y), v-y\rangle + J(v) - J(y) \geq 0 \quad \text{for all } v \in \Xi, \end{aligned} \tag{1.2}$$

where F is the same function as in (1.1), $J[\mathbb{R}^m \to \mathbb{R}]$ is a convex continuous function, and Ξ is a nonempty closed convex subset of $\mathbb{R}^m$. Since J is not differentiable at all points of Ξ, the VI (1.2) cannot be converted to the form (1.1).

Remark. (1.2) describes, e.g., a discretized contact problem with *given* friction; cf. [5]. The state variable corresponds then to the (vector-valued) displacement at the nodes of discretization and x to the body forces and surface tractions.

Further, when modeling the behavior of bodies made from composites (fiberglass or sandwich constructions), one arrives at so-called hemivariational inequalities; cf. [17]. After a discretization, they lead to models which cannot be converted to the form (1.1) either.

*Received by the editors March 3, 1999; accepted for publication (in revised form) December 3, 1999; published electronically May 26, 2000. This research was supported by grant A 1075707 of the Czech Academy of Sciences.

http://www.siam.org/journals/sicon/38-5/35291.html

†Institute of Information Theory and Automation, Academy of Sciences of the Czech Republic, Pod vodárenskou věží 4, 182 08 Prague, Czech Republic (outrata@utia.cas.cz).

Both the VI (1.2) as well as a discretized hemivariational inequality can be, however, rewritten into the form of the *generalized equation* (GE):

$$0 \in F_1(x,y) + Q(F_2(x,y)), \tag{1.3}$$

where F_1, F_2 map $\mathbb{R}^n \times \mathbb{R}^m$ into $\mathbb{R}^m$, $\mathbb{R}^\ell$, respectively, and $Q[\mathbb{R}^\ell \rightsquigarrow \mathbb{R}^m]$ is a multifunction with the closed graph. In the case of the VI (1.2) we put $F_1 := F,\ F_2(x,y) := y$, and

$$Q(y) := \partial J(y) + N_\Xi(y), \tag{1.4}$$

where $\partial J(y)$ is the standard subdifferential of J at y (in the sense of convex analysis), and

$$N_\Xi(y) := \begin{cases} \text{the standard normal cone to } \Xi \text{ at } y & \text{provided } y \in \Xi, \\ \emptyset & \text{otherwise.} \end{cases}$$

In the case of an NCP one has $F_1 := F,\ F_2(x,y) := y$,

$$Q(y) := N_{\mathbb{R}^m_+}(y),$$

and also the VI (1.1) can mostly be converted to the form (1.3), e.g., when Γ is given by means of (control-dependent) equalities and inequalities; cf. [16].

On the basis of these considerations we define the (generalized) MPEC as the optimization problem:

$$\begin{array}{ll} \text{minimize} & f(x,y) \\ \text{subject to} & \\ & 0 \in F_1(x,y) + Q(F_2(x,y)), \\ & (x,y) \in \omega, \end{array} \tag{1.5}$$

where $f[\mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}]$ is an objective and $\omega \subset \mathbb{R}^n \times \mathbb{R}^m$ is the set of *admissible control (decision)-state pairs*. The GE (1.3) models the *equilibrium constraint*.

The aim of this paper is

(i) to derive first-order necessary optimality conditions for (1.5);

(ii) to apply these conditions to two distinguished classes of (1.5) with equilibria governed by implicit complementarity problems and by hemivariational inequalities.

In these investigations our basic tool is the "nonconvex" generalized differential calculus of Mordukhovich [11], [12], [13]. The applied approach utilizes the idea of exact penalization of the equilibrium constraint used already in [22] (in case of equilibria given by (lower-level) optimization problems). Instead of penalizing $\|v\|$ for

$$v \in F_1(x,y) + Q(F_2(x,y)),$$

which would correspond to the approach in [21], our "equilibrium constraint violation" is given by

$$v \in \begin{bmatrix} -F_2(x,y) \\ F_1(x,y) \end{bmatrix} + \operatorname{Gph} Q, \tag{1.6}$$

where $\operatorname{Gph} Q$ denotes the graph of multifunction Q. This enables one to respect easily joint constraints in both variables x and y. Furthermore, it is advantageous when applying a penalty approach to the numerical solution of (1.5).

The paper is organized as follows. The next section is devoted to a general mathematical program, to which form problem (1.5) can be converted. In section 3 the optimality conditions for (1.5) are established. Attention is paid also to the polyhedral case, where the constraint qualification can be omitted. Section 4 deals with the application of the obtained conditions to MPECs with equilibria described by implicit complementarity problems and by hemivariational inequalities.

The following notation is employed: x^i is the ith component of a vector $x \in \mathbb{R}^n$, $\overline{\mathbb{R}}$ is the extended real line, and $\mathbb{R}^n_+$ denotes the nonnegative orthant of $\mathbb{R}^n$. E is the unit matrix. For an $[m \times n]$ matrix A and an index set $I \subset \{1, 2, \ldots, m\}$, A_I is the submatrix of A with rows specified by I. Similarly, for a vector $d \in \mathbb{R}^n$, d_I is the subvector composed from the components d^i, $i \in I$. Furthermore, $epi\, f$ is the epigraph of a function f and $\partial^c f(x)$ is the Clarke's subdifferential of f at x. For a multifunction $Q[\mathbb{R}^n \rightsquigarrow \mathbb{R}^m]$, $\mathrm{Ker} Q := \{x \in \mathbb{R}^n \,|\, 0 \in Q(x)\}$. If D is a cone with vertex at the origin, then D^0 is its negative polar cone. For x, $y \in \mathbb{R}^n$ the inequalities $x \geq y$, $x > y$ mean $x^i \geq y^i$ and $x^i > y^i$ for all i, respectively. $\mathbb{B}$ is the unit ball, $cl\,\Omega$ denotes the closure of a set Ω, and $\mathrm{dist}_\Omega(x)$ is the distance of x to Ω.

For the reader's convenience, we close this section with three fundamental definitions from Mordukhovich's generalized differential calculus used throughout the paper.

Consider a set $\Pi \subset \mathbb{R}^p$.

DEFINITION 1.1. *Let $a \in \mathrm{cl}\,\Pi$. The nonempty cone*

$$T_\Pi(a) := \limsup_{t \downarrow 0} \frac{\Pi - a}{t}$$

is called the contingent cone *to Π at a.*[1] *The* generalized normal cone *to Π at a, denoted $N_\Pi(a)$, is defined by*

$$N_\Pi(a) = \limsup_{a' \overset{cl\ \Pi}{\longrightarrow} a} T^0_\Pi(a').$$

If Π is convex, $N_\Pi(a)$ amounts to the standard normal cone to Π at a in the sense of convex analysis. The cone $N_\Pi(a)$ is generally nonconvex, but the multifunction $N_\Pi(\cdot)$ is upper semicontinuous at each point of $cl\Pi$ (with respect to $cl\Pi$), which is essential in the calculus of Mordukhovich's subdifferentials and coderivatives introduced below.

DEFINITION 1.2. *Let $\varphi[\mathbb{R}^p \to \overline{\mathbb{R}}]$ be an arbitrary extended real-valued function and $a \in dom\,\varphi$. The set*

$$\partial^- \varphi(a) := \{a^* \in \mathbb{R}^p | (a^*, -1) \in N_{epi\,\varphi}(a, \varphi(a))\}$$

is called Mordukhovich's subdifferential *of φ at a.*

DEFINITION 1.3. *Let $\Phi[\mathbb{R}^p \rightsquigarrow \mathbb{R}^q]$ be an arbitrary multifunction and $(a, b) \in cl\,\mathrm{Gph}\,\Phi$. The multifunction $D^*\Phi(a, b)\,[\mathbb{R}^q \to \mathbb{R}^p]$ defined by*

$$D^*\Phi(a, b)\,(b^*) := \{a^* \in \mathbb{R}^p | (a^*, -b^*) \in N_{Gph\,\Phi}(a, b)\}, \quad b^* \in \mathbb{R}^q,$$

is called the coderivative *of Φ at (a, b).*

Besides these notions we also will need several results of Mordukhovich from [11], [12], and [13]; for the reader's convenience three of them are stated in the appendix. For a thorough study of Mordukhovich's theory the reader is referred to [11].

[1] The "lim sup" in the definitions of $T_\Pi(a)$ and $N_\Pi(a)$ is the upper limit of multifunctions in the sense of Kuratowski–Painlevé; cf. [1].

2. Auxiliary results. Consider first a general mathematical program

$$\begin{array}{ll} \text{minimize} & \varphi(u) \\ \text{subject to} & \\ & u \in \Pi, \end{array} \tag{2.1}$$

where $\varphi[\mathbb{R}^p \to \mathbb{R}]$ is locally Lipschitz and Π is a nonempty and closed subset of $\mathbb{R}^p$. For such programs, first-order necessary optimality conditions attain the following form; cf. [11, Theorem 7.1].

PROPOSITION 2.1. *Let $\hat{u}$ be a (local) solution of* (2.1). *Then one has*

$$0 \in \partial^- \varphi(\hat{u}) + N_\Pi(\hat{u}). \tag{2.2}$$

Assume now that

$$\Pi = \{u \in \Theta \mid -F(u) \in \Lambda\}, \tag{2.3}$$

where $F[\mathbb{R}^p \to \mathbb{R}^q]$ is continuously differentiable and Θ, Λ are closed subsets of $\mathbb{R}^p$ and $\mathbb{R}^q$, respectively. By Φ we denote the multifunction defined by

$$\Phi(u) := F(u) + \Lambda.$$

Our next aim is to derive optimality conditions for problem (2.1) with Π given by (2.3). In this case, unfortunately, the direct computation of $N_\Pi(\hat{u})$ can be quite complicated and so we apply the mentioned penalization approach of [22], [21] and employ the following important concept.

DEFINITION 2.1. *A multifunction $\Psi[\mathbb{R}^q \rightsquigarrow \mathbb{R}^p]$ is said to be* pseudo–upper-Lipschitz *continuous at $(\overline{v}, \overline{u}) \in \operatorname{Gph} \Psi$ with modulus $L \geq 0$, provided there exists a neighborhood $\mathcal{V}$ of $\overline{v}$ and a neighborhood $\mathcal{U}$ of $\overline{u}$ such that*

$$\Psi(v) \cap \mathcal{U} \subset \Psi(\overline{v}) + L\|v - \overline{v}\| \, \mathbb{B} \quad \textit{for all } v \in \mathcal{V}.$$

This concept has been introduced in [7] and entitled pseudo–upper-Lipschitz continuity in [21]. If in the above definition one can put $\mathcal{U} = \mathbb{R}^p$, then Ψ is in fact *(locally) upper-Lipschitz* at $\overline{v}$ with modulus L; cf. [19].

In the further development we make use of the two lemmas given below. The first one comes from [21] and is actually valid without any structural assumptions on Φ.

LEMMA 2.2. *Assume that $\hat{u}$ is a (local) solution of* (2.1) *with Π given by* (2.3). *Further suppose that the multifunction $\Phi^{-1}(\cdot) \cap \Theta$ is pseudo–upper-Lipschitz continuous at $(0, \hat{u})$ with modulus L. Then there exist a neighborhood $\mathcal{V}$ of $0 \in \mathbb{R}^q$ and a neighborhood $\mathcal{U}$ of $\hat{u}$ such that $(v, u) = (0, \hat{u})$ solves the penalized program*

$$\begin{array}{ll} \textit{minimize} & \varphi(u) + r\|v\| \\ \textit{subject to} & \\ & v \in \Phi(u) \cap \mathcal{V}, \\ & u \in \Theta \cap \mathcal{U}, \end{array} \tag{2.4}$$

provided $r \geq L\lambda$, where $\lambda \geq 0$ is the Lipschitz modulus of φ near $\hat{u}$.

The next result is an easy consequence of Theorem A.2 from the appendix.

LEMMA 2.3. *Under the imposed assumptions for each $(u, v) \in \operatorname{Gph} \Phi$ one has*

$$D^*\Phi(u, v)\,(v^*) = \begin{cases} (\nabla F(u))^{\mathrm{T}} v^* & \textit{provided } v^* \in -N_\Lambda(v - F(u)), \\ \emptyset & \textit{otherwise.} \end{cases}$$

Proof. The set Λ can be viewed as the value of a constant multifunction defined on $\mathbb{R}^p$ with the graph $\mathbb{R}^p \times \Lambda$ so that for $\xi \in \Lambda$

$$N_{\operatorname{Gph}\Lambda}(u,\xi) = 0 \times N_\Lambda(\xi);$$

cf. [11, Proposition 1.6]. Hence, for $\xi \in \Lambda$ one has

$$D^*\Lambda(u,\xi)\,(v^*) = \begin{cases} 0 & \text{provided } -v^* \in N_\Lambda(\xi), \\ \emptyset & \text{otherwise,} \end{cases}$$

and the result follows directly from Theorem A.2. ∎

On the basis of the preceding two lemmas we can now state the next result, which is a specialization of [21, Theorem 3.1] to our structure of multifunction Φ.

THEOREM 2.4. *Assume that $\hat{u}$ is a (local) solution of* (2.1) *with* Π *given by* (2.3). *Further suppose that the multifunction $\Phi^{-1}(\cdot) \cap \Theta$ is pseudo–upper-Lipschitz continuous at $(0,\hat{u})$. Then there exists a* Karush–Kuhn–Tucker *(KKT) vector $\hat{v}^* \in N_\Lambda(-F(\hat{u}))$ such that*

$$0 \in \partial^- \varphi(\hat{u}) - (\nabla F(\hat{u}))^{\mathrm{T}} \hat{v}^* + N_\Theta(\hat{u}). \tag{2.5}$$

Proof. By Lemma 2.2 we can apply Proposition 2.1 to program (2.4), provided r is sufficiently large. This yields

$$\begin{bmatrix} 0 \\ 0 \end{bmatrix} \in \begin{bmatrix} \partial^- \varphi(\hat{u}) \\ r\mathbb{B} \end{bmatrix} + N_{\operatorname{Gph}\Phi \cap (\mathcal{U}\times\mathcal{V}) \cap (\Theta \times \mathbb{R}^q)}(\hat{u},0). \tag{2.6}$$

By Definition 1.3 one has

$$N_{\operatorname{Gph}\Phi\cap(\mathcal{U}\times\mathcal{V})}(\hat{u},0) = \{(u^*,w^*) \in \mathbb{R}^p \times \mathbb{R}^q \,|\, u^* \in D^*\Phi(\hat{u},0)\,(-w^*)\}$$

and thus, in virtue of Lemma 2.3,

$$N_{\operatorname{Gph}\Phi\cap(\mathcal{U}\times\mathcal{V})}(\hat{u},0) = \left\{(u^*,w^*) \in \mathbb{R}^p \times \mathbb{R}^q \,|\, u^* = -(\nabla F(\hat{u}))^{\mathrm{T}} w^*,\ w^* \in N_\Lambda(-F(\hat{u}))\right\}.$$

By [11, Proposition 1.6],

$$N_{\Theta\times\mathbb{R}^\ell}(\hat{u},0) = N_\Theta(\hat{u}) \times \{0\}.$$

Hence one has

$$N_{\operatorname{Gph}\Phi\cap(\mathcal{U}\times\mathcal{V})}(\hat{u},0) \cap (-N_{\Theta\times\mathbb{R}^q}(\hat{u},0)) = \{0\}. \tag{2.7}$$

By invoking Theorem A.3, relation (2.6) thus implies that

$$\begin{bmatrix} 0 \\ 0 \end{bmatrix} \in \begin{bmatrix} \partial^- \varphi(\hat{u}) \\ r\mathbb{B} \end{bmatrix} + \begin{bmatrix} -(\nabla F(\hat{u}))^{\mathrm{T}} \hat{v}^* \\ \hat{v}^* \end{bmatrix} + \begin{bmatrix} N_\Theta(\hat{u}) \\ 0 \end{bmatrix}$$

for some $\hat{v}^* \in N_\Lambda(-F(\hat{u}))$. In this way relation (2.5) has been established. ∎

Remark. Statements of the above type can be found in many works on mathematical programming with various subdifferentials and normal cones and even for nonsmooth maps F [20], [6]. In [21] conditions of this type have been derived for the program

$$\begin{array}{ll} \text{minimize} & \varphi(u) \\ \text{subject to} & \\ & 0 \in \Phi(u), \\ & u \in \Theta \end{array}$$

under no structural assumptions on Φ. Then, however, one needs an additional requirement to ensure the validity of the appropriate counterpart to (2.5).

3. Optimality conditions. We return now to problem

$$\begin{array}{ll} \text{minimize} & f(x,y) \\ \text{subject to} & \\ & 0 \in F_1(x,y) + Q(F_2(x,y)), \\ & (x,y) \in \omega, \end{array} \tag{3.1}$$

and pose the following additional assumptions:

(A1) f is locally Lipschitz on $\mathbb{R}^n \times \mathbb{R}^m$;

(A2) F_1, F_2 are continuously differentiable;

(A3) $\omega \subset \mathbb{R}^n \times \mathbb{R}^m$ is nonempty and closed.

To convert (3.1) to the form (2.1) with Π given by (2.3) it suffices to put $u := (x,y) \in \mathbb{R}^n \times \mathbb{R}^m$, $F(u) := \left[\begin{smallmatrix} -F_2(x,y) \\ F_1(x,y) \end{smallmatrix}\right]$, $\Theta := \omega$, and $\Lambda := \operatorname{Gph} Q$. Indeed, the equilibrium constraint can equivalently be written in the form

$$\left[\begin{array}{c} F_2(x,y) \\ -F_1(x,y) \end{array} \right] \in \operatorname{Gph} Q,$$

and so, consequently, the respective multifunction $\Phi[\mathbb{R}^n \times \mathbb{R}^m \rightsquigarrow \mathbb{R}^\ell \times \mathbb{R}^m]$ is defined by

$$\Phi(x,y) := \left[\begin{array}{c} -F_2(x,y) \\ F_1(x,y) \end{array} \right] + \operatorname{Gph} Q. \tag{3.2}$$

The pseudo–upper-Lipschitz continuity of a multifunction is implied by both the pseudo-Lipschitz as well as the upper-Lipschitz continuity [1], [19]. Let us consider first the case where the multifunction $\Phi^{-1}(\cdot) \cap \omega$ is pseudo-Lipschitz around $(0,\hat{x},\hat{y})$. On the basis of Theorem 2.4 and Theorem A.1 from the appendix we obtain directly the following optimality conditions.

THEOREM 3.1. *Let assumptions* (A1)–(A3) *be fulfilled and let* $(\hat{x},\hat{y})$ *be a (local) solution of* (3.1). *Assume further that the constraint qualification*

(CQ)

$$\left.\begin{array}{l} \left[\begin{array}{cc} (\nabla_x F_2(\hat{x},\hat{y}))^{\mathrm{T}} & -(\nabla_x F_1(\hat{x},\hat{y}))^{\mathrm{T}} \\ (\nabla_y F_2(\hat{x},\hat{y}))^{\mathrm{T}} & -(\nabla_y F_1(\hat{x},\hat{y}))^{\mathrm{T}} \end{array} \right] \left[\begin{array}{c} w \\ z \end{array} \right] \in -N_\omega(\hat{x},\hat{y}) \\ (w,z) \in N_{\operatorname{Gph} Q}(F_2(\hat{x},\hat{y}), -F_1(\hat{x},\hat{y})) \end{array}\right\} \text{ implies } \left\{ \begin{array}{l} w = 0, \\ z = 0 \end{array} \right.$$

holds true. Then there exist a pair $(\xi,\eta) \in \partial^- f(\hat{x},\hat{y})$, *a pair* $(\gamma,\delta) \in N_\omega(\hat{x},\hat{y})$, *and a KKT pair* $(\hat{w},\hat{z}) \in N_{\operatorname{Gph} Q}(F_2(\hat{x},\hat{y}), -F_1(\hat{x},\hat{y}))$ *such that*

$$\begin{aligned} 0 &= \xi + (\nabla_x F_2(\hat{x},\hat{y}))^{\mathrm{T}} \hat{w} - (\nabla_x F_1(\hat{x},\hat{y}))^{\mathrm{T}} \hat{z} + \gamma, \\ 0 &= \eta + (\nabla_y F_2(\hat{x},\hat{y}))^{\mathrm{T}} \hat{w} - (\nabla_y F_1(\hat{x},\hat{y}))^{\mathrm{T}} \hat{z} + \delta. \end{aligned} \tag{3.3}$$

Proof. Let us denote by Ψ the map $\Phi^{-1}(\cdot) \cap \omega$, i. e., for $v = (v_1,v_2) \in \mathbb{R}^\ell \times \mathbb{R}^m$,

$$\Psi(v) := \left\{ (x,y) \in \omega \,\middle|\, \left[\begin{array}{c} v_1 + F_2(x,y) \\ v_2 - F_1(x,y) \end{array} \right] \in \operatorname{Gph} Q \right\}.$$

By [13, Theorem 3.2] we know that Ψ is pseudo-Lipschitz around $(0,\hat{x},\hat{y})$ if and only if $D^*\Psi(0,\hat{x},\hat{y})\,(0) = \{0\}$. Map Ψ has exactly the structure of multifunction $\sum$ analyzed in Theorem A.1 with

$$H(v_1,v_2,x,y) = \left[\begin{array}{c} v_1 + F_2(x,y) \\ v_2 - F_1(x,y) \end{array} \right], \ \Omega = \mathbb{R}^\ell \times \mathbb{R}^m \times \omega \text{ and } \Lambda = \operatorname{Gph} Q.$$

Due to the specific structure of the Jacobian $\nabla H(0,0,\hat{x},\hat{y})$ one easily verifies that both qualification conditions of Theorem A.1 are fulfilled. Hence,

$$D^*\Psi(0,0,\hat{x},\hat{y})\,(x^*,y^*) \subset \left\{ \left(v_1^*, v_2^*\right) \in \mathbb{R}^\ell \times \mathbb{R}^m \Big| \right.$$

$$\left[\begin{array}{c} v_1^* \\ v_2^* \\ -x^* \\ -y^* \end{array} \right] = \left[\begin{array}{cc} E & 0 \\ 0 & E \\ (\nabla_x F_2(\hat{x},\hat{y}))^{\mathrm{T}} & -(\nabla_x F_1(\hat{x},\hat{y}))^{\mathrm{T}} \\ (\nabla_y F_2(\hat{x},\hat{y}))^{\mathrm{T}} & -(\nabla_y F_1(\hat{x},\hat{y}))^{\mathrm{T}} \end{array} \right] \left[\begin{array}{c} w \\ z \end{array} \right] + \{0\} \times \{0\} \times N_\omega(\hat{x},\hat{y}),$$

$$\left. (w,z) \in N_{\mathrm{Gph}\,Q}(F_2(\hat{x},\hat{y}), -F_1(\hat{x},\hat{y})) \right\}$$

and

$$D^*\Psi(0,0,\hat{x},\hat{y})\,(0,0) = \left\{ (w,z) \in N_{\mathrm{Gph}\,Q}(F_2(\hat{x},\hat{y}), -F_1(\hat{x},\hat{y})) \right|$$

$$\left. \left[\begin{array}{cc} (\nabla_x F_2(\hat{x},\hat{y}))^{\mathrm{T}} & -(\nabla_x F_1(\hat{x},\hat{y}))^{\mathrm{T}} \\ (\nabla_y F_2(\hat{x},\hat{y}))^{\mathrm{T}} & -(\nabla_y F_1(\hat{x},\hat{y}))^{\mathrm{T}} \end{array} \right] \left[\begin{array}{c} w \\ z \end{array} \right] \in -N_\omega(\hat{x},\hat{y}) \right\}.$$

We observe that (CQ) ensures the pseudo-Lipschitz continuity of $\Phi^{-1}(\cdot) \cap \omega$ at $(0,\hat{x},\hat{y})$. The rest follows readily from Theorem 2.4. □

Observe that, expectantly, x and y enter (3.3) in a fully symmetric way. Only a specific structure of F_1, F_2 causes possibly different roles of the control and the state variable known from standard MPECs.

Condition (CQ) amounts to the standard Mangasarian–Fromowitz constraint qualification for the constraint set

$$\left\{ (x,y) \in \omega \,\middle|\, \left[\begin{array}{c} F_2(x,y) \\ -F_1(x,y) \end{array} \right] \in \mathrm{Gph}\, Q \right\}$$

at $(\hat{x},\hat{y})$ in the so-called dual form. It prevents the existence of (abnormal) nonzero multipliers $(w,z) \in N_{\mathrm{Gph}\,Q}(F_2(\hat{x},\hat{y}), -F_1(\hat{x},\hat{y}))$ such that

$$0 \in \left[\begin{array}{cc} (\nabla_x F_2(\hat{x},\hat{y}))^{\mathrm{T}} & -(\nabla_x F_1(\hat{x},\hat{y}))^{\mathrm{T}} \\ (\nabla_y F_2(\hat{x},\hat{y}))^{\mathrm{T}} & -(\nabla_y F_1(\hat{x},\hat{y}))^{\mathrm{T}} \end{array} \right] \left[\begin{array}{c} w \\ z \end{array} \right] + N_\omega(\hat{x},\hat{y}).$$

Using a particular structure of Q and ω it is sometimes possible to derive also a primal version of (CQ); cf. [15].

The statement of Theorem 3.1 reduces to [21, Theorem 3.2 (c,d)] provided $F_2(x,y) = y$, $Q(\cdot) = N_\Xi(\cdot)$ with a closed convex set $\Xi \subset \mathbb{R}^m$, and the state variable is not subject to any constraints. To verify (CQ) and to be able to apply the derived optimality conditions, we need, however, some structural assumptions on Q (and ω). In [15] the constraint qualification as well as the optimality conditions have been converted to a workable form in the case of equilibria governed by nonlinear complementarity problems and for ω given by smooth inequalities. We return to this subject in the next section, where we analyze two special classes of (3.1).

Let $\tilde{\omega}$ be a closed subset of $\mathbb{R}^n$. For a problem of the type

$$\begin{array}{ll} \text{minimize} & f(x,y) \\ \text{subject to} & \\ & 0 \in F_1(x,y) + Q(y), \\ & x \in \tilde{\omega}, \end{array} \tag{3.4}$$

there is a possibility to ensure the validity of (CQ) via the strong regularity (cf. [18]) of the GE

$$0 \in F_1(\hat{x},y) + Q(y) \tag{3.5}$$

at $\hat{y}$.

PROPOSITION 3.2. *Let* (A1)–(A3) *(with* $\omega = \tilde{\omega} \times \mathbb{R}^m$*) be fulfilled and let* $(\hat{x},\hat{y})$ *be a (local) solution of* (3.4)*. Further assume that the GE* (3.5) *is strongly regular at* $\hat{y}$*, i.e., that the map*

$$\Delta : \eta \mapsto \{y \in \mathbb{R}^m \,|\, \eta \in F_1(\hat{x},\hat{y}) + \nabla_y F_1(\hat{x},\hat{y})\,(y-\hat{y}) + Q(y)\} \tag{3.6}$$

is locally single-valued and Lipschitz around $(0,\hat{y})$*. Then* (CQ) *is fulfilled.*

Proof. By the assumptions, Δ is pseudo-Lipschitz around $(0,\hat{y})$. Hence, again by invoking [13, Theorem 3.2], we have

$$D^*\Delta(0,\hat{y})\,(0) = \{0\}. \tag{3.7}$$

Since (cf. [13])

$$y^* \in D^*\Delta^{-1}(\hat{y},0)\,(\eta^*) \Leftrightarrow -\eta^* \in D^*\Delta(0,\hat{y})\,(-y^*),$$

relation (3.7) amounts to

$$\operatorname{Ker} D^*\Delta^{-1}(\hat{y},0) = \{0\}. \tag{3.8}$$

By Theorem A.3

$$D^*\Delta^{-1}(\hat{y},0)\,(\eta^*) = (\nabla_y F_1(\hat{x},\hat{y}))^{\mathrm{T}}\eta^* + D^*Q(\hat{y},-F_1(\hat{x},\hat{y}))\,(\eta^*).$$

Putting $z := -\eta^*$, condition (3.8) can thus be written in the form

$$\left.\begin{array}{l} w - (\nabla_y F_1(\hat{x},\hat{y}))^{\mathrm{T}} z = 0 \\ (w,z) \in N_{\operatorname{Gph} Q}(\hat{y},-F_1(\hat{x},\hat{y})) \end{array}\right\} \text{ implies } z = 0. \tag{3.9}$$

Since $z = 0$ implies $w = 0$ and

$$N_\omega(\hat{x},\hat{y}) = N_{\tilde{\omega}}(\hat{x}) \times \{0\},$$

condition (3.9) ensures the satisfaction of (CQ). □

The pseudo–upper-Lipschitz continuity of $\Phi^{-1}(\cdot) \cap \omega$ is also implied by the (locally) upper-Lipschitz continuity of this map. To simplify the corresponding statement, we will assume the following.

(A4) With a given $[s \times n]$ matrix C, a given $[s \times m]$ matrix D, and a given vector $c \in \mathbb{R}^s$,

$$\omega := \{(x,y) \in \mathbb{R}^n \times \mathbb{R}^m \,|\, Cx + Dy + c \le 0\}. \tag{3.10}$$

For a fixed $(\overline{x}, \overline{y}) \in \omega$ let

$$M(\overline{x}, \overline{y}) := \left\{ i \in \{1, 2, \ldots, s\} \,|\, (C\overline{x})^i + (D\overline{y})^i + c^i = 0 \right\}.$$

Consider now the optimization problem

$$\begin{array}{ll} \text{minimize} & f(x,y) \\ \text{subject to} & \\ & 0 \in A_1 x + B_1 y + a_1 + Q(A_2 x + B_2 y + a_2), \\ & (x,y) \in \omega, \end{array} \tag{3.11}$$

where A_1, A_2, B_1, B_2, a_1, and a_2 are given matrices and vectors of appropriate dimensions, and ω is given by (3.10).

THEOREM 3.3. *Let $(\hat{x}, \hat{y})$ be a (local) solution of problem* (3.11). *Suppose that Q is polyhedral (i.e.,* $\operatorname{Gph} Q$ *is a union of finitely many convex polyhedral sets), and* (A1), (A4) *are fulfilled. Then there exist a pair $(\xi, \eta) \in \partial^- f(\hat{x}, \hat{y})$, a KKT pair $(\hat{w}, \hat{z}) \in N_{\operatorname{Gph} Q}(A_2\hat{x} + B_2\hat{y} + a_2,\ -A_1\hat{x} - B_1\hat{y} - a_1)$, and a multiplier $\hat{\mu} \in \mathbb{R}^s_+$ such that*

$$\begin{aligned} 0 &= \xi + A_2^{\mathrm{T}}\hat{w} - A_1^{\mathrm{T}}\hat{z} + C^{\mathrm{T}}\hat{\mu}, \\ 0 &= \eta + B_2^{\mathrm{T}}\hat{w} - B_1^{\mathrm{T}}\hat{z} + D^{\mathrm{T}}\hat{\mu}, \\ \hat{\mu}^i &= 0 \quad \textit{for all } i \notin M(\hat{x}, \hat{y}). \end{aligned} \tag{3.12}$$

Proof. By the imposed assumptions, $\operatorname{Gph}\Phi$ is a union of finitely many convex polyhedral sets and ω is convex polyhedral. This implies in virtue of [19] that $\Phi^{-1}(\cdot) \cap \omega$ is (locally) upper-Lipschitz at 0, and hence the conditions of Theorem 2.4 are fulfilled. It remains to express relation (3.3) in terms of our problem data. By the convexity and polyhedrality of ω one has

$$N_\omega(\hat{x}, \hat{y}) = \left\{ \begin{bmatrix} C^{\mathrm{T}} \\ D^{\mathrm{T}} \end{bmatrix} \mu \,\middle|\, \mu \in \mathbb{R}^s_+,\ \mu^i = 0 \quad \text{for } i \notin M(\hat{x}, \hat{y}) \right\}. \tag{3.13}$$

Hence, it suffices to insert (3.13) into (3.3) and take into account the specific structure of functions F_1, F_2. ∎

We conclude this section with a simple academic example illustrating the optimality conditions of Theorem 3.3. Consider the MPEC

$$\begin{array}{ll} \text{minimize} & \dfrac{1}{2}x - y \\ \text{subject to} & \\ & 0 \in y - x + \partial|y|, \\ & x \in [-2, 0]. \end{array} \tag{3.14}$$

In this one-dimensional MPEC it is easy to construct the map S assigning x those y's which are feasible with respect to the equilibrium constraint; this map is depicted in Figure 1. Hence the unique global solution of (3.14) is the point $(\hat{x}, \hat{y}) = (-1, 0)$. One has $F_1(x,y) = y - x$, $F_2(x,y) = y$ and $Q(\cdot) = \partial|\cdot|$. Since $\operatorname{Gph} Q$ is polyhedral (cf. Figure 2), all assumptions of Theorem 3.3 are fulfilled. On the basis of Definition 1.1 and Figure 2 one easily deduces that

$$N_{\operatorname{Gph}\partial|\cdot|}(\hat{y}, \hat{x} - \hat{y}) = N_{\operatorname{Gph}\partial|\cdot|}(0, -1) = \{(w,z) \,|\, \text{either } wz = 0 \text{ or } w > 0 \text{ and } z < 0\}.$$

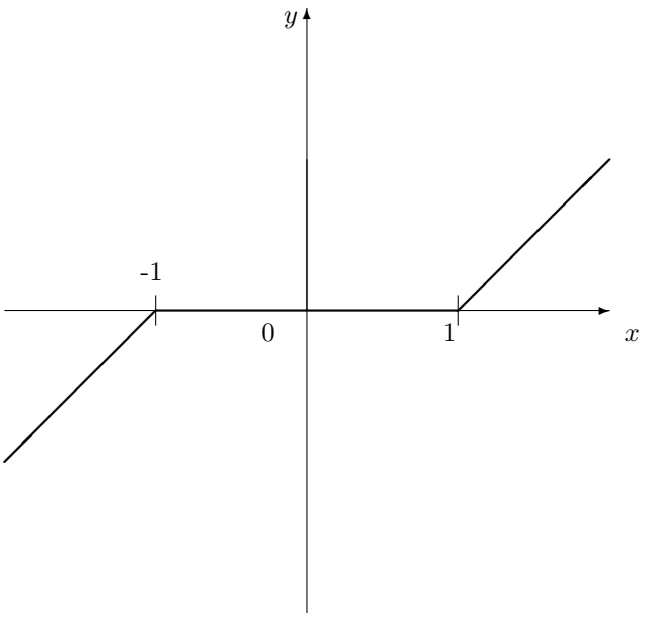


FIG. 1. Gph S.

Conditions (3.12) attain the form

$$0 = \frac{1}{2} + z,$$
$$0 = -1 + w - z,$$

and we observe that they are fulfilled with

$$(\hat{w}, \hat{z}) = \left(\frac{1}{2}, -\frac{1}{2}\right) \in N_{\operatorname{Gph} \partial|\cdot|}(0, -1).$$

The next section is devoted to two important classes of MPECs, where the optimality conditions of Theorems 3.1 and 3.3 can be converted to a workable form.

4. Applications. Assume that we are given two continuously differentiable functions F, $G[\mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}^m]$ and consider equlibria governed by the parameter-dependent *implicit complementarity problem* (ICP):

for a given $x \in \mathbb{R}^n$, find $y \in \mathbb{R}^m$ such that

$$F(x,y) \geq 0,\ y \geq G(x,y) \quad \text{and} \quad \langle y - G(x,y), F(x,y)\rangle = 0.$$

If $G \equiv 0$, the ICP reduces to the standard nonlinear complementarity problem. As an ICP one can model, e.g., (discretized) obstacle problems with compliant obstacles [8] or filtration through porous media [14]. In [16] optimality conditions have been derived for MPECs with such equilibria under the strong regularity assumption. Here we remove this assumption and strengthen the conditions from [16] using the approach of the preceding section.

It is well known that an ICP can equivalently be formulated as the GE of the type (1.3):

$$0 \in F(x,y) + N_{\mathbb{R}^m_+}(y - G(x,y)). \tag{4.1}$$

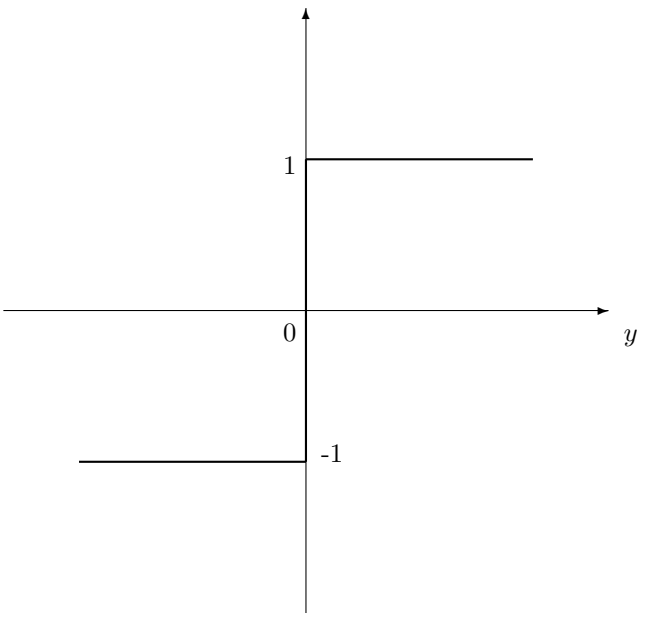


FIG. 2. Gph $\partial|y|$.

Analogously to [15] we associate with a pair (x, y), feasible with respect to (4.1), the index sets

$$\begin{aligned} L(x,y) &:= \left\{ i \in \{1,2,\ldots,m\} \,|\, y^i > G^i(x,y) \right\}, \\ I(x,y) &:= \left\{ i \in \{1,2,\ldots,m\} \,|\, y^i = G^i(x,y) \right\}, \\ I_+(x,y) &:= \left\{ i \in I(x,y) \,|\, F^i(x,y) > 0 \right\}, \\ I_0(x,y) &:= \left\{ i \in I(x,y) \,|\, F^i(x,y) = 0 \right\}. \end{aligned}$$

Evidently, $L(x,y) \cup I_+(x,y) \cup I_0(x,y) = \{1,2,\ldots,m\}$. Since we will associate these index sets only with a (locally) optimal pair $(\hat{x}, \hat{y})$, their arguments are dropped whenever they occur as subscripts.

On the basis of Theorem 3.1 we can now state the optimality conditions for the problem

$$\begin{array}{ll} \text{minimize} & f(x,y) \\ \text{subject to} & \\ & 0 \in F(x,y) + N_{\mathbb{R}^m_+}(y - G(x,y)), \\ & (x,y) \in \omega. \end{array} \tag{4.2}$$

THEOREM 4.1. *Let $(\hat{x}, \hat{y})$ be a (local) solution of problem* (4.2) *and let the assumptions* (A1), (A3) *be fulfilled. Further suppose that the constraint qualification*

(CQ*)

$$\left.\begin{array}{l} \left[\begin{array}{cc} -(\nabla_x G_I(\hat{x},\hat{y}))^{\mathrm{T}} & -(\nabla_x F_{L\cup I_0}(\hat{x},\hat{y}))^{\mathrm{T}} \\ (E_I)^{\mathrm{T}} - (\nabla_y G_I(\hat{x},\hat{y}))^{\mathrm{T}} & -(\nabla_y F_{L\cup I_0}(\hat{x},\hat{y}))^{\mathrm{T}} \end{array}\right] \left[\begin{array}{c} w_I \\ z_{L\cup I_0} \end{array}\right] \in -N_\omega(\hat{x},\hat{y}) \\ \text{for } i \in I_0(\hat{x},\hat{y}), \text{ either } w^i z^i = 0 \text{ or } w^i < 0 \text{ and } z^i > 0 \end{array}\right\} \text{implies } w_I = 0, z_{L\cup I_0} = 0$$

holds true. Then there exist a pair $(\xi,\eta)\in\partial^- f(\hat x,\hat y)$*, a pair* $(\gamma,\delta)\in N_\omega(\hat x,\hat y)$*, and a KKT pair* $(\hat w,\hat z)\in\mathbb{R}^m\times\mathbb{R}^m$ *such that* $\hat w_L=0,\hat z_{I_+}=0$*,*

$$\begin{aligned}0 &= \xi-(\nabla_x G_I(\hat x,\hat y))^{\mathrm T}\hat w_I-(\nabla_x F_{L\cup I_0}(\hat x,\hat y))^{\mathrm T}\hat z_{L\cup I_0}+\gamma,\\ 0 &= \eta+\hat w-(\nabla_y G_I(\hat x,\hat y))^{\mathrm T}\hat w_I-(\nabla_y F_{L\cup I_0}(\hat x,\hat y))^{\mathrm T}\hat z_{L\cup I_0}+\delta,\end{aligned}\tag{4.3}$$

and for $i\in I_0(\hat x,\hat y)$*, either* $\hat w^i\hat z^i=0$ *or* $\hat w^i<0$ *and* $\hat z^i>0$*.*

Proof. Evidently, the GE (4.1) is a special case of the GE (1.3) with

$$F_1(x,y)=F(x,y),\quad F_2(x,y)=y-G(x,y),\quad\text{and}\quad Q(\cdot)=N_{\mathbb{R}^m_+}(\cdot).$$

In virtue of [15, Lemma 2.2] we readily infer that

$$\begin{aligned}N_{\operatorname{Gph}N_{\mathbb{R}^m_+}}(\hat y-G(\hat x,\hat y),-F(\hat x,\hat y)) = \{(w,z)\in\mathbb{R}^m\times\mathbb{R}^m \mid w_L=0,\ z_{I_+}=0\\ \text{and for } i\in I_0(\hat x,\hat y) \text{ either } w^iz^i=0 \text{ or } w^i<0 \text{ and } z^i>0\}.\end{aligned}\tag{4.4}$$

Consequently, it suffices to apply Theorem 3.1 and realize that (CQ) becomes (CQ*) and conditions (3.3) attain the form (4.3). □

Remark. If $G\equiv 0$ and the state variable is not subject to any constraints, Theorem 4.1 reduces to the optimality conditions derived in [15]. Indeed, since $\hat w_{I_+}$ is arbitrary and the third term at the right-hand side of the second equation in (4.3) disappears, we can neglect all equations

$$0=\eta^i+\hat w^i-\big((\nabla_y F_{L\cup I_0}(\hat x,\hat y))^{\mathrm T}z_{L\cup I_0}\big)^i+\delta^i$$

for $i\in I_+(\hat x,\hat y)$. In the case of genuine ICPs this is, however, not possible.

For the mathematical description of the behavior of some modern materials like fiberglass or sandwich composites a theory has been developed in [17], leading to so-called hemivariational inequalities. After a suitable discretization, some of them attain the form of the GE

$$0\in A(x)\,y+b(x)+Q(y),\tag{4.5}$$

where the maps $A[\mathbb{R}^n\to\mathbb{R}^m\times\mathbb{R}^m]$ and $b[\mathbb{R}^n\to\mathbb{R}^m]$ are continuously differentiable and

$$Q(y)=\mathop{\mathrm{X}}_{i=1}^m Q_i(y^i),\quad Q_i(y^i)=\begin{cases}\partial^c J(y^i) & \text{for}\quad i\in I_1,\\ 0 & \text{for}\quad i\in\{1,2,\dots,m\}\setminus I_1\end{cases}\tag{4.6}$$

with a locally Lipschitz function $J[\mathbb{R}\to\mathbb{R}]$ and a given index set $I_1\subset\{1,2,\dots,m\}$. In (4.5) y corresponds, e.g., to displacement or stresses at single nodes of discretization and x describes, e.g., the shape of the body in question; cf. [9]. Since J is a function of one variable, in most cases the set $\operatorname{Gph}\partial^c J\subset\mathbb{R}^2$ can easily be constructed. Moreover, if J is a piecewise smooth function, then also the generalized normal cone can readily be computed at each point of $\operatorname{Gph}\partial^c J$.

Consider now the problem

$$\begin{aligned}&\text{minimize} && f(x,y)\\ &\text{subject to}\\ &&& (x,y) \text{ is feasible with respect to the GE (4.5)},\\ &&& (x,y)\in\omega.\end{aligned}\tag{4.7}$$

THEOREM 4.2. *Let $(\hat{x}, \hat{y})$ be a (local) solution of problem* (4.7) *and let the assumptions* (A1), (A3) *be fulfilled. Further suppose that the constraint qualification*

$$\left.\begin{array}{l}(\mathrm{CQ}^+)\\ \left[\begin{array}{cc} 0 & -(\nabla_x(A(\hat{x})\,\hat{y}+b(\hat{x})))^{\mathrm{T}} \\ (E_{I_1})^{\mathrm{T}} & -(A(\hat{x}))^{\mathrm{T}} \end{array}\right]\left[\begin{array}{c} w_{I_1} \\ z \end{array}\right] \in -N_\omega(\hat{x},\hat{y})\\ \text{for } i\in I_1 \text{ one has } (w^i,z^i)\in N_{\mathrm{Gph}\,\partial^c J}(\hat{y}^i,-(A(\hat{x})\,\hat{y}+b(\hat{x}))^i)\end{array}\right\} \text{implies} \left\{\begin{array}{l} w_{I_1}=0,\\ z=0\end{array}\right.$$

holds true. Then there exist a pair $(\xi,\eta)\in\partial^- f(\hat{x},\hat{y})$, a pair $(\gamma,\delta)\in N_\omega(\hat{x},\hat{y})$, and a KKT pair $(\hat{w},\hat{z})\in\mathbb{R}^m\times\mathbb{R}^m$ such that

$$\begin{aligned} \hat{w}^i &= 0 \quad \text{for } i\notin I_1,\\ 0 &= \xi-(\nabla_x(A(\hat{x})\,\hat{y}+b(\hat{x})))^{\mathrm{T}}\hat{z}+\gamma,\\ 0 &= \eta+\hat{w}-(A(\hat{x}))^{\mathrm{T}}\hat{z}+\delta, \end{aligned} \tag{4.8}$$

and for $i\in I_1$ one has $(\hat{w}^i,\hat{z}^i)\in N_{\mathrm{Gph}\,\partial^c J}(\hat{y}^i,-(A(\hat{x})\,\hat{y}+b(\hat{x}))^i)$.

Proof. It suffices to apply Theorem 3.1 with $F_1(x,y):=A(x)\,y+b(x)$, $F_2(x,y):=y$, and Q given by (4.6). If $i\notin I_1$, then clearly

$$N_{\mathrm{Gph}\,Q_i}(\hat{y}^i,0)=\left\{(w^i,z^i)\in\mathbb{R}\times\mathbb{R}\,|\,w^i=0\right\}. \qquad \square$$

We illustrate now the application of the above statement by a simple academic example. Consider the MPEC

$$\begin{array}{ll} \text{minimize} & -2xy-x\\ \text{subject to} & \\ & 0\in 2xy+x+\partial^c J(y),\\ & x\in[-2,2], \end{array} \tag{4.9}$$

where

$$J(y):=\begin{cases} 0 & \text{for} \quad |y|\geq 1,\\ \frac{1}{2}y^2+y+\frac{1}{2} & \text{for} \quad y\in(-1,0],\\ \frac{1}{2}y^2-y+\frac{1}{2} & \text{for} \quad y\in(0,1). \end{cases}$$

The set $\mathrm{Gph}\,\partial^c J$ is depicted in Figure 3. Due to the relation between our hemivariational inequality and the objective in (4.9), we deduce that $(\hat{x},\hat{y})=(1,0)$ is the (global) solution of (4.9). Condition $(\mathrm{CQ})^+$ is fulfilled because $N_\omega(\hat{x},\hat{y})=(0,0)$ and

$$\left[\begin{array}{cc} 0 & -1\\ 1 & -2 \end{array}\right]\left[\begin{array}{c} w\\ z \end{array}\right]=\left[\begin{array}{c} 0\\ 0 \end{array}\right] \text{ implies } w=z=0.$$

System (4.8) attains the form

$$\begin{aligned} 0&=-1-z,\\ 0&=-2+w-2z, \end{aligned}$$

and possesses the unique solution $(\hat{w},\hat{z})=(0,-1)$. Since $(0,-1)$ evidently belongs to $N_{\mathrm{Gph}\,\partial^c J}(0,-1)$, the optimality conditions of Theorem 4.2 are satisfied.

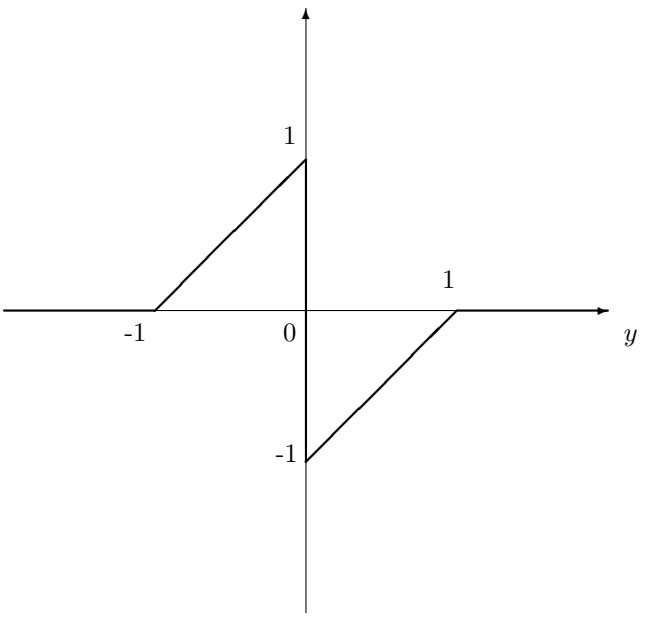


FIG. 3. $\operatorname{Gph} \partial^c J$.

5. Concluding remarks. The derived optimality conditions can well be used in testing of optimality (stationarity) of approximate solutions to problems of the type (1.5) computed by various available numerical methods.

The computation of the generalized normal cone to the graph of Q is substantially facilitated provided Q is polyhedral. Then, for a pair $(u, v) \in \operatorname{Gph} Q$, one has

$$N_{\operatorname{Gph} Q}(u, v) = \left\{ \bigcup T^0_{\operatorname{Gph} Q}(u', v') \,|\, (u', v') \in \mathcal{O} \cap \operatorname{Gph} Q \right\},$$

where $\mathcal{O}$ is any sufficiently small neighborhood of (u, v); cf. [2]. The situation is particularly simple provided

$$Q(y) = \bigtimes_{i=1}^{m} Q_i(y^i),$$

as in the case of both equilibria considered in the previous section. A different discretization of hemivariational inequalities leads, however, to a different situation, where

$$Q(y) = \bigtimes_{j=1}^{m} Q_j(y_j)$$

with the subvectors y_j of y having a "low" dimension. Nevertheless, even in such a situation the computation of $N_{\operatorname{Gph} Q}$ is still realistic.

In this paper the desired pseudo–upper-Lipschitz continuity of $\Phi^{-1}(\cdot) \cap \omega$ (with Φ given by (3.2)) is in the nonpolyhedral case ensured via condition (CQ) implying in fact the (more restrictive) pseudo-Lipschitz continuity. It seems that under some additional assumptions on Q and ω it is possible to guarantee this property by a weaker condition than (CQ) [4]. This issue is, however, rather complicated and deserves a separate study.

The last remark concerns the numerical solution of (1.5). From the above theory it is clear that under the assumptions of Theorem 3.1 or Theorem 3.3 the function

$$P : (x, y) \longmapsto \operatorname{dist}_{\operatorname{Gph} Q} \left(\begin{bmatrix} F_2(x, y) \\ -F_1(x, y) \end{bmatrix} \right) \tag{5.1}$$

is a Lipschitzian error bound of $\Phi^{-1}(0) \cap \omega$ with respect to test vectors from $\omega \cap \tilde{\mathcal{O}}$, where $\tilde{\mathcal{O}}$ is a neighborhood of $(\hat{x}, \hat{y})$. If we succeed to assign the graph of Q a suitable metric in $\mathbb{R}^\ell \times \mathbb{R}^m$ for which the values of P can easily be computed, we could try to solve (1.5) via the penalized problem

$$\begin{array}{ll} \text{minimize} & f(x, y) + \rho P(x, y) \\ \text{subject to} & \\ & (x, y) \in \omega \end{array} \tag{5.2}$$

with a suitably chosen penalty parameter $\rho \geq 0$.

Appendix. The generalized normal cones and the coderivatives of closed-graph multifunctions admit a rich calculus which is based mainly on the so-called extremal principle; cf. [11], [12]. Next we give three statements of this calculus which have been applied in sections 2 and 3. The first one concerns the multifunction

$$\Sigma(v) := \{u \in \mathbb{R}^p \mid H(v, u) \in \Lambda,\ (v, u) \in \Omega\}, \tag{A.1}$$

where $v \in \mathbb{R}^q$, H maps $\mathbb{R}^q \times \mathbb{R}^p$ into $\mathbb{R}^s$, $\Lambda \subset \mathbb{R}^s$, and $\Omega \subset \mathbb{R}^q \times \mathbb{R}^p$.

THEOREM A.1 (see [12, Theorem 6.10]). *Let Λ and Ω be closed and let H be continuously differentiable. For a pair $(v, u) \in \mathbb{R}^q \times \mathbb{R}^p$ assume that $(v, u) \in \Omega$, $H(v, u) \in \Lambda$, and the qualification conditions*

$$N_\Lambda(H(v, u)) \cap \operatorname{Ker}(\nabla H(v, u))^{\mathrm{T}} = \{0\},$$
$$(\nabla H(v, u))^{\mathrm{T}} N_\Lambda(H(v, u)) \cap (-N_\Omega(v, u)) = \{0\}$$

are fulfilled. Then for all $u^* \in \mathbb{R}^p$ one has the inclusion

$$D^*\Sigma(v, u)(u^*) \subset \left\{v^* \in \mathbb{R}^q \mid (v^*, -u^*) \in (\nabla H(v, u))^{\mathrm{T}} N_\Lambda(H(v, u)) + N_\Omega(v, u)\right\}. \tag{A.2}$$

The second statement concerns the sum of a function $\Phi_1[\mathbb{R}^p \to \mathbb{R}^\ell]$ and a multifunction $\Phi_2[\mathbb{R}^p \rightsquigarrow \mathbb{R}^\ell]$.

THEOREM A.2 (see [12, Corollary 4.4]). *Let Φ_1 be continuously differentiable on a neighborhood of $u \in \mathbb{R}^p$ and let Φ_2 have the closed graph. Then for each $v \in \Phi_1(u) + \Phi_2(u)$ and $v^* \in \mathbb{R}^\ell$ one has*

$$D^*(\Phi_1 + \Phi_2)(u, v)(v^*) = (\nabla \Phi_1(u))^{\mathrm{T}} v^* + D^*\Phi_2(u, v - \Phi_1(u))(v^*). \tag{A.3}$$

In some situations we would like to compute the generalized normal cone to an intersection of sets. The respective fundamental result from [11] can be, in the case of only two sets, simplified to the following form.

THEOREM A.3. *Consider two sets $\Omega_1, \Omega_2 \subset \mathbb{R}^s$ and a point $z \in cl\Omega_1 \cap cl\Omega_2$. Assume that*

$$N_{\Omega_1}(z) \cap (-N_{\Omega_2}(z)) = \{0\}.$$

Then one has the inclusion

$$N_{\Omega_1 \cap \Omega_2}(z) \subset N_{\Omega_1}(z) + N_{\Omega_2}(z).$$

Acknowledgments. The author would like to express his gratitude to B. Mordukhovich and D. Ralph for many helpful discussions on the subject of the paper. He is also deeply indebted to both referees for a number of constructive and helpful suggestions.

REFERENCES

[1] J.-P. Aubin and H. Frankowska, *Set-Valued Analysis*, Birkhäuser, Boston, 1990.
[2] A. L. Dontchev and R. T. Rockafellar, *Characterizations of strong regularity for variational inequalities over polyhedral convex sets*, SIAM J. Optim., 6 (1996), pp. 1087–1105.
[3] R. Glowinski, *Numerical Methods for Nonlinear Variational Problems*, Springer-Verlag, Heidelberg, 1984.
[4] R. Henrion and J. V. Outrata, *A Subdifferential Criterion for Calmness of Multifunctions*, Preprint 553, Weierstrass Inst. for Applied Analysis and Stochastics, Berlin, 2000.
[5] I. Hlaváček, J. Haslinger, J. Nečas, and J. Lovíšek, *Variational Inequalities in Mechanics*, Springer-Verlag, New York, 1988.
[6] A. Jourani and L. Thibault, *Approximations and metric regularity in mathematical programming in Banach space*, Math. Oper. Res., 18 (1993), pp. 390–401.
[7] D. Klatte, *On quantitative stability for non-isolated minima*, Control Cybernet., 23 (1994), pp. 183–200.
[8] M. Kočvara and J. V. Outrata, *On optimization of systems governed by implicit complementarity problems*, Numer. Funct. Anal. Optim., 15 (1994), pp. 869–887.
[9] M. Miettinen and J. Haslinger, *Approximation of nonmonotone multivalued differential inclusions*, IMA J. Numer. Anal., 15 (1995), pp. 475–503.
[10] Z.-Q. Luo, J.-S. Pang, and D. Ralph, *Mathematical Programs with Equilibrium Constraints*, Cambridge University Press, Cambridge, UK, 1996.
[11] B. S. Mordukhovich, *Approximation Methods in Problems of Optimization and Control*, Nauka, Moscow, 1988 (in Russian; Wiley-Interscience English translation to appear).
[12] B. S. Mordukhovich, *Generalized differential calculus for nonsmooth and set-valued mappings*, J. Math. Anal. Appl., 183 (1994), pp. 250–288.
[13] B. S. Mordukhovich, *Lipschitzian stability of constraint systems and generalized equations*, Nonlinear Anal., 22 (1994), pp. 173–206.
[14] V. Mosco, *Implicit variational problems and quasi-variational inequalities*, in Nonlinear Operators and the Calculus of Variations, Lecture Notes in Math. 543, Springer-Verlag, Berlin, 1976, pp. 83–156.
[15] J. V. Outrata, *Optimality conditions for a class of mathematical programs with equilibrium constraints*, Math. Oper. Res., 24 (1999), pp. 627–644.
[16] J. V. Outrata, M. Kočvara, and J. Zowe, *Nonsmooth Approach to Optimization Problems with Equilibrium Constraints*, Kluwer Academic Publishers, Dordrecht, The Netherlands, 1998.
[17] P. D. Panagiotopoulos, *Hemivariational Inequalities. Applications in Mechanics and Engineering*, Springer-Verlag, New York, 1993.
[18] S. M. Robinson, *Strongly regular generalized equations*, Math. Oper. Res., 5 (1980), pp. 43–62.
[19] S. M. Robinson, *Some continuity properties of polyhedral multifunctions*, Math. Programming Stud., 14 (1981), pp. 206–214.
[20] R. T. Rockafellar, *Extensions of subgradient calculus with applications to optimization*, Nonlinear Anal., 9 (1985), pp. 665–698.
[21] J. J. Ye and X. Y. Ye, *Necessary optimality conditions for optimization problems with variational inequality constraints*, Math. Oper. Res., 22 (1997), pp. 977–997.
[22] R. Zhang, *Problems of hierarchical optimization in finite dimensions*, SIAM J. Optim., 4 (1994), pp. 521–536.

SIAM J. CONTROL OPTIM.
Vol. 38, No. 5, pp. 1639–1654

OPTIMAL CONTROL OF PARABOLIC DIFFERENTIAL EQUATIONS WITH TWO POINT BOUNDARY STATE CONSTRAINTS*

GENGSHENG WANG†

Abstract. This paper deals with the necessary conditions of optimality for some optimal control problems governed by some parabolic differential equations involving monotone graphs. The two point boundary (time variable) state constraints are discussed. Some analyses on the state equations are given.

Key words. variational inequality, optimal control, state constraint, maximum principle

AMS subject classifications. 49K20, 35J65

PII. S0363012998338132

1. Introduction.

Let $H = L^2(\Omega)$, where $\Omega \subset R^N$ is a bounded domain with smooth boundary, $Q = \Omega \times (0,T)$.

We shall study the optimal control problems governed by nonlinear parabolic equations of the form

$$y' + Ay + \beta(y) \ni Bu \quad \text{almost everywhere (a.e.) in } Q, \tag{1.1}$$

with the state constraint

$$(y(0), y(T)) \in S \subset H \times H. \tag{1.2}$$

The pay-off functional is given by

$$L(y,u) = \int_0^T [g(t, y(t)) + h(u(t))]dt. \tag{1.3}$$

Note that y' in (1.1) is the strong derivative with respect to t of the function $y : Q \to R$ as a function of t from $[0,T]$ to $L^2(\Omega)$.

For the data in (1.1)–(1.3), we have the following assumptions.

(H_1) Let $V \subset H$ be a real Hilbert space such that V is dense in H and $V \subset H \subset V'$ algebraically and topologically, where V' is the dual of V. Further, the injection of V into H is compact.

$A : V \to V'$ is a linear continuous and symmetric operator from V to V' satisfying the coercivity condition

$$\langle Ay, y\rangle \geq w\|y\|_V^2 - \alpha\|y\|_H^2 \quad \text{for all } y \in V, \tag{1.4}$$

where $w > 0$ and $\alpha \in R$.

(H_2) β is a maximal monotone graph in $R \times R$ with $0 \in D(\beta)$. Moreover, there exists a constant C independent of ε such that

$$\langle Ay, \beta_\varepsilon(y)\rangle \geq -C(1 + \|\beta_\varepsilon(y)\|_H)(1 + \|y\|_H) \tag{1.5}$$

*Received by the editors May 1, 1998; accepted for publication (in revised form) May 28, 1999; published electronically May 26, 2000. This research was partially supported by the HuaCheng Foundation of China.

http://www.siam.org/journals/sicon/38-5/33813.html

†Mathematics Department, Huazhong Normal University, Wuhan 430079, P.R. of China (kyc@ccnu.edu.cn).

for all

$$y \in D(A_H) \equiv \{y \in V; Ay \in H\},$$

where $\beta_\varepsilon(r) = \varepsilon^{-1}(r - (1+\varepsilon\beta)^{-1}r)$ for all $\varepsilon > 0,\ r \in R$.

Let $\phi : H \to \overline{R} = (-\infty, +\infty]$ be the lower semicontinuous convex function defined by

$$\phi(y) = \int_\Omega j(y(x))dx,$$

where $j : R \to \overline{R}$ is such that $\beta = \partial j$.

Then $\partial\phi(y) = \{w \in H : w(x) \in \beta(y(x))$ a.e. $x \in \Omega\}$ (cf. [1] or [6]), where $\partial\phi$ denotes the subdifferential of ϕ.

Suppose

$$\overline{D(\phi)} = H, \tag{1.6}$$

where $D(\phi)$ denotes the effective domain of ϕ.

(H_3) $S \subset D(\phi) \times H \subset H \times H$ is a convex closed subset with finite codimensionality (cf. [3]).

(H_4) B is a linear continuous operator from a real Hilbert space U to H.

(H_5) The functional $h : U \to \overline{R}$ is convex and lower semicontinuous (l.s.c.). Moreover, there exist $c_1 > 0$ and $c_2 \in R$ such that $h(u) \geq c_1\|u\|_U^2 + c_2$ for all $u \in U$.

(H_6) $g : [0,T] \times H \to R^+$ is measurable in t, and for every $r > 0$, there exists $L_r > 0$ independent of t such that $g(t,0) \in L^\infty(0,T)$ and

$$|g(t,y) - g(t,z)| \leq L_r\|y - z\|_H$$

for all $t \in [0,T]$, $\|y\|_H + \|z\|_H \leq r$.

Note that, by (H_2), the state equation (1.1) is equivalent to

$$y'(t) + Ay(t) + \partial\phi(y(t)) \ni Bu(t) \text{ a.e. } t \in (0,T). \tag{1.7}$$

As we know (cf. [1]), for any $u \in L^2(0,T;U), y_0 \in D(\phi) \cap V$, (1.7) with the initial condition

$$y(0) = y_0 \tag{1.8}$$

has a unique solution

$$y \equiv y(t, y_0, u) \in W^{1.2}([0,T];H) \cap C([0,T];H) \cap L^2(0,T;D(A_H)).$$

If $y_0 \in H(= \overline{D(\phi)}$, by (1.6)), then (1.7) with (1.8) has a unique solution in $C([0,T];H)$ with $t^{1/2}y' \in L^2(0,T;H), t^{1/2}y \in L^2(0,T;D(A_H))$.

Now we formulate the optimal control problems as follows.

Let $A_{ad} = \{(y,u) \in W^{1,2}([0,T];H) \cap C([0,T;H) \cap L^2(0,T;D(A_H)) \times L^2(0,T;U) \mid y$ is the solution of (1.7) with (1.8) corresponding to $u, (y(0), y(T)) \in S\}$.

We are asked to find

(P) Min $L(y,u)$ over $(y,u) \in A_{ad}$.

Recently, many mathematicians have discussed the optimal control problems governed by linear and nonlinear parabolic differential equations. Barbu made many contributions in this field (cf. [1]). Li and Yong studied the maximal principle for optimal control governed by some nonlinear parabolic equations with two point boundary (time variable) state constraints (cf. [3], [4]). Pavel also discussed the necessary conditions for optimal control governed by linear parabolic equations with two-point boundary constraints (cf. [6]). However, there are some stronger restrictions on the nonlinear terms on the case studied by Li and Yong so that the state equation and its variational equation can be studied by linear semigroup. In Barbu's works, the two point boundary state constraints were considered less for the nonlinear state equation.

The present work in this paper is concerned with the optimal control problem governed by the state equation which is the same as what Barbu discussed in [1] and much more general than those in Li and Yong's case [3]. However, the problem studied in [1] does not involve state constraints.

Because of the involvement of the state constraint, we construct a new kind of penalty functional for the pay-off function. The idea is based on those in [1]. Also, some analyses for the state equation, like the sensitivity of the state with respect to the control, are discussed.

Due to the involvement of monotone graphs in the state equation, we need some restrictions on the initial data for the state equation, so we assume $S \subset D(\phi) \times H$ instead of $S \subset H \times H$ (Li and Yong's case). Also, the variational equations now involve measure, so the methods used in Li and Yong's case do not work here.

The plan of this paper is as follows. Section 2 gives an approximating control process. In section 3, we state and prove the necessary conditions on optimality for the problem (P). In section 4, some remarks are given.

2. The approximating control process. Let (y^*, u^*) be optimal for the problem (P). Then

$$y^{*\prime}(t) + Ay^*(t) + \beta(y^*(t)) \ni Bu^*(t), \qquad \text{a.e. } t \in (0,T), \qquad (y^*(0), y^*(T)) \in S$$

and

$$L(y^*, u^*) = \operatorname{Inf} L(y,u) \text{ over } (y,u) \in A_{ad}.$$

In order to introduce the approximating control process, we need the approximations β_ε and β^ε for β and related results. For the details we refer readers to [1].

Set $\beta_\varepsilon = \varepsilon^{-1}(1 - (1+\varepsilon\beta)^{-1})$ and

$$\beta^\varepsilon(r) = \int_{-\infty}^{+\infty} [\beta_\varepsilon(r - \varepsilon^2\theta) - \beta_\varepsilon(-\varepsilon^2\theta)]\rho(\theta)d\theta + \beta_\varepsilon(0), \tag{2.1}$$

where ρ is a C_0^∞-mollifier on R, i.e., $\rho \in C_0^\infty(R)$, $\rho(0) = 0$ for $|r| > 1, \rho(r) = \rho(-r)$, and $\int_{-\infty}^{\infty} \rho(r)dr = 1$.

Then β^ε is infinitely differentiable and Lipschitzian with Lipschitz constant ε^{-1}, and

$$\dot\beta^\varepsilon(r) \geq 0 \quad \text{for all } r \in R, \tag{2.2}$$

$$|\beta^\varepsilon(r) - \beta_\varepsilon(r)| \leq 2\varepsilon \quad \text{for all } r \in R. \tag{2.3}$$

Let $\phi^\varepsilon : H \to R$ be given by

$$\phi^\varepsilon(y) = \int_\Omega j^\varepsilon(y(x))dx \quad \text{for all } y \in H, \tag{2.4}$$

where $j^\varepsilon(r) = \int_0^r \beta^\varepsilon(s)ds$ for all $r \in R$.

Then ϕ^ε is Frechet differentiable on H and

$$\nabla\phi^\varepsilon(y(x)) = \beta^\varepsilon(y(x)) \text{ a.e. } x \in \Omega \quad \text{for all } y \in H. \tag{2.5}$$

Let $\phi_\varepsilon : H \to R$ be defined by

$$\phi_\varepsilon(y) = \operatorname{Inf}\left\{ \frac{\|y-z\|_H^2}{2\varepsilon} + \phi(z); \quad z \in H \right\}.$$

Then $\nabla\phi_\varepsilon(y(x)) = \beta_\varepsilon(y(x))$ a.e. $x \in \Omega$.

We also have

$$\|\nabla\phi_\varepsilon(y) - \nabla\phi^\varepsilon(y)\|_H \le 2\varepsilon \quad \text{for all } y \in H \text{ and } \varepsilon > 0, \tag{2.6}$$

and

$$\|\phi_\varepsilon(y) - \phi^\varepsilon(y)\|_H \le C_\varepsilon\|y\|_H \quad \text{for all } y \in H \text{ and } \varepsilon > 0. \tag{2.7}$$

Now consider the following approximating equations:

$$\begin{cases} y' + Ay + \beta^\varepsilon(y) = Bu, \\ y(0) = y_0. \end{cases} \tag{2.8}$$

As we know, for any $y_0 \in H, u \in L^2(0,T;U)$, (2.8) has a unique solution in $W^{1,2}((0,T];H) \cap C([0,T];H) \cap L^2(0,T;V)$ (cf. [1]).

We also have the following result on (2.8).

LEMMA 2.1. *For $\varepsilon > 0$ given, let $u_1, u_2 \in L^2(0,T;U)$ and $y_{1,0}$, $y_{2,0} \in H$. Suppose that y_1 and y_2 are the solutions of* (2.8) *corresponding to u_1, $y_{1,0}$ and $u_2, y_{2,0}$, respectively. Then*

$$\|y_1 - y_2\|_{C([0,T];H)} \le C[\ \|y_{1,0} - y_{2,0}\|_H + \|u_1 - u_2\|_{L^2(0,T;U)}\ ]$$

for some constant C independent of ε for $1 \ge \varepsilon > 0$.

Proof. Note that (2.8) is equivalent to

$$\begin{cases} y' + Ay + \nabla\phi_\varepsilon(y) = \nabla\phi_\varepsilon(y) - \nabla\phi^\varepsilon(y) + Bu, \\ y(0) = y_0. \end{cases} \tag{2.9}$$

We have

$$\begin{aligned} &(y_1 - y_2)' + A(y_1 - y_2) + \nabla\phi_\varepsilon(y_1) - \nabla\phi_\varepsilon(y_2) \\ &= \nabla\phi_\varepsilon(y_1) - \nabla\phi^\varepsilon(y_1) - [\nabla\phi_\varepsilon(y_2) - \nabla\phi^\varepsilon(y_2)] + B(u_1 - u_2). \end{aligned} \tag{2.10}$$

Multiplying (2.10) by $(y_1 - y_2)$ and integrating over $(0,t)$, by (H_1) and (2.6), one obtains

$$\begin{aligned} &\|(y_1 - y_2)(t)\|_H^2 + 2w\int_0^t \|y_1 - y_2\|_V^2 ds \\ &\le \|y_1(0) - y_2(0)\|_H^2 + 2\alpha\int_0^t \|y_1 - y_2\|_H^2 ds \\ &+ \int_0^t 4\varepsilon\|y_1 - y_2\|_H + C_\delta\int_0^t \|u_1 - u_2\|_V^2 + \delta\int_0^t \|y_1 - y_2\|_H^2 ds \end{aligned} \tag{2.11}$$

for any $\delta > 0$ and some constant C_δ depending on δ.

Applying Gronwall's inequality to (2.11), we get

$$\|(y_1 - y_2)(t)\|_{C([0,T];H)} \le C[\|y_{1,0} - y_{2,0}\|_H + \|u_1 - u_2\|_{L^2(0,T;U)}]$$

for some constant C independent of ε for $0 < \varepsilon \le 1$. This completes the proof.

Next, we recall the approximation g_ε of g and h_ε of h as follows. For the details, we refer readers to [1].

$g_\varepsilon : [0,T] \times H \to R$ is defined by

$$g_\varepsilon(t,y) = \int_{R^n} g(t, p_n y - \varepsilon \wedge_n \tau)\rho_n(\tau)d\tau, \tag{2.12}$$

where $n = [\varepsilon^{-1}]$, ρ_n is a mollifier in R^n, $p_n : H \to X_n$ is the projection of H on X_n which is the finite dimensional space generated by $\{e_i\}_{i=1}^n$, where $\{e_i\}_{i=1}^\infty$ is an orthonormal basis in H. $\wedge_n : R^n \to X_n$ is the operator defined by $\wedge_n(\tau) = \Sigma_{i=1}^n \tau_i e_i$, $\tau = (\tau_1, \dots, \tau_n)$.

$h_\varepsilon : H \to \overline{R}$ is defined by

$$h_\varepsilon(y) = \text{ Inf } \{\|y - x\|_H^2/(2\varepsilon) + h(x); x \in H\}. \tag{2.13}$$

Now we define the penalty $L_\varepsilon : H \times L^2(0,T;U) \to R$ by

$$\begin{aligned} L_\varepsilon(y_0,u) &= \int_0^T [g_\varepsilon(t,y_\varepsilon) + h_\varepsilon(u)]dt + \frac{1}{2}\|y_0 - y^*(0)\|_H^2 \\ &+ \tfrac{1}{2}\|u - u^*\|^2_{L^2(0,T;U)} + \tfrac{1}{2\varepsilon^{1/2}}[d_S(y_\varepsilon(0), y_\varepsilon(T)) + \varepsilon^{1/2}]^2, \end{aligned} \tag{2.14}$$

where y_ε is the solution of (2.8) and $h_\varepsilon, g_\varepsilon$ are given by (2.13) and (2.12), respectively. $d_S(y_\varepsilon(0), y_\varepsilon(T))$ denotes the distance of $(y_\varepsilon(0), y_\varepsilon(T))$ to S.

The approximating optimal control problems are as follows:

$$(\mathrm{P}^\varepsilon) \qquad \text{Minimize } L_\varepsilon(y_0,u) \quad \text{ over } (y_0,u) \in H \times L^2(0,T;U).$$

First of all, we show the existence of the optimalities for (P^ε).

THEOREM 2.2. (P^ε) *has at least one optimal solution.*

Proof. Let $\varepsilon > 0$ be fixed. It is clear that $\mathrm{Inf} L_\varepsilon(y_0,u) > -\infty$.

Let $d = \text{ Inf } L_\varepsilon(y_0,u), (y_0,u) \in H \times L^2(0,T;U)$ and $\{y_{0,n}, u_n\}$ be a minimizing sequence such that

$$d \le L_\varepsilon(y_{0,n}, u_n) \le d + n^{-1}. \tag{2.15}$$

By (H_5), (H_6), and (2.15), $\{u_n\}$ and $\{y_{0,n}\}$ are bounded in $L^2(0,T;U)$ and H, respectively. Without loss of generality, we may assume that $u_n \to \tilde{u}$ weakly in $L^2(0,T;U)$, and $y_{0,n} \to \tilde{y_0}^*$ weakly in H (relabeling if necessary).

By Lemma 2.1, it is clear that L_ε is continuous in $y_{0,n}$ in H and so by (1.6), for each $n \in N$. We may choose $\tilde{y}_{0,n} \in D(\phi) \cap V$ such that

$$\|y_{0,n} - \tilde{y}_{0,n}\|_H < 1/n \quad \text{and} \quad |L_\varepsilon(y_{0,n}, u_n) - L_\varepsilon(\tilde{y}_{0,n}, u_n)| < 1/n. \tag{2.16}$$

Let $\tilde{y}_{\varepsilon,n}$ be the solution of (2.9) corresponding to u_n and $\tilde{y}_{0,n}$, i.e., $\tilde{y}_{\varepsilon,n} \in W^{1,2}([0,T];H) \cap C([[0,T];H) \cap L^2(0,T;V)$ and satisfy

$$\begin{cases} \tilde{y}'_{\varepsilon,n} + A\tilde{y}_{\varepsilon,n} + \nabla\phi_\varepsilon(\tilde{y}_{\varepsilon,n}) = \nabla\phi_\varepsilon(\tilde{y}_{\varepsilon,n}) - \nabla\phi^\varepsilon(\tilde{y}_{\varepsilon,n}) + Bu_n, \\ \tilde{y}_{\varepsilon,n}(0) = \tilde{y}_{0,n}. \end{cases} \tag{2.17}$$

Multiplying (2.17) by $\tilde{y}_{\varepsilon,n}$ and integrating over $(0,t)$, by (H_1), we have

$$\|\tilde{y}_{\varepsilon,n}(t)\|_H^2 + 2w\int_0^t \|\tilde{y}_{\varepsilon,n}(s)\|_V^2 ds + 2\int_0^t \langle \nabla\phi_\varepsilon(\tilde{y}_{\varepsilon,n}), \tilde{y}_{\varepsilon,n}\rangle ds$$
$$\le \|y_{0,n}\|_H^2 + \int_0^t \langle Bu_n, \tilde{y}_{\varepsilon,n}\rangle ds + \alpha\int_0^t \|\tilde{y}_{\varepsilon,n}\|_H^2 ds + \int_0^t \langle \nabla\phi_\varepsilon(\tilde{y}_{\varepsilon,n}) - \nabla\phi^\varepsilon(\tilde{y}_{\varepsilon,n}), \tilde{y}_{\varepsilon,n}\rangle ds.$$

This implies by Gronwall's inequality and (2.6) that

$$\|\tilde{y}_{\varepsilon,n}\|_{C([0,T];H)} + \|\tilde{y}_{\varepsilon,n}\|_{L^2(0,T;V)} + \int_0^t \langle \nabla\phi_\varepsilon(\tilde{y}_{\varepsilon,n}), \tilde{y}_{\varepsilon,n}\rangle ds \le C, \tag{2.18}$$

where C is a constant independent of ε and n (cf. [1]).

Multiplying (2.17) by $\tilde{y}'_{\varepsilon,n}$ and integrating from 0 to t, we have

$$\begin{aligned}&\int_0^t \|\tilde{y}'_{\varepsilon,n}(s)\|_H^2 + \int_0^t \langle A\tilde{y}_{\varepsilon,n}, \tilde{y}_{\varepsilon,n}\rangle + \int_0^t \langle \nabla\phi_\varepsilon(\tilde{y}_{\varepsilon,n}), \tilde{y}'_{\varepsilon,n}\rangle \\ &= \int_0^t \langle Bu_n, \tilde{y}_{\varepsilon,n}\rangle + \int_0^t \langle \nabla\phi_\varepsilon(\tilde{y}_{\varepsilon,n}) - \nabla\phi^\varepsilon(\tilde{y}_{\varepsilon,n}), \tilde{y}'_{\varepsilon,n}\rangle.\end{aligned} \tag{2.19}$$

Note that

$$\begin{aligned}&\int_0^t \langle A\tilde{y}_{\varepsilon,n}, y'_{\varepsilon,n}\rangle ds = \frac{1}{2}\int_0^t \frac{d}{ds}\langle A\tilde{y}_{\varepsilon,n}, \tilde{y}_{\varepsilon,n}\rangle ds \\ &= \tfrac{1}{2}[\langle A\tilde{y}_{\varepsilon,n}(t), \tilde{y}_{\varepsilon,n}(t)\rangle - \langle A\tilde{y}_{0,n}, \tilde{y}_{0,n}\rangle].\end{aligned} \tag{2.20}$$

Since $\{\tilde{y}_{0,n}\} \subset V \cap D(\phi)$ is bounded, we have that $\langle A\tilde{y}_{0,n}, \tilde{y}_{0,n}\rangle$ is bounded.

$$\int_0^t \langle \nabla\phi_\varepsilon(\tilde{y}_{\varepsilon,n}), \tilde{y}'_{\varepsilon,n}\rangle ds = \int_0^t \frac{d}{ds}\phi_\varepsilon(\tilde{y}_{\varepsilon,n}(s)) ds = \phi_\varepsilon(\tilde{y}_{\varepsilon,n}(t)) - \phi_\varepsilon(\tilde{y}_{0,n}). \tag{2.21}$$

Since ϕ_ε is convex and continuously Frechet differentiable and $\{\tilde{y}_{0,n}\}$ is bounded, $\{\phi_\varepsilon(\tilde{y}_{0,n})\}$ is bounded. By (2.18)–(2.21), we obtain

$$\|\tilde{y}'_{\varepsilon,n}\|_{L^2(0,T;H)}^2 + \|\tilde{y}_{\varepsilon,n}\|_{L^2(0,T;V)}^2 + \|\tilde{y}_{\varepsilon,n}\|_{C([0,T];H)}^2 + \phi_\varepsilon(\tilde{y}_{\varepsilon,n}(t)) \le C, \tag{2.22}$$

where C is a constant independent of n and may be different from the constant in (2.18).

The convexity and continuity of ϕ_ε imply that ϕ_ε is bounded below by an affine functional in H. Thus (2.22) shows that

$$\|\tilde{y}'_{\varepsilon,n}\|_{L^2(0,T;H)} + \|\tilde{y}_{\varepsilon,n}\|_{L^2(0,T;V)} + \|\tilde{y}_{\varepsilon,n}\|_{C([0,T];H)} \le C. \tag{2.23}$$

Multiplying (2.17) by $A\tilde{y}_{\varepsilon,n}$ and integrating over $(0,t)$, by (2.18)–(2.23), (H_1), and (1.5) of (H_2), we may obtain

$$\|Ay_{\varepsilon,n}\|_{L^2(0,T;H)}^2 \le C. \tag{2.24}$$

Now by (2.18), (2.23), and (2.24), using the Arzela–Ascoli theorem, we obtain (cf. Lemma 5.1 of [1]) that there exist subsequences of $\{\tilde{y}_{\varepsilon,n}\}$ and $\{u_n\}$, still denoted by them, such that

$\tilde{y}_{\varepsilon,n} \to \tilde{y}_\varepsilon^*$ strongly in $C([0,T];H) \cap L^2(0,T;V)$ as $n \to \infty$,

$A\tilde{y}_{\varepsilon,n} \to A\tilde{y}_\varepsilon^*$ weakly in $L^2(0,T;H)$ as $n \to \infty$, and

$\tilde{y}'_{\varepsilon,n} \to \tilde{y}^*_\varepsilon$ weakly in $L^2(0,T;H)$ as $n \to \infty$.
Thus $\nabla\phi^\varepsilon(\tilde{y}_{\varepsilon,n}) \to \nabla\phi^\varepsilon(\tilde{y}^*_{\varepsilon,n})$ in $L^2(0,T;H)$.
So by taking the limit for $n \to \infty$ in (2.17), we have

$$\begin{cases} (\tilde{y}^*_\varepsilon)' + A\tilde{y}^*_\varepsilon + \nabla\phi^\varepsilon(\tilde{y}^*_\varepsilon) = B\tilde{u}^*, \\ \tilde{y}^*_\varepsilon(0) = \tilde{y}^*_0. \end{cases}$$

Since $d \le L_\varepsilon(y_{0,n}, u_n) \le d + 1/n$, by (2.16) we have

$$d \le L_\varepsilon(\tilde{y}_{0,n}, u_n) \le d + 2/n. \tag{2.25}$$

Since $\tilde{y}_{\varepsilon,n} \to \tilde{y}^*_\varepsilon$ strongly in $C([0,T];H)$, we obtain (note that h_ε is convex)

$$\int_0^T g_\varepsilon(t, \tilde{y}_{\varepsilon,n})dt \to \int_0^T g_\varepsilon(t, \tilde{y}^*_\varepsilon) \text{as } n \to \infty \quad \text{and} \quad \underline{\lim}_{n\to\infty} \int_0^T h_\varepsilon(u_\varepsilon) \ge \int_0^T h_\varepsilon(u^*).$$

This with (2.25) implies

$$L_\varepsilon(\tilde{y}^*_0, u^*) = d.$$

This completes the proof.

The following results are useful in discussing the approximating control problems.

LEMMA 2.3. *Let $u^\varepsilon \in L^2(0,T;U)$ and y^ε_0 be such that $u^\varepsilon \to \overline{u}$ weakly in $L^2(0,T;U)$ and $y^\varepsilon_0 \to \overline{y}_0$ weakly in H as $\varepsilon \to 0$. Let y^ε be the solution of (2.8) corresponding to u^ε and y^ε_0. Then $y^{\varepsilon_n} \to \overline{y}$ strongly in $L^2(0,T;H)$, $y^{\varepsilon_n}(t) \to \overline{y}(t)$ weakly in H for every $t \in [0,T]$ on a subsequence $\{\varepsilon_n\}$, where $\overline{y} \in W^{1,2}((0,T];H) \cap C([0,T];H)$ is the solution of the problem*

$$\begin{cases} \overline{y}' + A\overline{y} + \beta(\overline{y}) \ni B\overline{u} \quad \text{a.e.} \quad t \in (0,T), \\ \overline{y}(0) = \overline{y}_0. \end{cases}$$

Proof. We have

$$\begin{cases} (y^\varepsilon)' + Ay^\varepsilon + \nabla\phi_\varepsilon(y^\varepsilon) = \nabla\phi_\varepsilon(y^\varepsilon) - \nabla\phi^\varepsilon(y^\varepsilon) + Bu^\varepsilon, \\ y^\varepsilon(0) = y^\varepsilon_0. \end{cases} \tag{2.26}$$

Multiplying (2.26) by y^ε and integrating over $(0,t)$, by (H_1) and (2.6), we obtain

$$\|y^\varepsilon(t)\|^2_H + \int_0^T \|y^\varepsilon\|^2_V + \int_0^t \phi_\varepsilon(y^\varepsilon)ds \le C \tag{2.27}$$

for any $t \in (0,T)$.

Multiplying (2.26) by $s.(y^\varepsilon(s))'$ and integrating over $(0,t)$ by (2.6), we get

$$\begin{aligned} &\int_0^t s\|(y^\varepsilon)'(s)\|^2_H - \int_0^t s\langle Ay^\varepsilon, y^\varepsilon\rangle ds \\ &+t\langle Ay^\varepsilon(t), y^\varepsilon(t)\rangle - \int_0^t \phi_\varepsilon(y^\varepsilon)ds + t\phi_\varepsilon(y^\varepsilon(t)) \\ &\le \int_0^t \langle Bu^\varepsilon, s(y^\varepsilon)'\rangle + t\int_0^t 2\varepsilon\|(y^\varepsilon)'\|_H ds. \end{aligned}$$

By (2.27) and the boundedness of

$$\int_0^t \langle Ay^\varepsilon, y^\varepsilon\rangle,$$

which is from (H_1) and (2.27), we obtain

$$\int_0^t s\|(y^\varepsilon)'(s)\|_H^2 ds + t\langle Ay^\varepsilon(t), y^\varepsilon(t)\rangle + t\phi_\varepsilon(y^\varepsilon(t)) \le C \tag{2.28}$$

for each $t \in [0,T]$, where C is a constant independent of ε and $t \in [0,T]$.

Multiplying (2.26) by $t\phi_\varepsilon(y^\varepsilon(t))$ and integrating from 0 to t, by (2.27), (2.28), (H_1), and (1.5) of (H_2), we obtain

$$\int_0^t s\|\phi_\varepsilon(y^\varepsilon(t))\|_H^2 \le C. \tag{2.29}$$

Multiplying (2.26) by $sAy^\varepsilon(s)$ and integrating over $(0,t)$, by (2.6), using the same arguments as above, we obtain

$$\int_0^t s\|Ay^\varepsilon(s)\|_H^2 \le C. \tag{2.30}$$

Now by (2.27)–(2.30), using the Arzela–Ascoli theorem, there exists a $\overline{y} \in W^{1,2}((0,T];H) \cap L^2(0,T;V) \cap C([0,T];H)$ such that (cf. [1])

$y^{\varepsilon_n} \to \overline{y}$ weakly in $L^2(0,T;V)$ and strongly in $L^2(\delta,T;H)$ for each $\delta > 0$,

$Ay^{\varepsilon_n} \to A\overline{y}$ weakly in every $L^2(\delta,T;H)$,

$(y^{\varepsilon_n})' \to \overline{y}'$ weakly in every $L^2(\delta,T;H)$, and

$\nabla\phi_{\varepsilon_n}(y^{\varepsilon_n}) \to \xi$ weakly in every $L^2(\delta,T;H)$.

It follows that (cf. [1], [5]) $y^{\varepsilon_n} \to \overline{y}$ strongly in $L^2(0,T;H)$ and $\nabla\phi_{\varepsilon_n}(y^{\varepsilon_n}) \to \xi \in \partial\phi(\overline{y})$ weakly in every $L^2(\delta,T;H)$ on a subsequence $\{\varepsilon_n\}$. Also, $\overline{y}$ satisfies

$$\begin{cases} \overline{y}' + A\overline{y} + \beta(\overline{y}) \ni B\overline{u}, \\ \overline{y}(0) = \overline{y}_0. \end{cases}$$

Now by (2.27)–(2.30), using the same arguments as those in [1], (cf. Lemma 5.2 of [1]), we get $y^{\varepsilon_n}(t) \to \overline{y}(t)$ weakly in H for each $t \in [0,T]$.

This completes the proof.

LEMMA 2.4. *Let $y_0 \in D(\phi) \cap V$, $u \in L^2(0,T;U)$, and y^ε be the solution of the problem* (2.8). *Then $y^\varepsilon \to y$ in $C([0,T];H) \cap L^2(0,T;V)$, where $y \in W^{1,2}([0,T];H) \cap C([0,T];H)$ is the solution of the problem* (1.1) *with the initial condition $y(0) = y_0$. Moreover,*

$$\|y^\varepsilon - y\|_{C([0,T];H)} \le C\varepsilon^{1/2}, \tag{2.31}$$

where C is a constant independent of ε.

This result is due to Barbu (cf. Lemma 5.1 of [1]).

LEMMA 2.5. *Let $(y_{\varepsilon,0}, u_\varepsilon)$ be optimal for the problem (P^ε) and y_ε be the solution corresponding to u_ε and $y_{\varepsilon,0}$. Then $y_{\varepsilon_n} \to y^*$ in $C([0,T];H)$ and $u_{\varepsilon_n} \to u^*$ in $L^2(0,T;U)$ strongly as $\varepsilon_n \to 0$ on some subsequence $\{\varepsilon_n\}$.*

Proof. For each $\varepsilon > 0$, let y^ε be the solution of (2.8) corresponding to $y^*(0)$ and u^*. By Lemma 2.4, $y^\varepsilon \to y^*$ in $C([0,T];H)$. We have (cf. [1], [5])

$$g_\varepsilon(t, y^\varepsilon(t)) \to g(t, y^*(t)) \quad \text{for all } t \in [0,T], \quad h_\varepsilon(u^*(t)) \to h(u^*(t)) \text{ a.e. } t \in (0,T).$$

So

$$\lim_{\varepsilon\to 0} \int_0^t g_\varepsilon(t, y^\varepsilon(t)) = \int_0^t g(t, y^*(t))dt, \ \lim_{\varepsilon\to 0} \int_0^t h_\varepsilon(u^*(t)) = \int_0^t h(u^*(t)). \tag{2.32}$$

Since $(y_{\varepsilon,0}, u_\varepsilon)$ is optimal for the problem (P_ε), we have

$$\begin{aligned} &L_\varepsilon(y_{\varepsilon,0}, u_\varepsilon) \le L_\varepsilon(y^*(0), u^*) \\ &= \int_0^T [g_\varepsilon(t, y^\varepsilon) + h_\varepsilon(u^*)]dt + \frac{1}{2\varepsilon^{1/2}}[d_S(y^\varepsilon(0), y^\varepsilon(t)) + \varepsilon^{1/2}]^2. \end{aligned}$$

By (2.31), we get

$$\begin{aligned} &\tfrac{[ds(y^\varepsilon(0), y^\varepsilon(T)) + \varepsilon^{1/2}]^2}{2\varepsilon^{1/2}} \\ &\le \tfrac{\{[\|y^\varepsilon(0) - y^*(0)\|_H^2 + \|y^\varepsilon(T) - y^*(T)\|_H^2]^{1/2} + \varepsilon^{1/2}\}^2}{2\varepsilon^{1/2}} \\ &\le C\varepsilon^{1/2} \to 0 \ \text{ as } \ \varepsilon \to 0. \end{aligned} \tag{2.33}$$

By (2.32) and (2.33), we yield

$$\overline{\lim}_{\varepsilon\to 0} L_\varepsilon(y_{\varepsilon,0}, u_\varepsilon) \le L(y^*, u^*). \tag{2.34}$$

On the other hand, one can easily verify that $\{u_\varepsilon\}$ and $\{y_{\varepsilon,0}\}$ are bounded in $L^2(0,T;U)$ and H, respectively.

Without loss of generality, we may assume that $u_\varepsilon \to u_1^*$, $y_{\varepsilon,0} \to y_{1,0}^*$ weakly in $L^2(0,T;U)$ and H, respectively.

Let $y_1^*(t)$ be the solution of (2.8) corresponding to $y_{1,0}^*$ and u_1^*.

By Lemma 2.3, $y_\varepsilon \to y_1^*$ strongly in $L^2(0,T;H)$.

Consider $\underline{\lim}_{\varepsilon\to 0} L_\varepsilon(y_{\varepsilon,0}, u_\varepsilon)$.

By the arguments in [1],

$$\underline{\lim}_{\varepsilon\to 0} \int_0^T h_\varepsilon(u_\varepsilon) \ge \int_0^T h(u_1^*),\ \lim_{\varepsilon\to 0} \int_0^T g_\varepsilon(t, y_\varepsilon) = \int_0^T g(t, y_1^*).$$

Thus

$$\underline{\lim}_{\varepsilon\to 0} \int_0^T L_\varepsilon(y_{\varepsilon,0}, u_\varepsilon) \ge \int_0^T [g(t, y_1^*) + h(u_1^*)]. \tag{2.35}$$

By (2.14) and (2.34), one can check easily that

$$\frac{1}{2\varepsilon^{1/2}}[d_S(y_\varepsilon(0), y_\varepsilon(T)) + \varepsilon^{1/2}]^2 \le C.$$

Thus $d_S(y_\varepsilon(0), y_\varepsilon(T)) \to 0$ as $\varepsilon \to 0$.

By Lemma 2.3, there exists a subsequence of $\{\varepsilon\}_{\varepsilon>0}$, still denoted by itself, such that $y_\varepsilon(T) \to y_1^*(T)$ weakly in H.

Since S is convex and closed in $H \times H$, we have

$$(y_{1,0}^*, y_1^*(T)) \in S. \tag{2.36}$$

Indeed, suppose that (2.36) did not hold.

By the Hahn–Banach theorem, there exist $f \in H \times H$ and $\alpha \in R$ such that

$$f(y_{1,0}^*, y_1^*(T)) > \alpha \quad \text{and} \quad f(y_0, y_1) < \alpha \quad \text{for all } (y_0, y_1) \in S. \tag{2.37}$$

Choose $\gamma > 0$ such that

$$f(y_{1,0}^*, y_1^*(T)) > \alpha + \gamma. \tag{2.38}$$

Since $d_S(y_\varepsilon(0), y_\varepsilon(T)) \to 0$, there exists $(y'_\varepsilon(0), y'_\varepsilon(T)) \in S$ such that

$$\|y'_\varepsilon(0) - y_\varepsilon(0)\|_H + \|y'_\varepsilon(T) - y_\varepsilon(T)\|_H < \varepsilon.$$

It is clear that $(y'_\varepsilon(0), y'_\varepsilon(T)) \to (y^*_{1,0}, y^*_1(T))$ weakly in $H \times H$. Thus

$$f(y'_\varepsilon(0), y'_\varepsilon(T)) \to f(y^*_{1,0}, y^*_1(T)) \quad \text{as} \quad \varepsilon \to 0. \tag{2.39}$$

By (2.37)–(2.39), we obtain

$$\alpha \geq \overline{\lim}_{\varepsilon \to 0} f(y'_\varepsilon(0), y'_\varepsilon(T)) = f(y^*_{1,0}, y^*_1(T)) > \alpha + \gamma.$$

This contradiction shows (2.36).

Now (2.36) yields (y^*_1, u^*_1) is admissible, i.e., $(y^*_1, u^*_1) \in A_{ad}$ and

$$L(y^*_1, u^*_1) \geq L(y^*, u^*). \tag{2.40}$$

Thus by (2.34), (2.35), and (2.40), we imply

$$\lim_{\varepsilon \to 0} L_\varepsilon(y_{\varepsilon,0}, u_\varepsilon) = L(y^*, u^*).$$

By (2.14), we obtain

$$u_\varepsilon \to u^* \quad \text{strongly in } L^2(0,T;U) \quad \text{and} \quad y_{\varepsilon,0} \to y^*(0) \quad \text{strongly in } H. \tag{2.41}$$

Now we prove $y_\varepsilon \to y^*$ strongly in $C([0,T];H)$.

Since $y^*(0) \in D(\phi) \cap V$, by Lemma 2.1, we may assume without loss of generality that $y_{\varepsilon,0} \in D(\phi) \cap V$ (since $\overline{D(\phi) \cap V} = H$ by the assumption (1.6)).

Multiply the equation

$$\begin{cases} y'_\varepsilon + Ay_\varepsilon + \nabla\phi_\varepsilon(y_\varepsilon) = \nabla\phi_\varepsilon(y_\varepsilon) - \nabla\phi^\varepsilon(y_\varepsilon) + Bu_\varepsilon, \\ y_\varepsilon(0) = y_{\varepsilon,0} \qquad \text{a.e.} \quad t \in (0,T) \end{cases} \tag{2.42}$$

by y_ε, y'_ε and $\nabla\phi_\varepsilon(y_\varepsilon)$, respectively. Using the similar arguments in the proof of Lemma 2.3 and noting (2.41) and the fact that $y_{\varepsilon,0}$ and $y^*(0) \in D(\phi) \cap V$, we obtain after some manipulation

$$\|y_\varepsilon(t)\|^2_H + \int_0^T [\|y_\varepsilon(t)\|^2_V + \|y'_\varepsilon(t)\|^2_H + \|\nabla\phi_\varepsilon(y_\varepsilon(t))\|^2_H]dt \leq C \quad \text{for } t \in [0,T],\ \varepsilon > 0. \tag{2.43}$$

On the other hand, by (2.42) we have

$$\begin{cases} (y_\varepsilon - y_\lambda)' + A(y_\varepsilon - y_\lambda) + \nabla\phi_\varepsilon(y_\varepsilon) - \nabla\phi_\lambda(y_\lambda) \\ \quad = \nabla\phi_\varepsilon(y_\varepsilon) - \nabla\phi^\varepsilon(y_\varepsilon) + \nabla\phi^\lambda(y_\lambda) - \nabla\phi_\lambda(y_\lambda) + B(u_\varepsilon - u_\lambda), \\ y_\varepsilon - y_\lambda(0) = y_{\varepsilon,0} - y_{\lambda,0}. \end{cases} \tag{2.44}$$

Multiplying (2.44) by $(y_\varepsilon - y_\lambda)$ and using assumption (H_1) and (2.6), we obtain

$$\begin{aligned} &\tfrac{1}{2}\tfrac{d}{dt}\|y_\varepsilon(t) - y_\lambda(t)\|^2 + w\|y_\varepsilon(t) - y_\lambda(t)\|^2 \\ &+\langle \nabla\phi_\varepsilon(y_\varepsilon(t)) - \nabla\phi_\lambda(y_\lambda(t)), y_\varepsilon(t) - y_\lambda(t)\rangle \\ &\leq (1+\alpha)\|y_\varepsilon(t) - y_\lambda(t)\|^2_H + (\varepsilon + \lambda) \\ &+\langle B(u_\varepsilon - u_\lambda), y_\varepsilon - y_\lambda\rangle + \tfrac{1}{2}\|y_{\varepsilon,0} - y_{\lambda,0}\|^2_H. \end{aligned} \tag{2.45}$$

Recall (cf. [5]) that

$$\nabla\phi_\varepsilon(y_\varepsilon) = \varepsilon^{-1}(y_\varepsilon - (I+\varepsilon\partial\phi)^{-1}y_\varepsilon) \in \partial\phi((I+\varepsilon\partial\phi)^{-1}y_\varepsilon).$$

Since by (2.43) $\{\nabla\phi_\varepsilon(y_\varepsilon)\}$ is bounded in $L^2(0,T;H)$, we have by integrating (2.45) over $(0,t)$ that

$$\begin{aligned}
&\tfrac{1}{2}\|y_\varepsilon(t)-y_\lambda(t)\|_H^2 + w\int_0^t \|y_\varepsilon(s)-y_\lambda(s)\|_V^2 ds \\
&\le \tilde{\alpha}\int_0^t \|y_\varepsilon(s)-y_\lambda(s)\|_H^2 ds + \int_0^T \|B(u_\varepsilon-u_\lambda)\|_H^2 + \tfrac{1}{2}\|y_{\varepsilon,0}-y_{\lambda,0}\|_H^2 \\
&+\int_0^t \langle \nabla\phi_\varepsilon(y_\varepsilon(s)) - \nabla\phi_\lambda(y_\lambda(s)), \varepsilon\nabla\phi_\varepsilon(y_\varepsilon(s)) - \lambda\nabla\phi_\lambda(y_\lambda(s))ds\rangle + (\varepsilon+\lambda) \\
&\le \tilde{\alpha}\int_0^t \|y_\varepsilon(s)-y_\lambda(s)\|_H^2 ds + C(\varepsilon+\lambda) \\
&+\tfrac{1}{2}\|y_{\varepsilon,0}-y_{\lambda,0}\|_H^2 + \|B(u_\varepsilon-u_\lambda)\|^2_{L^2(0,T;H)}.
\end{aligned}$$

By Gronwall's inequality,

$$\begin{aligned}
&\|y_\varepsilon(t)-y_\lambda(t)\|_H^2 + \int_0^T \|y_\varepsilon(t)-y_\lambda(t)\|_V^2 dt \\
&\le C(\varepsilon+\lambda) + \|y_{\varepsilon,0}-y_{\lambda,0}\|_H^2 + \|B(u_\varepsilon-u_\lambda)\|^2_{L^2(0,T;H)}.
\end{aligned} \tag{2.46}$$

By (2.41), (2.43), and (2.46), using the same arguments in [1] (cf. Theorem 4.5 of [1]), we obtain

$$y_\varepsilon \to y^* \text{ in } C([0,T];H)\cap L^2(0,T;V).$$

This completes the proof.

3. Necessary conditions on optimality. Let ∂g be the generalized gradient of $y \to g(t,y)$ and ∂h be the subdifferential of h (cf. [2]).

Let $Y^* = (H^s(\Omega))' + V'$ which is the dual of $Y = H^s(\Omega)\cap V$ with $s > N/2$. We state the main results of the necessary conditions on optimality as follows.

THEOREM 3.1. *Let (y^*,u^*) be an optimal pair of problem* (P). *Suppose that* (H_1)–(H_6) *hold. Then there exist the function $p \in L^\infty(0,T;H)\cap L^2(0,T;V)\cap BV([0,T];Y^*)$, the measure $\mu \in (L^\infty(Q))^*$, and $\lambda_0 \in R$ with $\lambda_0 \ge 0$ satisfying*

$$\begin{gathered}
p' - Ap - \mu \in L^\infty(0,T;H), \\
p'(t) - Ap(t) - \mu \in \lambda_0\partial g(t,y^*(t)) \quad a.e. \quad in\ (0,T), \\
\langle\, p(0), x_0 - y^*(0)\,\rangle - \langle\, p(T), x_1 - y^*(T)\,\rangle \le 0
\end{gathered}$$

for all $(x_0,x_1) \in S$ (transversality condition).

$$B^*p(t) \in \lambda_0\partial h(u^*(t)) \quad a.e.\ t \in (0,T),$$

and $(\lambda_0,p) \ne 0$.

Proof. Let $(y_{\varepsilon,0},u_\varepsilon)$ be optimal for the problem (P^ε) and y_ε be the solution of (2.8) corresponding to $(y_{\varepsilon,0},u_\varepsilon)$.

For any $v \in L^2(0,T;U), \eta \in H$ given, let $u_\varepsilon^\lambda = u_\varepsilon + \lambda v$, $y_{\varepsilon,0}^\lambda = y_{\varepsilon,0} + \lambda\eta$, $\lambda > 0$, and let y_ε^λ be the solution of the problem (2.8) corresponding to $y_{\varepsilon,0}^\lambda$ and u_ε^λ.

By Lemma 2.1, $y_\varepsilon^\lambda \to y_\varepsilon$ strongly in $C([0,T];H)$ as $\lambda \to 0$.

Since $(y_{\varepsilon,0}, u_\varepsilon)$ is optimal, we have $L_\varepsilon(y_{\varepsilon,0}^\lambda, u_\varepsilon^\lambda) \geq L_\varepsilon(y_{\varepsilon,0}, u_\varepsilon)$ for any $\lambda > 0$, $v \in L^2(0,T;U)$, and $\eta \in H$.

Thus

$$\frac{L_\varepsilon(y_{\varepsilon,0}^\lambda, u_\varepsilon^\lambda) - L_\varepsilon(y_{\varepsilon,0}, u_\varepsilon)}{\lambda} \geq 0. \tag{3.1}$$

Let $\nabla g_\varepsilon(t, y_\varepsilon)$ be the gradient of g_ε to the second variable at y_ε, and let $\nabla h_\varepsilon(u_\varepsilon)$ be the gradient of h_ε at u_ε. After some simple calculations, we obtain

$$\lim_{\lambda\to 0} \int_0^T \frac{g_\varepsilon(t, y_\varepsilon^\lambda) - g_\varepsilon(t, y_\varepsilon)}{\lambda} = \int_0^T \langle \nabla g^\varepsilon(t, y_\varepsilon), z_\varepsilon \rangle \tag{3.2}$$

with $z_\varepsilon \in C([0,T];H) \cap W^{1,2}([0,T];H)$ satisfying

$$\begin{cases} z_\varepsilon' + Az_\varepsilon + \dot{\beta}^\varepsilon(y_\varepsilon) z_\varepsilon = Bv, \\ z_\varepsilon(0) = \eta, \end{cases} \tag{3.3}$$

$$\lim_{\lambda\to 0} \int_0^T \frac{h_\varepsilon(u_\varepsilon^\lambda) - h_\varepsilon(u_\varepsilon)}{\lambda} dt = \int_0^T \langle \nabla h_\varepsilon(u_\varepsilon), v \rangle dt, \tag{3.4}$$

$$\begin{aligned} &\lim_{\lambda\to 0} \tfrac{1}{2} \Big[\tfrac{\|u_\varepsilon^\lambda - u^*\|_H^2 - \|u_\varepsilon - u^*\|_H^2}{\lambda} + \tfrac{\|y_{\varepsilon,0}^\lambda - y^*(0)\|_H^2 - \|y_{\varepsilon,0} - y^*(0)\|_H^2}{\lambda} \Big] \\ &= \int_0^T \langle u_\varepsilon - u^*, v \rangle dt + \langle y_{\varepsilon,0} - y^*(0), \eta \rangle, \end{aligned} \tag{3.5}$$

and

$$\begin{aligned} &\lim_{\lambda\to 0} \tfrac{1}{2\varepsilon^{1/2}} \left\{ \tfrac{[d_S(y_\varepsilon^\lambda(0), y_\varepsilon^\lambda(T)) + \varepsilon^{1/2}]^2 - [d_S(y_\varepsilon(0), y_\varepsilon(T)) + \varepsilon^{1/2}]^2}{\lambda} \right\} \\ &= \tfrac{1}{\varepsilon^{1/2}} [\, d_S(y_\varepsilon(0), y_\varepsilon(T)) + \varepsilon^{1/2} \,] \, [\, \langle a_\varepsilon \, , \, \eta \rangle + \langle b_\varepsilon \, , \, z^\varepsilon(T) \rangle \,], \end{aligned} \tag{3.6}$$

where $(a_\varepsilon, b_\varepsilon) \in \partial d_S(y_\varepsilon(0), y_\varepsilon(T))$, the subdifferential of d_S at $(y_\varepsilon(0), y_\varepsilon(T))$.

Since S is convex and closed,

$$\partial ds(y_\varepsilon(0), y_\varepsilon(T)) = \begin{cases} \nabla ds(y_\varepsilon(0), y_\varepsilon(T)) & \text{if } (y_\varepsilon(0), y_\varepsilon(T)) \notin S, \\ 0 & \text{if } (y_\varepsilon(0), y_\varepsilon(T)) \in S \end{cases}$$

and

$$\|a_\varepsilon\|_H^2 + \|b_\varepsilon\|_H^2 = 1 \quad \text{if } (y_\varepsilon(0), y_\varepsilon(T)) \notin S. \tag{3.7}$$

Let

$$\lambda_\varepsilon = \frac{\varepsilon^{1/2}}{ds(y_\varepsilon(0), y_\varepsilon(T)) + \varepsilon^{1/2}}. \tag{3.8}$$

By (3.7) and (3.8)

$$2 \geq |\lambda_\varepsilon|^2 + \|a_\varepsilon\|_H^2 + \|b_\varepsilon\|_H^2 \geq 1. \tag{3.9}$$

By (3.1)–(3.6), we yield

$$\begin{aligned}(3.10)\quad &\lambda_\varepsilon\left[\int_0^T \langle \nabla g_\varepsilon(t,y_\varepsilon), z_\varepsilon\rangle dt + \int_0^T \langle \nabla h_\varepsilon(u_\varepsilon), v\rangle dt\right] + [\langle a_\varepsilon, \eta\rangle + \langle b_\varepsilon, z^\varepsilon(T)\rangle] \\ &\geq \int_0^T \langle u^* - u_\varepsilon\ ,\ v\rangle dt + \langle y^*(0) - y_{\varepsilon,0}\ ,\ \eta\rangle \equiv q(\varepsilon, v, \eta).\end{aligned}$$

By (3.9), we may assume that (relabeling if necessary)

$$(3.11)\qquad \lambda_\varepsilon \to \lambda_0 \quad \text{and} \quad a_\varepsilon \to a, b_\varepsilon \to b \text{ weakly in } H.$$

It follows by Theorem 1.14 of [1] that the boundary value problem

$$(3.12)\qquad \begin{cases} p_\varepsilon' - Ap_\varepsilon - \dot\beta^\varepsilon(y_\varepsilon)p_\varepsilon = \lambda_\varepsilon \nabla g_\varepsilon(t, y_\varepsilon) & \text{in } Q, \\ p_\varepsilon(T) = -b_\varepsilon & \text{in } \Omega \end{cases}$$

has a unique solution $p_\varepsilon \in L^2(0,T;V) \cap C([0,T];H)$ with $p_\varepsilon' \in L^2(0,T;V')$.

Since $y_\varepsilon \to y^*$ strongly in $C([0,T];H) \cap L^2(0,T;V)$ from Lemma 2.5, by the same arguments as those in [1] (cf. Lemma 5.3 of [1]), there exists $p \in BV([0,T];Y^*) \cap L^2(0,T;V) \cap L^\infty(0,T;H)$ and $\mu \in (L^\infty(Q))^*$ such that on some subsequence $\{\varepsilon_n\}$,

$$(3.13)\qquad p_{\varepsilon_n} \to p \quad \text{in } L^2(0,T;H), \qquad \text{weakly in } L^2(0,T;V), \quad \text{weak star in } L^\infty(0,T;H),$$

where $Y^* = (H^s(\Omega))' \cap V'$ with $s > N/2$ which is the dual of $Y = H^s(\Omega) \cap V$, and

$$(3.14)\qquad p_{\varepsilon_n}(t) \to p(t) \quad \text{strongly in } Y^* \text{ and weakly in } H \text{ for every } t \in [0,T],$$

$$(3.15)\qquad \dot\beta^{\varepsilon_n}(y_{\varepsilon_n})p_{\varepsilon_n} \to \mu \quad \text{weak star in } (L^\infty(Q))^*,$$

$$(3.16)\qquad \begin{aligned} &\nabla g_{\varepsilon_n}(t, y_{\varepsilon_n}) \to \xi_1 \text{ weak star in } L^\infty(0,T;H) \\ &\quad \text{with } \xi_1(t) \in \partial g(t, y^*(t)) \quad \text{a.e. } t \in (0,T), \end{aligned}$$

where $\partial g(t,y)$ is the generalized gradient of $y \to g(t,y)$.

Now letting $\varepsilon_n \to 0$ in (3.12), it follows by (3.11) and (3.13)–(3.16) that p satisfies the equations

$$(3.17)\qquad \begin{cases} p' - Ap - \mu \in \lambda_0 \partial g(t, y^*) & \text{a.e. in } (0,T), \\ p(T) = -b \end{cases}$$

and

$$(3.18)\qquad p' - Ap - \mu \in L^\infty(0,T;H),$$

where p' is the derivative in the sense of V'-valued distribution.

It follows from (3.3), (3.10), and (3.12) that

$$(3.19)\qquad -\int_0^T \langle B^* p_\varepsilon, v\rangle + \lambda_\varepsilon \int_0^T \langle \nabla h_\varepsilon(u_\varepsilon), v\rangle + \langle a_\varepsilon, \eta\rangle - \langle p_\varepsilon(0), \eta\rangle \geq q(\varepsilon, v, \eta)$$

for all $v \in L^2(0,T;U)$ and $\eta \in H$.

Since $u_\varepsilon \to u^*$ in $L^2(0,T;U)$, by a standard argument (cf. [1], [5]), we have

$$\int_0^T \langle \nabla h_\varepsilon(u_\varepsilon), v\rangle \to \int_0^T \langle \xi(t), v\rangle \text{ with } \xi(t) \in \partial h(u^*(t)) \quad \text{a.e. } t \in (0,T) \tag{3.20}$$

for all $v \in L^2(0,T;U)$, where ∂h denotes the subdifferential of h.

Letting $\varepsilon \to 0$ in (3.19), it follows from (3.13) and (3.20) that

$$-\int_0^T \langle B^*p, v\rangle + \lambda_0 \int_0^T \langle \xi(t), v\rangle + \langle a, \eta\rangle - \langle p(0), \eta\rangle \geq 0 \tag{3.21}$$

for all $v \in L^2(0,T;U)$ and $\eta \in H$.

Since $(a_\varepsilon, b_\varepsilon) \in d_S(y_{\varepsilon,0}, y_\varepsilon(T))$, we have

$$\langle a_\varepsilon, x_0 - y_{0,\varepsilon}\rangle + \langle b_\varepsilon, x_1 - y_\varepsilon(T)\rangle \leq 0 \text{ for all } (x_0, x_1) \in S.$$

Thus

$$\begin{aligned} &\langle a_\varepsilon, x_0 - y^*(0)\rangle + \langle b_\varepsilon, x_1 - y^*(T)\rangle \\ &\leq \langle a_\varepsilon, y_{\varepsilon,0} - y^*(0)\rangle + \langle b_\varepsilon, y_\varepsilon(T) - y^*(T)\rangle \\ &\equiv q_1(\varepsilon) \end{aligned} \tag{3.22}$$

for all $(x_0, x_1) \in S$.

By Lemma 2.5 and (3.9), $q_1(\varepsilon) \to 0$ as $\varepsilon \to 0$. We claim $(\lambda_0, a, b) \neq 0$. Indeed, if $\lambda_0 = 0$, by (3.9), $\|a_\varepsilon\|_H^2 + \|b_\varepsilon\|_H^2 \geq \delta > 0$.

Since S is finite codimensional (by (H_3)), so is $S - (y^*(0), y^*(T))$ (cf. Proposition 3.4 of [3]).

Thus by (3.22), we yield $(a, b) \neq 0$ and

$$\langle a, x_0 - y^*(0)\rangle + \langle b, x_1 - y^*(T)\rangle \leq 0 \tag{3.23}$$

for all $(x_0, x_1) \in S$ (cf. Lemma 3.6 [3]).

Now by (3.21) we have $(\lambda_0, p) \neq 0$. Indeed, if $\lambda_0 = 0$, then $(a, b) \neq 0$. By (3.17), $p \neq 0$ if $\mu \neq 0$ or $b \neq 0$ If $\mu = 0$ and $b = 0$, we have $p(0) = 0$. Then (3.21) makes $\langle a, \eta\rangle \geq 0$ for all $\eta \in H$. So $a = 0$. This contradiction implies that $p \neq 0$. Next we prove the transversality condition.

For any $(x_0, x_1) \in S$, let $v = 0, \eta = x_0 - y^*(0)$ in (3.21), which yields

$$\langle p(0) - a, x_0 - y^*(0)\rangle \leq 0. \tag{3.24}$$

By (3.23) and (3.24) and noting that $p(T) = -b$, we obtain

$$\langle p(0), x_0 - y^*(0)\rangle - \langle p(T), x_1 - y^*(T)\rangle \leq 0 \tag{3.25}$$

for all $(x_0, x_1) \in S$.

Finally in (3.21), by letting $\eta = 0$ and using a standard argument in [1], it follows from (3.20) that

$$B^*p \in \lambda_0 \partial h(u^*(t)) \quad \text{a.e.} \quad t \in (0,T). \tag{3.26}$$

Thus (3.17), together with (3.18), (3.25), and (3.26) completes the proof.

4. Some remarks.

Remark 4.1. Let $S = Q_1 \times Q_2$, where $Q_1 \subset D(\phi) \cap V \subset H$, $Q_2 \subset H$ are convex and closed sets. Suppose that either Q_1 or Q_2 is finite codimensional. Then all results in Theorem 3.1 remain true without assumption (H_3). (Note that one of Q_1 and Q_2 has finite codimensionality which cannot imply that S does.)

Indeed, we need only to show that $(\lambda_0, a, b) \neq 0$ in (3.11).

Suppose that $\lambda_0 = 0$ and $a = 0$. Then by (3.9) we have $\|b_\varepsilon\|_H^2 \geq \delta > 0$ for constant δ. Now by (3.22), we have $\langle b_\varepsilon, x_1 - y^*(T)\rangle \leq \tilde{q}_1(\varepsilon)$ for all $x_1 \in Q_2$. $\tilde{q}_1(\varepsilon) \to 0$ as $\varepsilon \to 0$.

Since Q_2 is finite codimensional, so is $Q_2 - y^*(T)$ (cf. [3]). By the same arguments in the proof of Theorem 3.1, one obtains $b_\varepsilon \to b \neq 0$.

Remark 4.2. Let β be locally Lipschitz continuous, monotone on R, and satisfying the growth condition

$$0 \leq \beta'(r) \leq c(|\beta(r)| + |r| + 1) \quad \text{a.e. } r \in R. \tag{4.1}$$

Then $D(\phi)$ is dense in H (cf. section 5.3 of [1]). Thus, the results of Theorem 3.1 remain true.

Moreover, $p \in AC([0,T], Y^*) \cap C_w([0,T]; H)$ and $\mu \in L^1(Q)$, where $C_w([0,T]; H)$ denotes the space of all weakly continuous functions from $[0,T]$ to H.

Remark 4.3. Let β be globally Lipschitz and monotone. Then $\{\dot{\beta}^\varepsilon\}$ is uniformly bounded on R (cf. Proposition 5.3 of [1]) and $\overline{D(\phi)} = H$. In this case, we may weaken the assumption (H_3) on S.

Start from the variational equation (3.3):

$$\begin{cases} z_\varepsilon' + Az_\varepsilon + \dot{\beta}^\varepsilon(y_\varepsilon) z_\varepsilon = Bv, \\ z_\varepsilon(0) = \eta. \end{cases} \tag{4.2}$$

Multiplying (4.2) by $z_\varepsilon, z_\varepsilon', Az_\varepsilon$ and integrating over $(0,t)$, by the uniform boundedness of $\{\dot{\beta}^\varepsilon\}$, we may have

$$\|z_\varepsilon\|_{C([0,T];H)} + \|z_\varepsilon'\|_{L^2(0,T;H)} + \|Az_\varepsilon\|_{L^2(0,T;H)} \leq C. \tag{4.3}$$

On the other hand, $\dot{\beta}^\varepsilon(y_\varepsilon) z_\varepsilon$ is weakly compact in $L^2(Q)$, so $\dot{\beta}(y_\varepsilon) z_\varepsilon \to \tilde{\mu}$ weakly in $L^2(Q)$.

Thus by (4.2) and (4.3), using the Arzela–Ascoli theorem, on some subsequence ε_n, we have

$$\begin{aligned} &z_{\varepsilon_n} \to z \text{ in } C([0,T]; H) \text{ strongly}, \\ &Az_{\varepsilon_n} \to Az \text{ weakly in } L^2(0,T;H), \\ &z_{\varepsilon_n}' \to z_{\varepsilon_n}' \text{ weakly in } L^2(0,T;H). \end{aligned}$$

So $z \in C([0,T]; H) \cap W^{1,2}([0,T]; H)$ and satisfies

$$\begin{cases} z' + Az + \mu = Bv, \\ z(0) = \eta. \end{cases} \tag{4.4}$$

Let

$$\tilde{R}_r = \big\{(\eta, z(T)) : \ \eta \in B_1(0), \text{ where } z \text{ is the solution of (4.3) with } v \in \tilde{B}_r(0)\big\},$$

where $B_1(0)$ is the unit ball of H, and where $\tilde{B}_r(0)$ is the ball in $L^2(0,T;U)$ with radius r, center 0.

Instead of (H_3), we may assume

$(\tilde{H}_3)$ $S \subset (D(\phi) \cap V) \times H \subset H \times H$ is closed and convex. $\tilde{R} - S$ has finite codimensionality.

The results of Theorem 3.1 remain true if we use $(\tilde{H}_3)$ instead of (H_3).

Indeed, by (3.10) and (3.22), we have

$$\begin{aligned}&\lambda_\varepsilon \left[\int_0^T \langle \Delta g_\varepsilon(t, y_\varepsilon), z_\varepsilon\rangle dt + \int_0^T \langle \Delta h_\varepsilon(u_\varepsilon), v\rangle dt\right]\\ &+\langle a_\varepsilon, \eta - x_0 + y^*(0)\rangle + \langle b_\varepsilon, z^\varepsilon(T) - x_1 + y^*(T)\rangle\\ &\geq q_2(\varepsilon, v, \eta),\end{aligned}$$

where $q_2(\varepsilon, v, \eta) = q(\varepsilon, v, \eta) - q_1(\varepsilon)$.

Thus

$$\begin{aligned}&\lambda_\varepsilon \left[\int_0^T \langle \Delta g_\varepsilon(t, y_\varepsilon), z_\varepsilon\rangle + \int_0^T \langle \Delta h_\varepsilon(u_\varepsilon), v\rangle dt\right]\\ &+\langle a_\varepsilon, \eta - x_0 + y^*(0)\rangle + \langle b_\varepsilon, z(T) - x_1 + y^*(T)\rangle\\ &\geq q_2(\varepsilon, v, \eta) + \langle b_\varepsilon, z(T) - z^\varepsilon(T)\rangle\\ &\equiv \tilde{q}(\varepsilon, v, \eta),\end{aligned}$$

where $\tilde{q}(\varepsilon, v, \eta) \to 0$ uniformly in $\eta \in B_1(0), v \in \tilde{B}_r(0)$, as $\varepsilon \to 0$.

Then $(\tilde{H}_3)$ guarantees $(\lambda_0, a, b) \neq 0$ (cf. [4]).

At last, we give an example as follows.

Example 1. Let β be a locally Lipschitz continuous monotone graph on R and satisfy the growth condition (4.1), $A = -\Delta$, $V = H_0^1(\Omega)$, and $S = \{y_0\} \times Q_2$ with $y_0 \in H_0^1(\Omega)$, and $Q_2 \subset H$ has finite codimension.

Then problem (P) reduces to the problem as follows:

(P_1) Min $L(y, u)$
subject to all
$(y, u) \in W^{1,2}([0, T]; H) \cap C([0, T]; H) \cap L^2(0, T; D(A_H)) \times L^2(0, T; U)$
satisfying the parabolic equation

$$\begin{cases} y_t - \Delta y + \beta(y) = Bu & \text{a.e. } Q,\\ y(x, t) = 0 & \text{in } \partial\Omega,\\ y(x, 0) = y_0(x), y(x, T) \in Q_2. \end{cases}$$

By Remark 4.1 and Remark 4.2, we may apply the theorem to problem (P_1) to get the necessary conditions of the pair for the problem (P_1).

REFERENCES

[1] V. Barbu, *Optimal Control of Variational Inequalities*, Res. Notes Math. 100, Pitman (Advanced Publishing Program), Boston, MA, London, UK, 1984.

[2] Z. Cai, N. Pavel, and S. L. Wen, *Optimal control of some partial differential equations with two-point boundary conditions*, in Optimal Control of Differential Equations (Athens, OH, 1993), Lecture Notes in Pure and Appl. Math. 160, Dekker, New York, 1994, pp. 49–68.

[3] F.H. Clarke, *Optimization and Nonsmooth Analysis*, John Wiley and Sons, New York, 1983.

[4] X. Li and J. Yong, *Optimal Control Theory for Infinite-Dimensional Systems*, in Systems Control Found. Appl., Birkhäuser Boston, Boston, MA, 1995.

[5] X. Li and J. Yong, *Necessary conditions for optimal control of distributed parameter systems*, SIAM J. Control Optim., 29 (1991), pp. 895–908.

[6] N. Pavel, *Nonlinear Evolution Operators and Semigroups. Applications to Partial Differential Equations,* Lecture Notes in Math. 1260, Springer-Verlag, Berlin, New York, 1987.

SIAM J. CONTROL OPTIM.
Vol. 38, No. 6, pp. 1655–1682

DELAY SENSITIVITY OF QUADRATIC CONTROLLERS: A SINGULAR PERTURBATION APPROACH*

PAPA MOMAR NDIAYE† AND MICHEL SORINE†

Abstract. We study the variations of the quadratic performance associated to a linear differential system of retarded type for small values of the delays. From an interpretation of delays as singular perturbations of abstract evolution operators, we revisit the usual theory of representation and optimal control of retarded systems. This leads to a new parameterization of associated Riccati operators for which insight is gained in the dependence on the delays. This explicit parameterization of Riccati operators by the delays enables us to prove differentiability at zero for performance viewed as a function of the delays, in the LQ-optimal or $\mathcal{H}_\infty$ suboptimal control. In each case, the gradient is explicitly computed in terms of the nonnegative solution of the finite dimensional Riccati equation associated to the nondelay control problem. A thorough treatment is stated for the linear quadratic optimal case, and the $\mathcal{H}_\infty$ suboptimal control is presented as an application.

Key words. optimal control, delay, sensitivity, singular perturbation, robust control

AMS subject classifications. 93C25, 49K35, 49K40, 93C73, 93B36

PII. S0363012997329858

1. Introduction.

Small time delay often appears as a side effect of network control of physical systems. In the specific case where the sensors, actuators, and processors share a single channel, the true open loop system has delayed inputs and outputs and possibly state delays resulting from delayed output feedback. In view of real time control, such delays are small but strongly dependent on ordering decisions, and it would be desirable for a given control law to be able to have an a priori estimate of the sensitivity of the performance of the controlled system to small delays.

This paper mainly addresses the issue of the variations of the optimal value of a quadratic cost associated to a linear system perturbed by small delays in its inputs and state. Therefore, given some delay-vectors $\vec{k} = (k_1, \dots, k_I)$, $\vec{h} = (h_1, \dots, h_J)$ with $(k_i, h_j) \in [0, K] \times [0, H]$ for K, $H > 0$, we consider the following equation:

$$\text{(1.1)} \qquad \begin{cases} \dot{x}(t) = A_0 x(t) + \displaystyle\sum_{i=1}^{I} A_i x(t-k_i) + B_0 u(t) + \sum_{j=1}^{J} B_j u(t-h_j) \\ \text{almost everywhere (a.e.) } t \in [0,T], \\ x(0) = x_{00} \in \mathbb{R}^n, \quad x = x_0 \text{ a.e. } t \in [-K, 0], \quad x_0 \in L^2(-K, 0; \mathbb{R}^m), \\ u \in L^2(-H, T; \mathbb{R}^m), \quad u = u_0 \text{ a.e. } t \in [-H, 0], \end{cases}$$

where $A_i \in \mathcal{L}(\mathbb{R}^n)$ and $B_j \in \mathcal{L}(\mathbb{R}^m, \mathbb{R}^n)$ for $i \in \{0, \dots, I\}$ and $j \in \{0, \dots, J\}$. Also, to the solution $x(.) = x(.; \vec{k}; \vec{h}; u)$ of (1.1), we associate the following quadratic cost:

$$\text{(1.2)} \qquad J_T(x) = \frac{1}{2}\langle x(T), G(T)x(T)\rangle_{\mathbb{R}^n} + \frac{1}{2}\int_0^T \{\|Cx\|^2_{\mathbb{R}^p} + \langle u, Ru\rangle_{\mathbb{R}^m}\}dt,$$

where $G(T) \in \mathcal{L}(\mathbb{R}^n)$ is nonnegative for $T \in \mathbb{R}_+$ and $G(\infty) = 0$; $C \in \mathcal{L}(\mathbb{R}^n, \mathbb{R}^p)$; R

*Received by the editors November 17, 1997; accepted for publication (in revised form) October 18, 1999; published electronically June 15, 2000.
http://www.siam.org/journals/sicon/38-6/32985.html
†INRIA Rocquencourt, BP 105 Domaine de Voluceau, F-78153 Le Chesnay Cedex, France (papa.ndiaye@inria.fr, michel.sorine@inria.fr).

positive. Moreover, by $\widehat{J}_T(\vec{k},\vec{h})$ we denote the infimum

$$\min_{u\in L^2(0,T;\mathbb{R}^m)} J_T(x(.;\vec{k};\vec{h};u))$$

for $T \in \mathbb{R}_+ \cup \{+\infty\}$. Then the central problem of interest is the computation of the gradient of the function $(\vec{k},\vec{h}) \longrightarrow \widehat{J}_T(\vec{k},\vec{h})$ at the point $(\vec{k},\vec{h}) = (0,0)$.

We recall that, in a more general framework of systems of evolution with small changes in some parameter, Pritchard [20] has shown that there exists some constant $0 < c < 1$, depending on the nondelay optimal feedback and such that $c\widehat{J}_T(0,0) \leq \widehat{J}_T(\vec{k},0)$, whereas Dontchev [13, 14, 15], with singular perturbation techniques, computes $|\widehat{J}_T(\vec{k},0) - \widehat{J}_T(0,0)|$ in terms of the multipliers occurring in the optimality condition. More recently, for a single state delay case, Clarke and Wolenski [3, Theorem 3.2] prove the differentiability of the optimum of some associated performance and provide a characterization of the derivative in the multipliers occurring in the optimality condition. In this paper, we propose a simplification of these results in the case of quadratic minimization, and extend them to the optimum of the min-max problem associated to $\mathcal{H}_\infty$ disturbance attenuation via state feedback controller. We prove the existence of the partial derivative at zero under weaker conditions than those of Clarke and Wolenski [3]. Moreover, and this is our major contribution, we provide an explicit gradient formula for the quadratic optimal cost with partial derivatives at zero, simply expressed in terms of the solution of the finite dimensional Riccati equation solving the nondelay optimal control problem, for both finite and infinite horizons. We also state a similar result for the $\mathcal{H}_\infty$ robust performance.

The keynote of this paper is an interpretation of the delays as singular perturbations of some evolution operators that leads to an unusual variant of classical state space representations via semigroup techniques. For the representation of solutions, we refer to Ichikawa [16], Delfour and Karrakchou [9, 10], and Delfour [8]. Then using a well-known compensation trick—see, e.g., Kokotović and Yackel [18] for the finite dimensional case—we parameterize Riccati operators with the delays when revisiting optimal control of delay systems in the Pritchard and Salamon class [4, 21, 22]. In view of that parameterization, the problem of the sensitivity analysis of the optimal performance becomes a study, in the neighborhood of zero, of continuity of the components of some four-block decomposition associated to Riccati operators.

Section 3 introduces the proposed state space representation and defines the framework in which the control problem will be solved. Next, linear quadratic optimal control theory for retarded equations is revisited in section 4. New regularity theorems for Riccati operators—some weak differentiability-like results—are presented in section 5. We believe that these are of independent interest, so we state them as separate theorems. They make the proof of sensitivity results given in section 6 rather elementary. Finally, in section 7 we present the $\mathcal{H}_\infty$ case.

2. Main results.

THEOREM 2.1. *Let us assume $T < \infty$. Then*

(i) *with u_0 continuous on the left side of 0, we have*

$$\frac{\partial \widehat{J_T}}{\partial h_j}(0,0) = \langle x_{00},\ \Delta_j(P_T)x_{00}\rangle_{\mathbb{R}^n}, \tag{2.1}$$

where $\Delta_j(P_T) = P_T(0)B_jR^{-1}B^TP_T(0)$ for P_T, the nonnegative symmetric

solution of the Riccati equation

$$\frac{dP_T}{dt} + A^T P_T + P_T A - P_T B R^{-1} B^T P_T + C^T C = 0 \text{ in } C(0,T;\mathcal{L}(\mathbb{R}^n)),$$
$$P_T(T) = G(T). \tag{2.2}$$

(ii) *With x_0 continuous on the left side of 0, we have*

$$\frac{\partial \widehat{J_T}}{\partial k_i}(0,0) = -\langle P_T x_{00},\ A_i x_{00}\rangle_{\mathbb{R}^n}. \tag{2.3}$$ □

THEOREM 2.2. *Assume the triple (A,B,C) stabilizable and detectable. Then*

(i) *with u_0 continuous on the left side of 0, we have*

$$\frac{\partial \widehat{J_\infty}}{\partial h_j}(0,0) = \langle x_{00},\ \Delta_j(P_\infty) x_{00}\rangle_{\mathbb{R}^n}, \tag{2.4}$$

where $\Delta_j(P_\infty) = P B_j R^{-1} B^T P_\infty$ for P_∞, the nonnegative symmetric solution of the Riccati equation

$$A^T P_\infty + P_\infty A - P_\infty B R^{-1} B^T P_\infty + C^T C = 0 \text{ in } \mathcal{L}(\mathbb{R}^n). \tag{2.5}$$

(ii) *with x_0 continuous on the left side of 0, we have*

$$\frac{\partial \widehat{J_\infty}}{\partial k_i}(0,0) = -\langle P_\infty x_{00},\ A_i x_{00}\rangle_{\mathbb{R}^n}. \tag{2.6}$$ □

Remark 2.1. For the existence of partial derivatives with respect to state delays, we need only continuity of x_0 on the left side of 0, a condition which is weaker than the existence of $\|\dot{x}_0\|_\infty$ as required in Clarke and Wolenski [3, Theorem 3.2]. □

Example 2.1. *Sensitivity to small input delay.* If we take $J = 1$ and assume $B_0 = 0$, then $\Delta_1(P_\infty) = P_\infty B_1 R^{-1} B_1^T P_\infty = K_{opt} R K_{opt} \geq 0$ where the matrix $K_{opt} = -R^{-1} B_1^T P_\infty$ is the optimal gain for $h = 0$. Therefore formula (2.4) becomes

$$\frac{\partial \widehat{J_\infty}}{\partial h}(0,0) = \langle K_{opt} x_{00},\ R K_{opt} x_{00}\rangle_{\mathbb{R}^n} \geq 0, \tag{2.7}$$

from which we have $\widehat{J_T}(0,0) \leq \widehat{J_T}(0,h)$ for a sufficiently small h. Moreover, we see that the degradation of the optimal performance increases with the optimal gain. □

3. Singular perturbations and product-space approach for representation of solutions. First of all, we point out that for computation of the partial derivatives at the point $(k,h) = (0,0)$, we need only to compare optimal cost with solely one state or input delay to nondelay optimal cost. So it is sufficient to restrict further analysis to the cases $(I = 0; J = 1)$ and $(I = 1; J = 0)$ which in fact will be studied separately. By doing so, we do not lose any generality while we simplify statements of intermediate steps and proofs. Henceforth, we are justified in assuming that the system of interest is

$$\begin{cases} \dot{x}(t) = A_0 x(t) + A_1 x(t-k) + B_0 u(t) + B_1 u(t-h) \text{ a.e. } t \in [0,T], \\ x(0) = x_{00} \in \mathbb{R}^n,\ x = x_0 \text{ a.e. } t \in [-k,0], \quad x_0 \in L^2(-k,0;\mathbb{R}^n), \\ u \in L^2(-h,T;\mathbb{R}^m), \quad u = u_0 \text{ a.e. } t \in [-h,0]. \end{cases} \tag{3.1}$$

So, in the following, with $A = A_0 + A_1$ and $B = B_0 + B_1$, we shall look separately for state space representations when $(k = 0,\ h > 0)$ and when $(k > 0,\ h = 0)$. These situations will be referred to as the retarded input system and retarded state system, respectively. We start with the retarded input case, since it is enough to provide a comprehensive statement of the approaches that will be taken in this paper.

3.1. Retarded input system: $(k = 0,\ h > 0)$. First recall that in this case, the solution of (3.1) is given in the space $H^1_{loc}(0,\infty;\mathbb{R}^n)$ by the variation of the constants formula

$$x(t) = e^{At}x_{00} + \int_0^t e^{A(t-s)}\{B_0u(s) + B_1u(s-h)\}ds. \tag{3.2}$$

But for controller synthesis, it is more convenient to have a description of this solution with a variation of the constants formula associated to a nondelay evolution system. A common way is to follow the choice originated by Ichikawa [16] which consists in choosing, as a support for description, the extended state defined by the pair $(x(t), v(t,.))' \in \mathcal{H}^m_h = \mathbb{R}^n \times L^2(-h,0;\mathbb{R}^m)$ with the segment function v given by $v(t,\theta) = u(t+\theta)$ for a.e. $t \in [0,T]$ and $\theta \in [-h,0]$. Such a choice leads to an equivalent linear system of evolution in $\mathcal{H}^m_h$ with unbounded input but bounded solutions as explained in [16] (see also Bensoussan et al. [1]). That line is well known to provide both optimal feedback and cost.

Now we point out that it is nothing but the presence of the delay that has brought the system from finite to infinite dimensions, so the delay may be interpreted as a singular perturbation in a convenient framework. Moreover, in view of sensitivity analysis, it is desirable for simplicity to have a non-delay-dependent state space. A natural way to do this is to rescale the variable θ, in order to choose $\mathcal{H}^m_1 = \mathbb{R}^n \times L^2(-1,0;\mathbb{R}^m)$ as state space. The proposed rescaling will transform our problem with varying space to one with evolution operators singularly parameterized by the delay.

Introduce the scaled functions u^h and φ^h, defined in $L^2(0,T;L^2(-1,0;\mathbb{R}^m))$ and $L^2(-1,0;\mathbb{R}^m)$, respectively, by $u^h(t,\sigma) = v(t,\sigma h)$ and $\varphi^h(\sigma) = u_0(\sigma h)$ for a.e. $t \in [0,T]$ and $\sigma \in [-1,0]$. Then observing that $h\frac{\partial u^h}{\partial t} = \frac{\partial u^h}{\partial \sigma}$ in $L^2(0,T;L^2(-1,0;\mathbb{R}^m))$ with $u^h(0,.) = \varphi^h \in L^2(-1,0;\mathbb{R}^m)$ and $u^h(.,0) = u \in L^2(0,T;\mathbb{R}^m)$, the use of standard transposition techniques yields

$$\left\{\begin{array}{lcl} h\dfrac{\partial u^h}{\partial t} & = & Du^h + \delta^*_{\sigma=0}u \quad \text{in } L^2(0,T;H^1(-1,0;\mathbb{R}^m)') \\ u^h(0,.) & = & \varphi^h(.) \in L^2(-1,0;\mathbb{R}^m), \end{array}\right. \tag{3.3}$$

where $D = \frac{\partial}{\partial\sigma}$ with domain $\mathcal{D}(D) = \{u \in H^1(-1,0;\mathbb{R}^m),\ u(0) = 0\}$.

Then, using the shorthand $\mathcal{H}^m = \mathcal{H}^m_1$, setting $X_h(t,.) = \binom{x(t)}{u^h(t,.)} \in \mathcal{H}^m$, and $\mathcal{W} = \mathbb{R}^n \times H^1(-1,0;\mathbb{R}^m)$, we get, for any $h > 0$, the following equation of evolution:

$$\left\{\begin{array}{lcl} \dfrac{\partial X_h}{\partial t} & = & \mathcal{A}_hX_h + \mathcal{B}_hu \quad \text{in } L^2(0,T;\mathcal{W}^{m'}), \\ X_h(0,.) & = & \begin{pmatrix} x_{00} \\ \varphi^h \end{pmatrix} \in \mathcal{H}^m, \end{array}\right. \tag{3.4}$$

where

$$\mathcal{A}_h = \begin{pmatrix} A & B_1\delta_{\sigma=-1} \\ 0 & \dfrac{1}{h}D \end{pmatrix}, \quad \mathcal{D}(\mathcal{A}_h) = \mathbb{R}^n \times \mathcal{D}(D), \quad \mathcal{B}_h = \begin{pmatrix} B_0 \\ \dfrac{1}{h}\delta^*_{\sigma=0} \end{pmatrix} \in \mathcal{L}(\mathbb{R}^m, \mathcal{W}^{m'}).$$

Note that if we have chosen $\mathcal{H}_H^m$ with $H \geq h$ as initial state space as it is done sometimes in the literature, then the rescaling will lead to $\mathcal{H}_{\frac{H}{h}}^m$ instead of $\mathcal{H}_1^m$.

Now it clear that $h = 0$ is a singularity of operators $\mathcal{A}_h$ and $\mathcal{B}_h$. The next step is the well-posedness of that equation. At this stage, observe that if—by construction—the extended state space is $\mathcal{H}^m$, then the input operator $\mathcal{B}_h$ is unbounded in $\mathcal{H}^m$. Further insight into the characteristics of the operator $\mathcal{A}_h$ will show that despite that unboundedness, the solution of (3.4) will remain inside $\mathcal{H}^m$.

PROPOSITION 3.1. *Henceforth let the space $\mathcal{H}^m$ be identified with its dual and take it as a pivot space. Then we have the following:*

(i) *$\mathcal{A}_h$ is the infinitesimal generator of a C_0-semigroup $\mathcal{S}_h(t)$ on $\mathcal{H}^m$. In addition, the $\mathcal{H}^m$-adjoint of the operator $\mathcal{A}_h$ is given with $\mathcal{D}(\mathcal{A}_h^*) = \{\binom{y}{w} \in \mathcal{W}^m,\ w(-1) = hB_1^T y\}$ by*

$$\mathcal{A}_h^* \begin{pmatrix} y \\ w \end{pmatrix} = \begin{pmatrix} A_0^T y \\ -\dfrac{1}{h} Dw \end{pmatrix}.$$

(ii) *Furthermore, for $t \geq 0$, the restriction to $\mathcal{D}(\mathcal{A}_h^*)$ of $\mathcal{S}_h^*$, the adjoint of the semigroup $\mathcal{S}_h$, defines a C_0-semigroup on $\mathcal{H}^m$ and then $\mathcal{S}_h(t)$ may be extended to a C_0-semigroup on the Hilbert space $\mathcal{D}(\mathcal{A}_h^*)'$.* □

That means that the operator $\mathcal{A}_h$ is the infinitesimal generator of a C_0-semigroup $\mathcal{S}_h(t)$ on Hilbert spaces ordered by dense and continuous injections, namely $\mathcal{D}(\mathcal{A}_h) \hookrightarrow \mathcal{H}^m \hookrightarrow \mathcal{D}(\mathcal{A}_h^*)'$. Therefore the weak solution of (3.4) may be defined by the variation of the constants formula

$$X_h(t,.) = \mathcal{S}_h(t) \begin{pmatrix} x_{00} \\ \varphi^h(.) \end{pmatrix} + \int_0^t \mathcal{S}_h(t-s)\mathcal{B}_h u(s) ds, \tag{3.5}$$

where, at first sight, it seems that we only have $X_h \in C(0,T;\mathcal{D}(\mathcal{A}_h^*)')$. But in fact, due to the structure of the semigroup $\mathcal{S}_h$, we have additional regularity in view of the following property of the convolution term.

LEMMA 3.2. *The operator $u \longmapsto \{t \to \int_0^t \mathcal{S}_h(t-s)\mathcal{B}_h u(s) ds\}$ is linear and continuous from $L^2(0,T;\mathbb{R}^m)$ to $C(0,T;\mathcal{H}^m)$.* □

The proofs of Proposition 3.1 and Lemma 3.2 are given in the appendix. Their combination is useful towards a complete characterization of the introduced infinite dimensional state representation which ultimately satisfies the following.

THEOREM 3.3.

(i) *The function X_h given by formula* (3.5) *is the unique solution in*

$$\left\{ X \in C(0,T;\mathcal{H}^m) \ :\ \frac{dX}{dt} \in L^2(0,T;\mathcal{D}(\mathcal{A}_h^*)') \right\} \tag{3.6}$$

of the weak equation

$$\begin{cases} \dfrac{dX}{dt} &= \mathcal{A}_h X + \mathcal{B}_h u \quad \text{in } L^2(0,T;\mathcal{D}(\mathcal{A}_h^*)'), \\ X(0) &= \begin{pmatrix} x_{00} \\ \varphi^h \end{pmatrix} \in \mathcal{H}^m. \end{cases} \tag{3.7}$$

In addition, there exists a constant $c > 0$ such that

$$\|X_h\|_{C(0,T;\mathcal{H}^m)} + \left\| \frac{\partial X_h}{\partial t} \right\|_{L^2(0,T;\mathcal{D}(\mathcal{A}_h^*)')} \leq c \left[\left\| \begin{pmatrix} x_{00} \\ \varphi^h \end{pmatrix} \right\|_{\mathcal{H}^m} + \|u\|_{L^2(0,T;\mathbb{R}^m)} \right]. \tag{3.8}$$

(ii) *Moreover, the first component of this weak solution X_h is equal to x, given by the formula* (3.2)*, which is the unique solution of the retarded equation* (3.1). □

The derivation of this theorem is immediate. The first item comes from the lemma, when using, e.g., the reference [1, Theorem 3.1, p. 173]. The second one is the outcome of a detailed rewriting of formula (3.5).

Remark 3.1. It is worth recalling that the property $X_h \in C(0,T;\mathcal{H}^m)$ follows easily from continuity of the translation in L^2 combined with the estimation

$$\|x\|_{W^{1,2}(0,T;\mathbb{R}^n)} \leq C\left[|x_{00}|_{\mathbb{R}^n} + \|x_0\|_{L^2(-K,0;\mathbb{R}^n)} + \|u\|_{L^p(-H,T;\mathbb{R}^m)}\right] \tag{3.9}$$

which can be found, e.g., in [1]. So the interest of the relation $\frac{\partial X_h}{\partial t} \in L^2(0,T;\mathcal{D}(\mathcal{A}_h^*)')$ is in additional information on the segment function u^h, which satisfies equation $\frac{\partial u^h}{\partial t} = h\frac{\partial u^h}{\partial \sigma}$ with $u^h(t,0) = u(t)$, and belongs to $L^2(0,T;H^1(-1,0;\mathbb{R}^m)')$. □

3.2. Retarded state system: ($k > 0$, $h = 0$). In this case the solution is given in $H^1_{loc}(0,\infty;\mathbb{R}^n)$ by the variation of the constants formula

$$x(t) = e^{A_0 t}x_{00} + \int_0^t e^{A_0(t-s)}A_1 x(s-k)ds + \int_0^t e^{A_0(t-s)}Bu(s)ds. \tag{3.10}$$

It is well known that this solution may be described in a product-space framework by means of the pair $(x(t), z(t,.))'$ with $z(t,\theta) = x(t+\theta)$, $t \geq 0$, $\theta \in [-k,0]$. Moreover, when considering only the homogeneous part of the retarded equation, this pair is generated by a C_0-semigroup acting on the extended initial condtion $(x_{00}, x_0)'$. See, e.g., Delfour [8] or Staffans [25, 26] for further details.

To emphasize the singular perturbation effect of the delay, we rescale the segment function z before following the product-space lines for state representation. To this end, we introduce the functions defined by $x^k(t,\sigma) = x(t+\sigma k)$ for $t \geq 0$ and $\sigma \in [-1,0]$, and $\psi^k(\sigma) = x_0(\sigma k)$. Then the solution of the homogeneous part of the retarded equation may be described from $(x_{00}, \psi^k)'$ with some linear operator given in $\mathcal{H}^n = \mathbb{R}^n \times L^2(-1,0;\mathbb{R}^n)$ by

$$\mathcal{S}_k(t)\begin{pmatrix} x_{00} \\ \psi^k(.) \end{pmatrix} = \begin{pmatrix} x(t) \\ x^k(t,.) \end{pmatrix} \ \forall\ t \geq 0 \ \text{ and } \ \begin{pmatrix} x_{00} \\ \psi^k \end{pmatrix} \in \mathcal{H}^n. \tag{3.11}$$

This next statement is a characterization of $\mathcal{S}_k$ that comes from the usual properties of the extended C_0-semigroup associated to a delay equation of retarded type [1, p. 61].

PROPOSITION 3.4. *$\mathcal{S}_k$ is a C_0-semigroup on $\mathcal{H}^n$ which is generated by the operator defined on $\mathcal{H}^n$ with domain $\mathcal{D}(\mathcal{A}_k) = \{\binom{x}{y} \in \mathcal{W}^n = \mathbb{R}^n \times H^1(-1,0;\mathbb{R}^n),\ y(0) = x\}$ as*

$$\mathcal{A}_k = \begin{pmatrix} A_0 & A_1\delta_{\sigma=-1} \\ 0 & \dfrac{1}{k}\dfrac{\partial}{\partial\sigma} \end{pmatrix}. \qquad \square$$

Thus setting $X_k = \binom{x(t)}{x^k(t,.)}$ and $\mathcal{B}_k = \binom{B}{0} \in \mathcal{L}(\mathbb{R}^m, \mathcal{H}^n)$, we have the following evolution equation

$$\begin{cases} \dfrac{\partial X_k}{\partial t} &= \mathcal{A}_k X_k + \mathcal{B}_k u \quad \text{in } L^2(0,T;\mathcal{H}^n), \\ X_k(0,.) &= \begin{pmatrix} x_{00} \\ \psi^k(.) \end{pmatrix} \in \mathcal{H}^n. \end{cases} \tag{3.12}$$

Now we have a bounded equation of evolution in the state space $\mathcal{H}^n$. It is obvious that for any initial condition in $\mathcal{H}^n$, its solution satisfies $X_k \in C(0,T;\mathcal{H}^n)$. Moreover, (3.12) may be extended by means of the method of transposition. In addition, easy computations lead to the following characterization of the adjoint of $\mathcal{A}_k$ that will be useful here.

LEMMA 3.5. *The $\mathcal{H}^n$-adjoint of the operator $\mathcal{A}_k$ is given with domain $\mathcal{D}(\mathcal{A}_k^*) = \{\binom{v}{w} \in \mathcal{W}^n,\ w(-1) = kA_1^T v\}$ as*

$$\mathcal{A}_k^* \begin{pmatrix} v \\ w \end{pmatrix} = \begin{pmatrix} A_0^T v + \dfrac{1}{k}\delta_{\sigma=0} w \\ -\dfrac{1}{k}\dfrac{\partial}{\partial \sigma} w \end{pmatrix}. \qquad \square$$

Then we have the following theorem.

THEOREM 3.6. *Given any $t \in [0,T]$, $X_0 \in \mathcal{H}^n$, and $u \in L^2(0,T;\mathbb{R}^m)$,*

$$X(t) = \mathcal{S}_k(t)X_0 + \int_0^t \mathcal{S}_k(t-s)\mathcal{B}_k u(s)ds \ \in \mathcal{H}^n \tag{3.13}$$

is the unique solution in the space

$$\left\{ X \in C(0,T;\mathcal{H}^n) : \frac{dX}{dt} \in L^2(0,T;\mathcal{D}(\mathcal{A}_k^*)') \right\} \tag{3.14}$$

of the transposed adjoint equation

$$\begin{cases} \dfrac{dX}{dt} &= \widetilde{\mathcal{A}}_k X + \mathcal{B}_k u \quad \text{in } L^2(0,T;\mathcal{D}(\mathcal{A}_k^*)'), \\ X(0) &= X_0 \in \mathcal{H}^n, \end{cases} \tag{3.15}$$

where

$$\widetilde{\mathcal{A}}_k \in \mathcal{L}(\mathcal{D}(\mathcal{A}_k),\mathcal{H}^n) \ \cap \ \mathcal{L}(\mathcal{W}^n,\mathcal{W}^{n'}) \ \cap \ \mathcal{L}(\mathcal{H}^n,\mathcal{D}(\mathcal{A}_k^*)') \tag{3.16}$$

is the extension of $\mathcal{A}_k$ given on $\mathcal{W}^n$ by

$$\widetilde{\mathcal{A}}_k = \begin{pmatrix} A_0 & A_1\delta_{\sigma=-1} \\ \dfrac{1}{k}\delta^*_{\sigma=0} & \dfrac{1}{k}D \end{pmatrix}. \tag{3.17}$$

Extension $\widetilde{\mathcal{A}}_k$ generates $\mathcal{S}_k$, henceforth considered as a C_0-semigroup on any of the three spaces $\mathcal{D}(\mathcal{A}_k) \hookrightarrow \mathcal{H}^n \hookrightarrow \mathcal{D}(\mathcal{A}_k^*)'$ which are ordered by continuous and dense injections. Specifically, for $X_0 = \binom{x_{00}}{\psi^k}$, X coincides with the augmented state given from x by $\binom{x}{x^k}$. $\square$

Proof. Theorem 3.6 is given in Appendix A.3.

Remark 3.2. Since $\mathcal{B}_k u \in L^2(0,T;\mathcal{H}^n) \Longrightarrow \{t \to \int_0^t \mathcal{S}_k(t-s)\mathcal{B}_k u(s)ds\} \in L^2(0,T;\mathcal{D}(\mathcal{A}_k))$ and in view of [1, Proposition 3.2], the different concepts of solution (strong, mild, and weak) coincide for the equation of evolution (3.12), so we just say "the solution." Furthermore, we may notice that

$$X_0 \in \mathcal{D}(\mathcal{A}_k) \Longrightarrow X_k \in L^2(0,T;\mathcal{D}(\mathcal{A}_k)), \tag{3.18}$$

$$X_0 \in \mathcal{W}^n \Longrightarrow X_k \in L^2(0,T;\mathcal{W}^n). \qquad \square \tag{3.19}$$

4. Linear quadratic optimal control revisited. To begin with, we state in the first subsection the solution of the linear quadratic minimization problem as a special case of the general result for systems belonging to the abstract class of Pritchard and Salamon [21]. The interest in such a setting is that it notably simplifies the form of well-known associated Riccati equations that can be found in many places in the literature. Without attempting to give a complete overview of references on existence and representation of the solution of the Riccati equation for delay systems, we mention the works of Bensoussan et al. [1], Delfour and Karrakchou [9, 10], Pritchard and Salamon [21, 22], Salamon [23, 24], Staffans [27], and Vinter and Kwong [30]. The second subsection introduces, for these infinite dimensional Riccati operators, a parameterization by the delays which corresponds to a singular perturbation analysis as proposed by Kokotović and Yackel [18] in finite dimensions. Thanks to such a parameterization, the question of the dependance of the solution of the Riccati equation with respect to the delays just turns to a problem of existence of weak limits for some appropriate operators.

Before going further, we recall that the Pritchard and Salamon class is a class of systems satisfying the regularity conditions stated in the following definition [4].

DEFINITION 4.1. *Consider W, V, U, and K four real and separable Hilbert spaces such that $W \hookrightarrow V$, and consider $\mathcal{S}(t)$ a C_0-semigroup on V and W. Then the linear system $(\mathcal{S},\mathcal{B},\mathcal{C})$*

$$\left\{\begin{array}{rcl} x(t) &=& \mathcal{S}(t)x_0 + \displaystyle\int_0^t \mathcal{S}(t-s)\mathcal{B}u(s)ds, \quad t \geq 0, \\ y(t) &=& \mathcal{C}x(t), \quad (x_0,u) \in V \times L^2_{loc}(0,\infty;U) \end{array}\right. \tag{4.1}$$

is said to belong to the abstract Pritchard and *Salamon class for W and V with respect to $\mathcal{S}$—in short $(\mathcal{S},\mathcal{B},\mathcal{C}) \in \mathcal{C}_{PS}(W,V,U,K)$—if we have the following.*

(1) *$\mathcal{B} \in \mathcal{L}(U,V)$ is such that there exists t and $b > 0$ satisfying*

$$\int_0^t \mathcal{S}(t-s)\mathcal{B}u(s)ds \in W \quad \text{and} \quad \| \int_0^t \mathcal{S}_h(t-s)\mathcal{B}u(s)ds\|_W \leq b\|u\|_{L^2(0,t;U)}.$$

(2) *$\mathcal{C} \in \mathcal{L}(W,K)$ is such that there exists t and $c > 0$ satisfying*

$$\|\mathcal{C}\mathcal{S}(.)x\|_{L^2(0,t;K)} \leq c\|x\|_V \quad \forall x \in W.$$

If, moreover, $\mathcal{D}_V(\mathcal{A}) \hookrightarrow W$, then $(\mathcal{S},\mathcal{B},\mathcal{C})$ is said to be regular. ∎

4.1. Solution of the control problem. We have the following.

PROPOSITION 4.2.

(i) *Denote the operator $(C,0) \in \mathcal{L}(\mathcal{H}^m,\mathbb{R}^p)$ as $\mathcal{C}_h$ and assume $h > 0$. Then the triple $(\mathcal{S}_h,\mathcal{B}_h,\mathcal{C}_h) \in \mathcal{C}_{PS}(\mathcal{H}^m,\mathcal{D}(\mathcal{A}_h^*)',\mathbb{R}^m,\mathbb{R}^p)$ and is regular.*

(ii) *Denote the operator $(C,0) \in \mathcal{L}(\mathcal{H}^n,\mathbb{R}^p)$ as $\mathcal{C}_k$ and assume $k > 0$. Then the triple $(\mathcal{S}_k,\mathcal{B}_k,\mathcal{C}_k) \in \mathcal{C}_{PS}(\mathcal{H}^n,\mathcal{D}(\mathcal{A}_k^*)',\mathbb{R}^n,\mathbb{R}^p)$ and is regular.*

Then, with the notation

$$\sigma_s^+[V] = \{P \in \mathcal{L}_s(V,V'),\ \langle P\phi,\phi\rangle_{V,V'} \geq 0\ \forall \phi \in V\}, \tag{4.2}$$

we can state the next theorem as the particular case of [21, Proposition 2.8].

THEOREM 4.3 (Finite horizon control: $\boldsymbol{T < \infty}$).

(i) *Assume $h > 0$ and denote $\left(\begin{smallmatrix} G(T) & 0 \\ 0 & 0 \end{smallmatrix}\right) \in \mathcal{L}(\mathcal{H}^m)$ as $\mathcal{G}_h(T)$. Then there exists a unique operator $\mathcal{P}_{T,h}$ such that*

$$\widehat{J_T}(0,h) = \frac{1}{2}\left\langle X_h(0,.), \mathcal{P}_{T,h}(0)X_h(0,.)\right\rangle_{\mathcal{H}^m}, \tag{4.3}$$

and the optimal control is $u_{opt}(t) = -R^{-1}\mathcal{B}_h^*\mathcal{P}_{T,h}(t)X_h(t,.)$, with X_h given by Theorem 3.3. That operator $\mathcal{P}_{T,h} \in C^1(0,T,\mathcal{L}_s(\mathcal{H}^m)) \cap C(0,T,\sigma_s^+[\mathcal{D}(\mathcal{A}_h^*)'])$ is the unique nonnegative self-adjoint solution of the Riccati equation

$$\begin{aligned} &\frac{\partial \mathcal{P}_{T,h}}{\partial t} + \mathcal{A}_h^*\mathcal{P}_{T,h} + \mathcal{P}_{T,h}\mathcal{A}_h - \mathcal{P}_{T,h}\mathcal{B}_h R^{-1}\mathcal{B}_h^*\mathcal{P}_{T,h} + \mathcal{C}_h^*\mathcal{C}_h = 0, \\ &\mathcal{P}_{T,h}(T) = \mathcal{G}_h(T). \end{aligned} \tag{4.4}$$

(ii) *Assume $k > 0$ and denote $\left(\begin{smallmatrix} G(T) & 0 \\ 0 & 0 \end{smallmatrix}\right) \in \mathcal{L}(\mathcal{H}^n)$ as $\mathcal{G}_k(T)$. Then there exists a unique operator $\mathcal{Q}_{T,k}$ such that*

$$\widehat{J_T}(k,0) = \frac{1}{2}\left\langle X_k(0,.), \mathcal{Q}_{T,k}(0)X_k(0,.)\right\rangle_{\mathcal{H}^n}, \tag{4.5}$$

and the optimal control is $u_{opt}(t) = -R^{-1}\mathcal{B}_k^*\mathcal{Q}_{T,k}(t)X_k(t,.)$, where X_k is given by Theorem 3.6. $\mathcal{Q}_{T,k} \in C^1(0,T,\mathcal{L}_s(\mathcal{H}^n)) \cap C(0,T,\sigma_s^+[\mathcal{D}(\mathcal{A}_k^*)'])$ is the unique nonnegative self-adjoint solution of the Riccati equation

$$\begin{aligned} &\frac{\partial \mathcal{Q}_{T,k}}{\partial t} + \widetilde{\mathcal{A}}_k^*\mathcal{Q}_{T,k} + \mathcal{Q}_{T,k}\widetilde{\mathcal{A}}_k - \mathcal{Q}_{T,k}(t)\mathcal{B}_k R^{-1}\mathcal{B}_k^*\mathcal{Q}_{T,k} + \mathcal{C}_k^*\mathcal{C}_k = 0, \\ &\mathcal{Q}_{T,k}(T) = \mathcal{G}_k(T). \end{aligned} \tag{4.6}$$

∎

Below, we recall the weak stabilizability and detectability assumptions required to state the infinite horizon case as an application of [21, Theorem 3.4].

DEFINITION 4.4. *Set $J_\infty(u,x,y) = \|y\|^2_{L^2(0,\infty,K)} + \|\mathcal{R}^{\frac{1}{2}}u\|^2_{L^2(0,\infty,K)}$ for some $\mathcal{R}$ positive in U and (x,u,y) satisfying (4.1). $(\mathcal{S},\mathcal{B},\mathcal{C}) \in \mathcal{C}_{PS}(W,V,U,K)$ is said to be*

(1) stabilizable *if $\forall x_0 \in V, \exists u_{x_0} \in L^2(0,\infty;\mathbb{R}^m)$ such that $J_\infty(u_{x_0},x,y) < \infty$;*
(2) detectable *if $\forall \left(\begin{smallmatrix} x_0 \\ u \end{smallmatrix}\right) \in V \times L^2(0,\infty;\mathbb{R}^m)$ such that $J_\infty(u,x,y) < \infty$, we have $x \in L^2(0,\infty;V)$).*

THEOREM 4.5 (Infinite horizon control).

(i) *Assume that $(\mathcal{S}_h,\mathcal{B}_h,\mathcal{C}_h)$ is stabilizable and detectable in the sense of Definition 4.4. Then there exists a unique operator $\mathcal{P}_{\infty,h}$ such that*

$$\widehat{J_\infty}(0,h) = \frac{1}{2}\left\langle X_h(0,.), \mathcal{P}_{\infty,h}X_h(0,.)\right\rangle_{\mathcal{H}^m} \tag{4.7}$$

and the optimal control is $u_{opt}(t) = -R^{-1}\mathcal{B}_h^*\mathcal{P}_{\infty,h}X_h(t,.)$. That $\mathcal{P}_{\infty,h} \in \sigma_s^+[\mathcal{D}(\mathcal{A}_h^*)']$ is the unique nonnegative self-adjoint solution of the Riccati equation (in $\mathcal{H}^m$)

$$\mathcal{A}_h^*\mathcal{P}_{\infty,h} + \mathcal{P}_{\infty,h}\mathcal{A}_h - \mathcal{P}_{\infty,h}\mathcal{B}_h R^{-1}\mathcal{B}_h^*\mathcal{P}_{\infty,h} + \mathcal{C}_h^*\mathcal{C}_h = 0. \tag{4.8}$$

(ii) *Assume that $(\mathcal{S}_k,\mathcal{B}_k,\mathcal{C}_k)$ is stabilizable and detectable in the sense of Definition 4.4. Then there exists a unique operator $\mathcal{Q}_{\infty,k}$ such that*

$$\widehat{J_\infty}(k,0) = \frac{1}{2}\left\langle X_k(0,.), \mathcal{Q}_{\infty,k}X_k(0,.)\right\rangle_{\mathcal{H}^n}, \tag{4.9}$$

and the optimal control is $u_{opt}(t) = -R^{-1}\mathcal{B}_k^*\mathcal{Q}_{\infty,k}X_k(t,.)$. *That* $\mathcal{Q}_{\infty,k} \in \sigma^+[\mathcal{D}(\mathcal{A}_k^*)']$ *is the unique nonnegative self-adjoint solution of the Riccati equation (in* $\mathcal{H}^n$*)*

$$\widetilde{\mathcal{A}}_k^*\mathcal{Q}_{\infty,k} + \mathcal{Q}_{\infty,k}\widetilde{\mathcal{A}}_k - \mathcal{Q}_{\infty,k}\mathcal{B}_k R^{-1}\mathcal{B}_k^*\mathcal{Q}_{\infty,k} + \mathcal{C}_k^*\mathcal{C}_k = 0. \qquad \square \tag{4.10}$$

Remark 4.1. If we define the functions $\mathcal{N}_h$ and Δ_k by $\mathcal{N}_h(\lambda) = B_0 + B_1e^{-\lambda h}$ and $\Delta_k(\lambda) = \lambda I_d - A_0 - A_1e^{\lambda k}$ for $k,\ h \geq 0$, and $\lambda \in \mathbb{C}$, then using, e.g., [22, Theorem 3.5], stabilizability and detectability in Theorem 4.5 can be reformulated as more usual rank tests for $[\Delta_\bullet(\lambda),\ \mathcal{N}_\bullet(\lambda)]$ and $[{}^{\Delta_\bullet(\lambda)}_{C}]$. Moreover, for the input delay case, a simpler test is given by Chyung [2]: Rank $\left[D\ AD\ A^2D \dots A^{n-1}D\right] = n$, with $D = B_0 + e^{-Ah}B_1$.

4.2. Parameterization and decomposition of Riccati operators. In this subsection, we propose for the previously given Riccati operators a four-block decomposition combined with a parameterization by the delays h or k. This type of parameterization was introduced by Kokotović and Yackel [18] for the Riccati matrix solving the LQ control problem of singular perturbed finite dimensional systems. Here, it makes available some regularity results on Riccati operators and this leads to a decomposition of Riccati equations which will be useful for sensitivity analysis. We restrict detailed computations and proofs to the finite horizon input delay case. The state delay and stationary cases may be derived in a similar way. Input delay results are stated in Propositions 4.7 and 4.8 and the state delay results are in Propositions 4.9 and 4.6.

PROPOSITION 4.6. *For any* $h > 0$, $t \in [0,T]$, *and* $T < \infty$, *the Riccati operator* $\mathcal{P}_{T,h}(t)$ *has in* $\mathcal{L}_s(\mathcal{H}^m)$ *the decomposition*

$$\mathcal{P}_{T,h}(t) = \begin{pmatrix} \mathcal{P}^1_{T,h}(t) & h\mathcal{P}^2_{T,h}(t) \\ h\mathcal{P}^{2*}_{T,h}(t) & h\mathcal{P}^3_{T,h}(t) \end{pmatrix} \tag{4.11}$$

with $\mathcal{P}^1_{T,h} \in C^1(0,T;\mathcal{L}_s(\mathbb{R}^n))$*; and operators* $\mathcal{P}^2_{T,h} \in C^1(0,T;\mathcal{L}_s(L^2(-1,0;\mathbb{R}^m),\mathbb{R}^n))$ *and* $\mathcal{P}^3_{T,h} \in C^1(0,T;\mathcal{L}_s(L^2(-1,0;\mathbb{R}^m))$ *having representations*

$$[\mathcal{P}^{2*}_{T,h}(t)x](\sigma) = \mathcal{P}^{2*}_{T,h}(t,\sigma)x \quad \forall (x,\sigma) \in \mathbb{R}^n \times [-1,0], \tag{4.12}$$

$$[\mathcal{P}^3_{T,h}(t)y](\sigma) = \int_{-1}^{0} \mathcal{P}^3_{T,h}(t,\sigma,r)y(r)dr \quad \forall y \in L^2(-1,0;\mathbb{R}^m) \tag{4.13}$$

that satisfy, for t *and* $r \in [0,T]$,

$$\mathcal{P}^{2*}_{T,h}(t,.) \in H^1(-1,0;\mathcal{L}(\mathbb{R}^n,\mathbb{R}^m)), \tag{4.14}$$

$$\mathcal{P}^3_{T,h}(t,.,r) \in H^1([-1,0];\mathcal{L}(\mathbb{R}^m)). \tag{4.15}$$

Moreover, $\mathcal{P}^3_{T,h}(t)$ *maps* $L^2(-1,0;\mathbb{R}^m)$ *to* $H^1(-1,0;\mathbb{R}^m)$ *and the following limiting conditions hold:*

$$\mathcal{P}^{2*}_{T,h}(t,-1) = B_1^T\mathcal{P}^1_{T,h}(t), \tag{4.16}$$

$$[\mathcal{P}^3_{T,h}(t)y](-1) = hB_1^T\mathcal{P}^2_{T,h}(t)y \quad \forall y \in L^2(-1,0;\mathbb{R}^m). \qquad \square \tag{4.17}$$

Proof. First, we point out that the existence of a four-block decomposition is a consequence of linearity of $\mathcal{P}_{T,h}(t)$ and the product-space structure of $\mathcal{H}^m = \mathbb{R}^n \times$

$L^2(-1,0,\mathbb{R}^m)$. We impose the presence of the factor h in this decomposition with the aim of simplifying the singularity which appears in $\mathcal{B}_h$ for small values of h. Then the representations of the blocks with the kernels of some continuous operator is a classical trick (see, e.g., Ichikawa [16], Delfour, McCalla, and Mitter [11], or Delfour and Mitter [12]). The only remaining point is in regard to regularity results and limiting conditions. For that, we may remember that the integral form of the Riccati operator arising in the solution of the control problem is obtained via a quasi-evolution operator (see, e.g., Curtain and Pritchard [5, 6], Pritchard and Salamon [21]) in such a way that the desired result will follow from a decomposition of the integral Riccati equation.

The quasi-evolution operator $U_h \in \mathcal{L}(\mathcal{H}^m)$ we mention is linked to $\mathcal{P}_h$ in the following way:

$$U_h(t,s)X = \mathcal{S}_h(t-s)X - \int_s^t \mathcal{S}_h(t-r)\mathcal{B}_h R^{-1}\mathcal{B}_{\vec{h}}^* \mathcal{P}_{T,h} U_{\vec{h}}(r,s)X\,dr, \tag{4.18}$$

$$\mathcal{P}_{T,h}(t)X = \mathcal{S}_h^*(T-t)\mathcal{G}_h(T)U_h(T,t)X + \int_t^T \mathcal{S}_h^*(r-t)\mathcal{C}_h^*\mathcal{C}_h U_h(r,t)X\,dr, \tag{4.19}$$

$$for\ X \in \mathcal{H}^m\ and\ 0 \le s \le t \le T.$$

Now let us decompose U_h as $U_h = \left(\begin{smallmatrix} U_{11} & U_{12} \\ U_{21} & U_{22} \end{smallmatrix}\right)$ and set $X = \left(\begin{smallmatrix} x \\ y \end{smallmatrix}\right) \in \mathcal{H}^m$. Then, since $\mathcal{G}_h = \left(\begin{smallmatrix} G & 0 \\ 0 & 0 \end{smallmatrix}\right)$ and $\mathcal{C}_h = (C,0)$, when writing $\mathcal{S}_h = \left(\begin{smallmatrix} \mathcal{S}_{11} & \mathcal{S}_{12} \\ 0 & \mathcal{S}_{22} \end{smallmatrix}\right)$, we get

$$\mathcal{S}_h^*(T-t)\mathcal{G}_h U_h(T,t)\begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} \mathcal{S}_{11}^*(T-t)G\,[\,U_{11}(T,t)x + U_{12}(T,t)y\,] \\ \mathcal{S}_{12}^*(T-t)G\,[\,U_{11}(T,t)x + U_{12}(T,t)y\,] \end{pmatrix}.$$

Therefore the decomposition of $\mathcal{P}_{T,h}$ produces the relations

$$\mathcal{P}_{T,h}^1(t)x = \mathcal{S}_{11}^*(T-t)GU_{11}(T,t)x + \int_t^T \mathcal{S}_{11}^*(r-t)C^*CU_{11}(r,t)x\,dr, \tag{4.20}$$

$$h\mathcal{P}_{T,h}^2(t)y = \mathcal{S}_{11}^*(T-t)GU_{12}(T,t)y + \int_t^T \mathcal{S}_{11}^*(r-t)C^*CU_{12}(r,t)y\,dr; \tag{4.21}$$

and, for $\sigma \in [-1,0]$,

$$h\mathcal{P}_{T,h}^{2*}(t,\sigma)x = \mathcal{S}_{12}^*(T-t)(\sigma)GU_{11}(T,t)x + \int_t^T \mathcal{S}_{12}^*(r-t)(\sigma)C^*CU_{11}(r,t)x\,dr, \tag{4.22}$$

$$h\mathcal{P}_{T,h}^3(t)y(\sigma) = \mathcal{S}_{12}^*(T-t)(\sigma)GU_{12}(T,t)y + \int_t^T \mathcal{S}_{12}^*(r-t)(\sigma)C^*CU_{12}(r,t)y\,dr. \tag{4.23}$$

But for $y \in \mathbb{R}^n$, we have $\mathcal{S}_{11}^*(t) = e^{A^T t}$ and $\mathcal{S}_{12}^*(t)y(\sigma) = h\chi_{\{t-h_j \ge 0\}} e^{-h(\sigma+1)} B_1 \mathcal{S}_{11}^*(t)y$. Then the regularities of $\mathcal{P}_{T,h}^{2*}$ and $\mathcal{P}_{T,h}^3$ as well as the limiting conditions follow from the property: $\sigma \longrightarrow \mathcal{S}_{11}^*(t-h(\sigma+1)) \in H^1(-1,0;\mathbb{R}^n)$. □

As a consequence of the Proposition 4.6, we may decompose the Riccati differential equation (4.4) in order to identify the components $\mathcal{P}_{T,h}^1$, $\mathcal{P}_{T,h}^2$ and $\mathcal{P}_{T,h}^3$. Then we get (we set $H^1 = H^1(-1,0;\mathbb{R}^m)$ and $L^2 = L^2(-1,0;\mathbb{R}^m)$)

$$\left\{\begin{array}{lcl} \langle \dot{\mathcal{P}}_{T,h}^1(t)x_1, x_2\rangle_{\mathbb{R}^n} & = & -\langle (A - B_0R^{-1}\delta_{\sigma=0}\mathcal{P}_{T,h}^{2*}(t,\sigma))x_1, \mathcal{P}_{T,h}^1(t)x_2\rangle_{\mathbb{R}^n} \\ & & -\langle \mathcal{P}_{T,h}^1(t)x_1, (A - B_0R^{-1}\delta_{\sigma=0}\mathcal{P}_{T,h}^{2*}(t,\sigma))x_2\rangle_{\mathbb{R}^n} \\ & & +\langle B_0^T\mathcal{P}_{T,h}^1(t)x_1, R^{-1}B_0^T\mathcal{P}_{T,h}^1(t)x_2\rangle_{\mathbb{R}^n} \\ & & +\langle \delta_{\sigma=0}\mathcal{P}_{T,h}^{2*}(t,\sigma)x_1, R^{-1}\delta_{\sigma=0}\mathcal{P}_{T,h}^{2*}(t,\sigma)x_2\rangle_{\mathbb{R}^m} \\ & & -\langle Cx_1, Cx_2\rangle_{\mathbb{R}^p}, \\ \mathcal{P}_{T,h}^1(T)x_1 & = & G(T)x_1 \quad \forall x_1, x_2 \in \mathbb{R}^n; \end{array}\right. \tag{4.24}$$

$$
(4.25)\quad \begin{cases} h\langle \dot{\mathcal{P}}^2_{T,h}(t)y,x\rangle_{\mathbb{R}^n} &= -\langle \mathcal{P}^1_{T,h}(t)B_1\delta_{\sigma=-1}y,x\rangle_{\mathbb{R}^n} \\ &\quad +\langle \mathcal{P}^1_{T,h}(t)B_0R^{-1}\delta_{\sigma=0}\mathcal{P}^3_{T,h}(t)y,x\rangle_{\mathbb{R}^n} \\ &\quad +\langle \delta_{\sigma=0}\mathcal{P}^3_{T,h}(t)y,R^{-1}\delta_{\sigma=0}\mathcal{P}^{2*}_{T,h}(t,\sigma)x\rangle_{\mathbb{R}^m} \\ &\quad -\langle Dy,\mathcal{P}^{2*}_{T,h}(t,.)x\rangle_{L^2}+\ h\{\langle \mathcal{P}^2_{T,h}(t)y,Ax\rangle_{\mathbb{R}^n} \\ &\quad -\langle B_0^T\mathcal{P}^2_{T,h}(t)y,R^{-1}B_0^T\mathcal{P}^1_{T,h}(t)x\rangle_{\mathbb{R}^n} \\ &\quad -\langle B_0^T\mathcal{P}^2_{T,h}(t)y,R^{-1}\delta_{\sigma=0}\mathcal{P}^{2*}_{T,h}(t,\sigma)x\rangle_{\mathbb{R}^m}\}, \\ \mathcal{P}^{2*}_{T,h}(t,-1)x &= B_1^T\mathcal{P}^1_{T,h}(t)x \quad \forall x\in\mathbb{R}^n\ \forall y\in H^1; \end{cases}
$$

$$
(4.26)\quad \begin{cases} h\langle \dot{\mathcal{P}}^3_{T,h}(t)y,z\rangle_{L^2} &= -\langle Dy,\mathcal{P}^3_{T,h}(t)z\rangle_{L^2}-\langle \mathcal{P}^3_{T,h}(t)y,Dz\rangle_{L^2} \\ &\quad +\langle \delta_{\sigma=0}\mathcal{P}^3_{T,h}(t)y,R^{-1}\delta_{\sigma=0}\mathcal{P}^3_{T,h}(t)z\rangle_{\mathbb{R}^m} \\ &\quad +h\{\langle B_0^T\mathcal{P}^2_{T,h}(t)y,R^{-1}\delta_{\sigma=0}\mathcal{P}^3_{T,h}(t)z\rangle_{\mathbb{R}^m} \\ &\quad +\langle B_0^T\mathcal{P}^2_{T,h}(t)z,R^{-1}\delta_{\sigma=0}\mathcal{P}^3_{T,h}(t)y\rangle_{\mathbb{R}^m} \\ &\quad -\langle B_1\delta_{\sigma=-1}y,\mathcal{P}^2_{T,h}(t)z\rangle_{\mathbb{R}^n} \\ &\quad -\langle \mathcal{P}^2_{T,h}(t)y,B_1\delta_{\sigma=-1}z\rangle_{\mathbb{R}^n}\} \\ &\quad +h^2\langle B_0^T\mathcal{P}^2_{T,h}(t)z,R^{-1}B_0^T\mathcal{P}^2_{T,h}(t)y\rangle_{\mathbb{R}^m}, \\ \mathcal{P}^3_{T,h}(t)y(-1) &= hB_1^T\mathcal{P}^2_{T,h}(t)y \quad \forall y,z\in H^1. \end{cases}
$$

For infinite horizon, we can derive the associated decomposition in a similar way.

PROPOSITION 4.7. *For any $h>0$, $\mathcal{P}_{\infty,h}$ may be written in $\mathcal{L}(\mathcal{H}^m)$ as*

$$
(4.27)\qquad \mathcal{P}_{\infty,h}=\begin{pmatrix} \mathcal{P}^1_{\infty,h} & h\mathcal{P}^2_{\infty,h} \\ h\mathcal{P}^{2*}_{\infty,h} & h\mathcal{P}^3_{\infty,h}\end{pmatrix},
$$

where $\mathcal{P}^1_{\infty,h}\in\mathcal{L}(\mathbb{R}^n)$; $\mathcal{P}^3_{\infty,h}\in\mathcal{L}(L^2(-1,0;\mathbb{R}^m)$ with range $H^1(-1,0;\mathbb{R}^m)$ satisfies

$$
(4.28)\qquad [\mathcal{P}^3_{\infty,h}y](-1)=hB_1^T\ \mathcal{P}^2_{\infty,h}\ y\quad \forall y\in L^2(-1,0;\mathbb{R}^m)
$$

and $\mathcal{P}^2_{\infty,h}\in\mathcal{L}(L^2(-1,0;\mathbb{R}^m),\mathbb{R}^n)$ has the following representation:

$$
(4.29)\qquad \forall(x,\sigma)\in\mathbb{R}^n\times[-1,0]\ :\ [\mathcal{P}^{2*}_{\infty,h}x](\sigma)=\mathcal{P}^{2*}_{\infty,h}(\sigma)x,
$$

with $\mathcal{P}^{2*}_{\infty,h}\in H^1(-1,0;\mathcal{L}(\mathbb{R}^n,\mathbb{R}^m))$ satisfying the limiting condition

$$
(4.30)\qquad \mathcal{P}^{2*}_{\infty,h}(-1)=B_1^T\mathcal{P}^1_{\infty,h}.\qquad \square
$$

For the state delay, we have analogues of Proposition 4.6 and 4.7.

PROPOSITION 4.8. *For any $k>0$, $t\in[0,T]$, and $T<\infty$, the Riccati operator $\mathcal{Q}_{T,k}(t)$ has in $\mathcal{L}_s(\mathcal{H}^n)$ the following decomposition:*

$$
(4.31)\qquad \mathcal{Q}_{T,k}(t)=\begin{pmatrix} \mathcal{Q}^1_{T,k}(t) & k\mathcal{Q}^2_{T,k}(t) \\ k\mathcal{Q}^{2*}_{T,k}(t) & k\mathcal{Q}^3_{T,k}(t)\end{pmatrix},
$$

with $\mathcal{Q}^1_{T,k}\in C^1(0,T;\mathcal{L}_s(\mathbb{R}^n))$ and operators $\mathcal{Q}^2_{T,k}\in C^1(0,T;\mathcal{L}_s(L^2(-1,0;\mathbb{R}^m),\mathbb{R}^n))$ and $\mathcal{Q}^3_{T,k}\in C^1(0,T;\mathcal{L}_s(L^2(-1,0;\mathbb{R}^n))$ having representations

$$
(4.32)\qquad [\mathcal{Q}^{2*}_{T,k}(t)x](\sigma)=\mathcal{Q}^{2*}_{T,k}(t,\sigma)x\quad\forall(x,\sigma)\in\mathbb{R}^n\times[-1,0],
$$

$$
(4.33)\qquad [\mathcal{Q}^3_{T,k}(t)y](\sigma)=\int_{-1}^0\mathcal{Q}^3_{T,k}(t,\sigma,r)y(r)dr\quad\forall y\in L^2(-1,0;\mathbb{R}^n)
$$

that satisfy, for t *and* $r \in [0,T]$,

$$\mathcal{Q}^{2*}_{T,k}(t,.) \in H^1(-1,0;\mathcal{L}(\mathbb{R}^n,\mathbb{R}^n)), \tag{4.34}$$

$$\mathcal{Q}^{3}_{T,k}(t,.,r) \in H^1([-1,0];\mathcal{L}(\mathbb{R}^n)). \tag{4.35}$$

Moreover, $\mathcal{Q}^3_{T,k}(t)$ *maps* $L^2(-1,0;\mathbb{R}^n)$ *to* $H^1(-1,0;\mathbb{R}^n)$ *and we also have*

$$\mathcal{Q}^{2*}_{T,k}(t,-1) = A_1^T \mathcal{Q}^1_{T,k}(t), \tag{4.36}$$

$$[\mathcal{Q}^3_{T,k}(t)y](-1) = kA_1^T \mathcal{Q}^2_{T,k}(t)y \quad \forall y \in L^2(-1,0;\mathbb{R}^n). \qquad \square \tag{4.37}$$

PROPOSITION 4.9. *For any* $k>0$, *the Riccati operator* $\mathcal{Q}_{\infty,k}$ *may be written in* $\mathcal{L}(\mathcal{H}^n)$ *in the following way:*

$$\mathcal{Q}_{\infty,k} = \begin{pmatrix} \mathcal{Q}^1_{\infty,k} & k\mathcal{Q}^2_{\infty,k} \\ k\mathcal{Q}^{2*}_{\infty,k} & k\mathcal{Q}^3_{\infty,k} \end{pmatrix}, \tag{4.38}$$

where $\mathcal{Q}^1_{\infty,k} \in \mathcal{L}(\mathbb{R}^n)$, *and* $\mathcal{Q}^3_{\infty,k} \in \mathcal{L}(L^2(-1,0;\mathbb{R}^n)$ *with range* $H^1(-1,0;\mathbb{R}^n)$ *satisfies*

$$[\mathcal{Q}^3_{\infty,k}y](-1) = kA_1^T\ \mathcal{Q}^2_{\infty,k}\ y \quad \forall y \in L^2(-1,0;\mathbb{R}^n); \tag{4.39}$$

and $\mathcal{Q}^2_{\infty,k} \in \mathcal{L}(L^2(-1,0;\mathbb{R}^n),\mathbb{R}^n)$ *has the following representation:*

$$\forall (x,\sigma) \in \mathbb{R}^n \times [-1,0]\ :\ [\mathcal{Q}^{2*}_{\infty,k}x](\sigma) = \mathcal{Q}^{2*}_{\infty,k}(\sigma)x \tag{4.40}$$

with $\mathcal{Q}^{2*}_{\infty,k} \in H^1(-1,0;\mathcal{L}(\mathbb{R}^n,\mathbb{R}^n))$ *satisfying the limiting condition*

$$\mathcal{Q}^{2*}_{\infty,k}(-1) = {A_1}^T \mathcal{Q}^1_{\infty,k}. \qquad \square \tag{4.41}$$

Then we may decompose the Riccati equation (4.6) in the following way (we set $H^1 = H^1(-1,0;\mathbb{R}^n)$ and $L^2 = L^2(-1,0;\mathbb{R}^n)$):

$$\left\{\begin{array}{lcl} \forall\ x_1,x_2 \in \mathbb{R}^n\ : & & \\ \langle \dot{\mathcal{Q}}^1_k(t)x_1,x_2\rangle_{\mathbb{R}^n} &=& -\langle A_0x_1, \mathcal{Q}^1_{T,k}(t)x_2\rangle_{\mathbb{R}^n} - \langle \mathcal{Q}^1_{T,k}(t)x_1, A_0x_2\rangle_{\mathbb{R}^n} \\ & & -\langle x_1, R^{-1}\delta_{\sigma=0}\mathcal{Q}^{2*}_{T,k}(t,\sigma)x_2\rangle_{\mathbb{R}^n} \\ & & -\langle R^{-1}\delta_{\sigma=0}\mathcal{Q}^{2*}_{T,k}(t,\sigma)x_1,x_2\rangle_{\mathbb{R}^n} \\ & & +\langle B^T\mathcal{Q}^1_{T,k}(t)x_1, B^T\mathcal{Q}^1_{T,k}(t)x_2\rangle_{\mathbb{R}^n} \\ & & -\langle Cx_1,Cx_2\rangle_{\mathbb{R}^p}, \\ \mathcal{Q}^1_{T,k}(T)x_1 &=& G(T)x_1, \end{array}\right. \tag{4.42}$$

$$\left\{\begin{array}{lcl} \forall x \in \mathbb{R}^n, \forall y \in H^1\ : & & \\ k\langle \dot{\mathcal{Q}}^2_{T,k}(t)y,x\rangle_{\mathbb{R}^n} &=& -\langle \mathcal{Q}^1_{T,k}(t)A_1\delta_{\sigma=-1}y,x\rangle_{\mathbb{R}^n} \\ & & +\langle R^{-1}\delta_{\sigma=0}\mathcal{Q}^3_{T,k}(t)y,x\rangle_{\mathbb{R}^n} - \langle Dy, \mathcal{Q}^{2*}_{T,k}(t,.)x\rangle_{L^2} \\ & & +k\langle B^T\mathcal{Q}^2_{T,k}(t)y, B^T\mathcal{Q}^1_{T,k}(t)x\rangle_{\mathbb{R}^m} \\ & & +k\langle \mathcal{Q}^2_{T,k}(t)y, A_0x\rangle_{\mathbb{R}^n}, \\ \mathcal{Q}^{2*}_k(t,-1)x &=& A_1^T\mathcal{Q}^1_{T,k}(t)x, \end{array}\right. \tag{4.43}$$

$$\left\{\begin{array}{lcl} \forall y \in H^1, \forall z \in H^1\ : & & \\ \langle k\dot{\mathcal{Q}}^3_{T,k}(t)y,z\rangle_{L^2} &=& -\langle Dy, \mathcal{Q}^3_{T,k}(t)z\rangle_{L^2} - \langle \mathcal{Q}^3_{T,k}(t)y, Dz\rangle_{L^2} \\ & & +k\{\langle A_1\delta_{\sigma=-1}y, \mathcal{Q}^2_{T,k}(t)z\rangle_{\mathbb{R}^n} \\ & & -\langle \mathcal{Q}^2_{T,k}(t)y, A_1\delta_{\sigma=-1}z\rangle_{\mathbb{R}^n}\} \\ & & +k^2\langle B^T\mathcal{Q}^2_{T,k}(t)y, B^T\mathcal{Q}^2_{T,k}(t)z\rangle_{\mathbb{R}^m} \\ \mathcal{Q}^3_{T,k}(t)y(-1) &=& kA_1^T\mathcal{Q}^2_{T,k}(t)y. \end{array}\right. \tag{4.44}$$

Some closing remarks follow.

Remark 4.2. In view of Proposition 4.9, we may derive the equations satisfied by the components $\mathcal{Q}^i_{\infty,k}$, for $i \in \{1,2,3\}$, from (4.42), (4.43), (4.44). Obviously it is sufficient to take 0 as a substitute for their right-hand side and the ordered triple $(\mathcal{Q}^1_{\infty,k}, \mathcal{Q}^2_{\infty,k}, \mathcal{Q}^3_{\infty,k})$ for $(\mathcal{Q}^1_{T,k}, \mathcal{Q}^2_{T,k}, \mathcal{Q}^3_{T,k})$. □

Remark 4.3. A similar remark holds for the triples $(\mathcal{P}^1_{\infty,h}, \mathcal{P}^2_{\infty,h}, \mathcal{P}^3_{\infty,h})$ and $(\mathcal{P}^1_{T,h}, \mathcal{P}^2_{T,h}, \mathcal{P}^3_{T,h})$ when recalling that the triple $(\mathcal{P}^1_{T,h}, \mathcal{P}^2_{T,h}, \mathcal{P}^3_{T,h})$ satisfies the system of equations (4.24), (4.25), (4.26). □

5. Differentiability results for Riccati operators. For regularity results on Riccati operators $\mathcal{P}_{\bullet,k}$ and $\mathcal{Q}_{\bullet,k}$, we shall use the shorthand

$$\mathcal{E}_q = \mathcal{L}(\mathcal{H}^q) \times \mathcal{L}(L^2(-1,0;\mathbb{R}^q),\mathbb{R}^n) \times \mathcal{L}(L^2(-1,0;\mathbb{R}^q)) \quad \text{for} \quad q \in \{m,n\} \tag{5.1}$$

$$\text{and } \mathcal{I}_m \text{ defined as } \mathcal{I}_m(y) = -\int_{-1}^{0} y(\sigma)d\sigma \text{ for } y \in \mathbb{R}^m. \tag{5.2}$$

The first theorem is on the sensitivity of the Riccati operator solving the input delay case.

THEOREM 5.1. *We have the following items.*

(i) *For $T < \infty$ and $t \in [0,T]$, the mapping*

$$\Phi : \begin{array}{ccc} \mathbb{R}^*_+ & \longrightarrow & \mathcal{L}(\mathcal{H}^m) \times \mathcal{E}_m, \\ h & \longmapsto & \left(\dfrac{d\mathcal{P}^1_{T,h}}{dt}(t),\ \mathcal{P}^1_{T,h}(t),\ \mathcal{P}^2_{T,h}(t),\ \mathcal{P}^3_{T,h}(t) \right) \end{array} \tag{5.3}$$

satisfies

$$w - \lim_{h\to 0} \Phi(h) = \left(\frac{dP_T}{dt}(t),\ P_T(t),\ P_T(t)B_1\,\mathcal{I}_m,\ 0 \right), \tag{5.4}$$

where $P_T \in \mathcal{L}(\mathbb{R}^n)$ is the nonnegative symmetric solution of the Riccati equation

$$\begin{aligned} &\frac{dP_T}{dt} + A^T P_T + P_T A - P_T B R^{-1} B^T P_T + C^T C = 0 \text{ in } C(0,T;\mathcal{L}(\mathbb{R}^n)), \\ &P_T(T) = G(T). \end{aligned} \tag{5.5}$$

Moreover $\mathcal{P}^1_{T,h}$, the finite dimensional 1-1*-block of $\mathcal{P}_{T,h}$, satisfies*

$$\left(\frac{d\mathcal{P}^1_{T,h}}{dt},\ \mathcal{P}^1_{T,h} \right) = \left(\frac{dP_T}{dt},\ P_T \right) + o(h) \text{ on } [0,T]. \tag{5.6}$$

(ii) *For $T = \infty$, the mapping*

$$\overline{\Phi} : \begin{array}{ccc} \mathbb{R}^*_+ & \longrightarrow & \mathcal{E}_m, \\ h & \longmapsto & (\mathcal{P}^1_{\infty,h},\ \mathcal{P}^2_{\infty,h},\ \mathcal{P}^3_{\infty,h}) \end{array} \tag{5.7}$$

satisfies

$$w - \lim_{h\to 0} \overline{\Phi}(h) = (P_\infty,\ P_\infty B_1\,\mathcal{I}_m,\ 0), \tag{5.8}$$

where $P_\infty \in \mathcal{L}(\mathbb{R}^n)$ is the nonnegative symmetric solution of the equation

$$A^*P_\infty + P_\infty A - P_\infty B R^{-1} B^* P_\infty + C^*C = 0. \tag{5.9}$$

Moreover $\mathcal{P}^1_{\infty,h}$, the finite dimensional 1-1*-block of $\mathcal{P}_{\infty,h}$, satisfies*

$$\mathcal{P}^1_{\infty,h} = P_\infty + o(h). \qquad \square \tag{5.10}$$

Note first that Theorem 5.1 implies that for any $T \in \mathbb{R} \cup \{+\infty\}$:

$$w - \lim_{h\to 0} \mathcal{P}_{T,h} = \begin{pmatrix} P_T & 0 \\ 0 & 0 \end{pmatrix} \quad \text{in} \quad \mathcal{H}^m, \tag{5.11}$$

which is a classical result from Delfour [8]. So by continuity, we can give a meaning to the number $\mathcal{P}_{T,0}$ by setting $\mathcal{P}_{T,0} = w - \lim_{h\to 0}\mathcal{P}_{T,h}$. The novelty is that this theorem may be interpreted as a weak differentiability result for Riccati operators $\mathcal{P}_{T,h}$, for any $T \in \mathbb{R} \cup \{+\infty\}$. Indeed, we can evaluate the limiting value of the ratio

$$V_T(h) = \frac{\mathcal{P}_{T,h} - \mathcal{P}_{T,0}}{h}, \tag{5.12}$$

since Theorem 5.1 simply says that in $\mathcal{H}^m = \mathbb{R}^n \times L^2(0,T;\mathbb{R}^m)$,

$$w - \lim_{h\to 0} V_T(h) = \begin{pmatrix} 0 & P_T B_1 \mathcal{I}_m \\ -B_1^T P_T & 0 \end{pmatrix} \quad \text{for any} \quad T \in \mathbb{R} \cup \{+\infty\}. \tag{5.13}$$

Likewise for the state delay case, we have the following.

THEOREM 5.2. *We have the following items.*

(i) *For $T < \infty$ and $t \in [0,T]$, the mapping*

$$\Psi : \begin{array}{ccc} \mathbb{R}^*_+ & \longrightarrow & \mathcal{L}(\mathcal{H}^n) \times \mathcal{E}_n, \\ k & \longmapsto & \left(\dfrac{d\mathcal{Q}^1_{T,k}}{dt}(t),\ \mathcal{Q}^1_{T,k}(t),\ \mathcal{Q}^2_{T,k}(t),\ \mathcal{Q}^3_{T,k}(t) \right) \end{array} \tag{5.14}$$

satisfies

$$w - \lim_{k\to 0} \Psi(k) = \left(\frac{dP_T}{dt}(t),\ P_T(t),\ P_T(t)A_1 \mathcal{I}_n,\ 0 \right), \tag{5.15}$$

where $P_T \in \mathcal{L}(\mathbb{R}^n)$ is the nonnegative symmetric solution of the Riccati equation (5.5)*. Moreover, $\mathcal{Q}^1_{T,k}$, the finite dimensional* 1-1*-block of $\mathcal{Q}_{T,k}$, satisfies*

$$\left(\frac{d\mathcal{Q}^1_{T,k}}{dt},\ \mathcal{Q}^1_{T,k} \right) = \left(\frac{dP_T}{dt},\ P_T \right) + o(k) \text{ on } [0,T]. \tag{5.16}$$

(ii) *For $T = \infty$, the mapping*

$$\overline{\Psi} : \begin{array}{ccc} \mathbb{R}^*_+ & \longrightarrow & \mathcal{E}_n, \\ k & \longmapsto & (\mathcal{Q}^1_{\infty,k},\ \mathcal{Q}^2_{\infty,k},\ \mathcal{Q}^3_{\infty,k}) \end{array} \tag{5.17}$$

satisfies

$$w - \lim_{k\to 0} \overline{\Psi}(k) = (P_\infty,\ P_\infty A_1 \mathcal{I}_n,\ 0), \tag{5.18}$$

where $P_\infty \in \mathcal{L}(\mathbb{R}^n)$ is the nonnegative symmetric solution of the Riccati equation (5.9)*. Moreover, $\mathcal{Q}^1_{\infty,k}$, the finite dimensional* 1-1*-block of $\mathcal{Q}_{\infty,k}$, satisfies*

$$\mathcal{Q}^1_{\infty,k} = P_\infty + o(k). \qquad \square \tag{5.19}$$

Obviously, we have $w-\lim_{k\to 0}\mathcal{Q}_{T,k}=\left(\begin{smallmatrix}P_T & 0\\ 0 & 0\end{smallmatrix}\right)$ in $\mathcal{H}^n$. Then setting $\mathcal{Q}_{T,0}=P_T$ and $W_T(k)=\frac{\mathcal{Q}_{T,k}-\mathcal{Q}_{T,0}}{k}$, it results from Theorem 5.2 that we have, in $\mathcal{H}^n=\mathbb{R}^n\times L^2(0,T;\mathbb{R}^n)$,

$$w-\lim_{h\to 0}W_T(k)=\begin{pmatrix}0 & P_TA_1\mathcal{I}_n\\ -A_1^TP_T & 0\end{pmatrix}\quad\text{for any } T\in\mathbb{R}\cup\{+\infty\}. \tag{5.20}$$

The proofs of Theorems 5.1 and 5.2 are identical. We restrict ourselves to the first one.

Proof of Theorem 5.1. Proof of (i). Denote

$$(\mathcal{P}_0^1(t),\ \dot{\mathcal{P}}_0^1(t),\ \mathcal{P}_0^2(t),\ \mathcal{P}_0^3(t))=w-\lim_{h\to 0}\Phi(h). \tag{5.21}$$

Because $\forall h>0,\ \forall t\in[0,T]\ :\ \dot{\mathcal{P}}_h(t)\in\mathcal{L}(\mathcal{H}^m)$, there exists some $M>0$ such that $\|\dot{\mathcal{P}}_h(t)\|\le M$. Using the shorthand $H^1=H^1(-1,0;\mathbb{R}^m)$ and $L^2=L^2(-1,0;\mathbb{R}^m)$, we get

$$\forall x\in\mathbb{R}^n,\ \forall y\in H^1:\lim_{h\to 0}h\langle\dot{\mathcal{P}}_h^2(t)y,x\rangle_{\mathbb{R}^n}=0, \tag{5.22}$$

$$\forall y,z\in H^1:\lim_{h\to 0}h\langle\dot{\mathcal{P}}_h^3(t)y,z\rangle_{L^2}=0. \tag{5.23}$$

Thus it follows from (4.24) to (4.26) that $(\mathcal{P}_0^1(t),\ \dot{\mathcal{P}}_0^1(t),\ \mathcal{P}_0^2(t),\ \mathcal{P}_0^3(t))$ satisfies the following:

$$\begin{cases}\langle\dot{\mathcal{P}}_0^1(t)x_1,x_2\rangle_{\mathbb{R}^n} &= -\langle(A-B_0R^{-1}\delta_{\sigma=0}\mathcal{P}_0^{2*}(t,\sigma))x_1,\mathcal{P}_0^1(t)x_2\rangle_{\mathbb{R}^n}\\ &\quad -\langle\mathcal{P}_0^1(t)x_1,(A-B_0R^{-1}\delta_{\sigma=0}\mathcal{P}_0^{2*}(t,\sigma))x_2\rangle_{\mathbb{R}^n}\\ &\quad +\langle B_0^T\mathcal{P}_0^1(t)x_1,R^{-1}B_0^T\mathcal{P}_0^1(t)x_2\rangle_{\mathbb{R}^n}\\ &\quad +\langle\delta_{\sigma=0}\mathcal{P}_0^{2*}(t,\sigma)x_1,R^{-1}\delta_{\sigma=0}\mathcal{P}_0^{2*}(t,\sigma)x_2\rangle_{\mathbb{R}^m}\\ &\quad -\langle Cx_1,Cx_2\rangle_{\mathbb{R}^p},\\ \mathcal{P}_0^1(T)x_1 &= G(T)x_1\quad\forall x_1,x_2\in\mathbb{R}^n;\end{cases} \tag{5.24}$$

$$\begin{cases}0=-\langle\mathcal{P}_0^1(t)B_1\delta_{\sigma=-1}y,x\rangle_{\mathbb{R}^n}+\langle\mathcal{P}_0^1(t)B_0R^{-1}\delta_{\sigma=0}\mathcal{P}_0^3(t)y,x\rangle_{\mathbb{R}^n}\\ \qquad+\langle\delta_{\sigma=0}\mathcal{P}_0^3(t)y,R^{-1}\delta_{\sigma=0}\mathcal{P}_0^{2*}(t,\sigma)x\rangle_{\mathbb{R}^m}-\langle Dy,\mathcal{P}_0^{2*}(t,.)x\rangle_{L^2},\\ \mathcal{P}_0^{2*}(t,-1)x=B_1^T\mathcal{P}_0^1(t)x\quad\forall x\in\mathbb{R}^n,\forall y\in H^1;\end{cases} \tag{5.25}$$

$$\begin{cases}0=-\langle Dy,\mathcal{P}_0^3(t)z\rangle_{L^2}-\langle\mathcal{P}_0^3(t)y,Dz\rangle_{L^2}+\langle\delta_{\sigma=0}\mathcal{P}_0^3(t)y,R^{-1}\delta_{\sigma=0}\mathcal{P}_0^3(t)z\rangle_{\mathbb{R}^m}\\ \mathcal{P}_0^3(t)y(-1)=0\quad\forall y,z\in H^1.\end{cases} \tag{5.26}$$

This last equation may be viewed as the stationary Riccati equation associated to the control problem

$$\begin{cases}\dfrac{\partial Y}{\partial t}(t,\sigma)=DY(t,\sigma)+\delta_{\sigma=0}^*u(t),\ \text{a.e. } t\ge 0\\ u\in L^2(0,\infty;\mathbb{R}^m),\ Y(0,\sigma)\in L^2(-1,0;\mathbb{R}^m),\\ \min\limits_{u\in L^2(0,\infty;\mathbb{R}^m)}J(Y,u)=\displaystyle\int_0^\infty\frac{1}{2}\langle u,Ru(t)\rangle_{\mathbb{R}^m}dt.\end{cases} \tag{5.27}$$

Now we point out that this function Y we wish to control satisfies the equation of transport associated to a unit delay ($h=1$). That is, Y is given by

$$Y(t,\sigma)=u(t+\sigma h)_{|_{h=1}}=u(t+\sigma),\quad -1\le\sigma\le 0. \tag{5.28}$$

Therefore the optimal command for the quadratic control problem (5.27) is the input $u \equiv 0$ in $L^2(-1,0;\mathbb{R}^m)$, which implies that the solution of (5.26) is

$$\mathcal{P}_0^3(t) \equiv 0 \text{ in } L^2(-1,0;\mathbb{R}^m), \tag{5.29}$$

which agrees with the limiting condition $[\mathcal{P}_0^3(t)y](-1) = 0$.

Now let us clarify the relationship between $\mathcal{P}_0^2$ and $\mathcal{P}_0^1$, which satisfy the system

$$\left\{ \begin{array}{l} 0 = \langle \mathcal{P}_0^1(t) B_1 \delta_{\sigma=-1} y, x\rangle_{\mathbb{R}^n} + \langle Dy, \mathcal{P}_0^{2*}(t,.)x\rangle_{L^2}, \\ \mathcal{P}_0^{2*}(t,-1)x = B_1^T \mathcal{P}_0^1(t)x \quad \forall x \in \mathbb{R}^n,\ \forall y \in H^1. \end{array} \right. \tag{5.30}$$

By transposition, we get $\mathcal{P}_0^2(t,.)D = -\mathcal{P}_0^1(t)B_1\delta_{\sigma=-1}$ so that multiplying on the right side with the inverse of the operator $D = \frac{\partial}{\partial\sigma}$ yields

$$\begin{array}{rcl} \mathcal{P}_0^2(t)y & = & -\mathcal{P}_0^1(t)B_1\delta_{\sigma=-1}\displaystyle\int_\sigma^0 y(s)ds \\ & = & -\mathcal{P}_0^1(t)B_1\displaystyle\int_{-1}^0 y(s)ds = -\mathcal{P}_0^1(t)B_1\mathcal{I}_m y \quad \forall y \in L^2(-1,0;\mathbb{R}^m). \end{array} \tag{5.31}$$

There remains the computation of $\mathcal{P}_0^1$. For this we need further insight in the expression $\mathcal{P}_0^{2*}(t,0)$, so we shall apply the integration by parts formula to the factor $\langle Dy, \mathcal{P}_0^{2*}(t,.)x\rangle_{L^2}$. Taking $y \in \mathcal{D}(D) = \{y \in H^1(-1,0;\mathbb{R}^m,\ y(0) = 0\}$, we get

$$\langle Dy, \mathcal{P}_0^{2*}(t,.)x\rangle_{L^2} = -\langle y, D\mathcal{P}_0^{2*}(t,.)x\rangle_{L^2} - \langle \varphi(-1), \mathcal{P}_0^{2*}(t,-1)x\rangle_{\mathbb{R}^m}, \tag{5.32}$$

so (5.30) becomes

$$\left\{ \begin{array}{l} 0 = \langle y, D\mathcal{P}_0^{2*}(t,.)x\rangle_{L^2}\ \forall \begin{pmatrix} x \\ y \end{pmatrix} \in \mathbb{R}^n \times \mathcal{D}(D), \\ \mathcal{P}_0^{2*}(t,-1)x = B_1^T\mathcal{P}_0^1(t)x. \end{array} \right. \tag{5.33}$$

With a density argument, we deduce that the linear form

$$y \to \langle y(\sigma), D\mathcal{P}_0^{2*}(t,\sigma)x\rangle_{L^2} = \int_{-1}^0 \langle y(\sigma), \frac{\partial}{\partial\sigma}\mathcal{P}_0^{2*}(t,\sigma)x\rangle_{\mathbb{R}^m} d\sigma$$

is identically equal to zero on $H^1(-1,0;\mathbb{R}^m)$, and then $\int_{-1}^0 \frac{\partial}{\partial\sigma}\mathcal{P}_0^{2*}(t,\sigma)x d\sigma = 0$. So finally we have

$$\mathcal{P}_0^{2*}(t,0)x = \mathcal{P}_0^{2*}(t,-1)x = B_1^T\mathcal{P}_0^1(t)x \quad \forall x \in \mathbb{R}^n. \tag{5.34}$$

Because $\forall x \in \mathbb{R}^n$, we have $\delta_{\sigma=0}\mathcal{P}_0^{2*}(t,\sigma)x = B_1^T\mathcal{P}_0^1(t)x$, we see that $\mathcal{P}_0^1$ satisfies

$$\left\{ \begin{array}{lcl} \langle \dot{\mathcal{P}}_0^1(t)x_1, x_2\rangle_{\mathbb{R}^n} & = & -\langle (A - B_0R^{-1}B_1^T\mathcal{P}_0^1(t))x_1, \mathcal{P}_0^1(t)x_2\rangle_{\mathbb{R}^n} \\ & & -\langle \mathcal{P}_0^1(t)x_1, (A - B_0R^{-1}B_1^T\mathcal{P}_0^1(t)x_2\rangle_{\mathbb{R}^n} \\ & & +\langle B_0^T\mathcal{P}_0^1(t)x_1, R^{-1}B_0^T\mathcal{P}_0^1(t)x_2\rangle_{\mathbb{R}^n} \\ & & +\langle B_1^T\mathcal{P}_0^1(t)x_1, R^{-1}B_1^T\mathcal{P}_0^1(t)x_2\rangle_{\mathbb{R}^m} - \langle Cx_1, Cx_2\rangle_{\mathbb{R}^p}, \\ \mathcal{P}_0^1(T)x_1 & = & G(T)x_1 \quad \forall\, x_1, x_2 \in \mathbb{R}^n. \end{array} \right. \tag{5.35}$$

Thus, recalling that $B = B_0 + B_1$, we may write

$$\left\{ \begin{array}{lcl} \langle \dot{\mathcal{P}}_0^1(t)x_1, x_2\rangle_{\mathbb{R}^n} & = & -\langle \mathcal{P}_0^1(t)x_1, Ax_2\rangle_{\mathbb{R}^n} - \langle \mathcal{P}_0^1(t)x_1, A\mathcal{P}_0^1(t)x_2\rangle_{\mathbb{R}^n} \\ & & +\langle B^T\mathcal{P}_0^1(t)x_1, R^{-1}B^T\mathcal{P}_0^1(t)x_2\rangle_{\mathbb{R}^n} - \langle Cx_1, Cx_2\rangle_{\mathbb{R}^p}, \\ \mathcal{P}_0^1(T)x_1 & = & Gx_1 \quad \forall\, x_1, x_2 \in \mathbb{R}^n, \end{array} \right. \tag{5.36}$$

which means that $\mathcal{P}_0^1 = P_T$, the unique solution of (5.5), thus completing the proof of formula (5.4).

For approximation of the 1-1-block of $\mathcal{P}_{T,h}$ (equation (5.6)) let us assume that

$$\mathcal{P}_h^1(t) = [\mathcal{P}_0^1 + h\mathcal{P}_h^{1,r}](t), \quad t \in [0,T].$$

Then we have

$$(5.37)\begin{cases} \langle \dot{\mathcal{P}}_h^{1,r}(t)x_1, x_2\rangle_{\mathbb{R}^n} &= -\langle (A - B_0R^{-1}\delta_{\sigma=0}\mathcal{P}_h^{2*}(t,\sigma))x_1, \mathcal{P}_h^{1,r}(t)x_2\rangle_{\mathbb{R}^n} \\ & \quad -\langle \mathcal{P}_h^{1,r}(t)x_1, (A - B_0R^{-1}\delta_{\sigma=0}\mathcal{P}_h^{2*}(t,\sigma))x_2\rangle_{\mathbb{R}^n} \\ & \quad +\langle B_0^T\mathcal{P}_h^1(t)x_1, R^{-1}B_0^T\mathcal{P}_h^{1,r}(t)x_2\rangle_{\mathbb{R}^n} \\ & \quad +\langle B_0^T\mathcal{P}_h^{1,r}(t)x_1, R^{-1}B_0^T\mathcal{P}_h^{1,r}(t)x_2\rangle_{\mathbb{R}^n} \\ & \quad +h\,\langle B_0^T\mathcal{P}_h^{1,r}(t)x_1, R^{-1}B_0^T\mathcal{P}_0^1(t)x_2\rangle_{\mathbb{R}^n}, \\ \mathcal{P}_h^{1,r}(T)x_1 &= 0 \quad \forall\, x_1, x_2 \in \mathbb{R}^n, \end{cases}$$

and passing to the limit with $(\mathcal{P}_0^{1,r}(t),\ \dot{\mathcal{P}}_0^{1,r}(t)) = w - \lim_{h\to 0}(\mathcal{P}_h^{1,r}(t),\ \dot{\mathcal{P}}_h^{1,r}(t))$, and recalling that $\mathcal{P}_0^{2*}(t,0) = B_1^T\mathcal{P}_0^1(t)$, we get

$$(5.38)\quad \begin{cases} \langle \dot{\mathcal{P}}_0^{1,r}(t)x_1, x_2\rangle_{\mathbb{R}^n} &= -\langle (A - BR^{-1}B^T\mathcal{P}_0^1(t))x_1, \mathcal{P}_0^{1,r}(t)x_2\rangle_{\mathbb{R}^n} \\ & \quad -\langle \mathcal{P}_0^{1,r}(t)x_1, (A - BR^{-1}B^T\mathcal{P}_0^1(t))x_2\rangle_{\mathbb{R}^n}, \\ \mathcal{P}_0^{1,r}(T)x_1 &= 0\ \forall\, x_1, x_2 \in \mathbb{R}^n, \end{cases}$$

so that $(\frac{d\mathcal{P}_{T,h}^1}{dt}(t),\ \mathcal{P}_0^{1,r}(t)) \equiv (0,0)$. This implies

$$(5.39)\qquad \left(\frac{d\mathcal{P}_{T,h}^1}{dt}(t),\ \mathcal{P}_{T,h}^1(t)\right) = \left(\frac{dP_T}{dt}(t),\ P_T(t)\right) + o(h) \ \text{ for } \ t \in [0,T],$$

which ends the proof of item (i).

Finally, we observe that item (ii) can be derived in a similar way, thus completing the proof. □

6. Proof of main results. The proofs of Theorems 2.1 and 2.2, identical in principle, are based on Theorems 5.1 and 5.2, respectively. We restrict ourselves to the first one.

Proof of Theorem 2.1. As pointed out in section 3, we may also restrict ourselves to $I = J = 1$.

Proof of (i). Recall that $\widehat{J_T}(0,0) = \frac{1}{2}\langle x_{00}, P_T(0)x_{00}\rangle_{\mathbb{R}^n}$ and that we also have

$$(6.1)\ \widehat{J_T}(0,h) = \frac{1}{2}\langle x_{00}, \mathcal{P}_h^1(0)x_{00}\rangle_{\mathbb{R}^n} + \frac{h}{2}[2\langle x_{00}, \mathcal{P}_h^2(0)\varphi^h\rangle_{L^2} + \langle \varphi^h, \mathcal{P}_h^3(0)\varphi^h\rangle_{L^2}].$$

Now, in view of Theorem 5.1(i), which ensures that $\mathcal{P}_h^1(0) = P_T(0) + o(h)$, we have

$$(6.2)\qquad \widehat{J_T}(0,0) = \frac{1}{2}\left\langle x_{00},\ \mathcal{P}_h^1(0)x_{00}\right\rangle_{\mathbb{R}^n} + o(h)\langle x_{00},\ x_{00}\rangle_{\mathbb{R}^n}.$$

Moreover, from Theorem 5.1, we also have

$$(6.3)\qquad \lim_{h\to 0}\langle y, \mathcal{P}_h^3(0)y\rangle_{L^2} = 0\ \forall y \in L^2$$

so that

$$\lim_{h\to 0}\frac{\widehat{J_T}(0,h)-\widehat{J_T}(0,0)}{h}=\lim_{h\to 0}\langle x_{00},\ \mathcal{P}_h^2(0)\varphi^h\rangle_{L^2} \tag{6.4}$$

$$=-\left\langle B_1^TP_T(0)x_{00},\ \lim_{h\to 0}\int_{-1}^0 u(\sigma h)d\sigma\right\rangle_{\mathbb{R}^m} \tag{6.5}$$

(recall that $\varphi^h(\sigma)=u(\sigma h)$)

$$=-\left\langle B_1^TP_T(0)x_{00},\ \lim_{h\to 0}\frac{1}{h}\int_{-h}^0 u(\theta)d\theta\right\rangle_{\mathbb{R}^m}. \tag{6.6}$$

The continuity of u on the left side of 0 allows us to write

$$\lim_{h\to 0}\frac{1}{h}\int_{-h}^0 u(\theta)d\theta=u(0)=-R^{-1}B^TP_T(0)x_{00}. \tag{6.7}$$

Therefore we have

$$\lim_{h\to 0}\frac{\widehat{J_T}(0,h)-\widehat{J_T}(0,0)}{h}=\langle x_{00},\Delta_1(P_T)x_{00}\rangle_{\mathbb{R}^n} \tag{6.8}$$

with $\Delta_1(P_T)=P(0)B_1R^{-1}B^TP_T(0)$, which we wished to prove.

Proof of (ii) Clearly, in view of Theorem 5.2(i), we have

$$\lim_{k\to 0}\frac{\widehat{J_T}(k,0)-\widehat{J_T}(0,0)}{k}=\lim_{k\to 0}\left\langle x_{00},\ \mathcal{Q}_k^2(0)\psi^k\right\rangle_{L^2}, \tag{6.9}$$

and this readily turns into

$$\lim_{k\to 0}\frac{\widehat{J_T}(k,0)-\widehat{J_T}(0,0)}{k}=-\langle A_1^TP_T(0)x_{00},\ \lim_{k\to 0}\int_{-1}^0\psi^k(\sigma)d\sigma\rangle_{\mathbb{R}^n} \tag{6.10}$$

$$=-\langle A_1^TP_T(0)x_{00},\ \lim_{k\to 0}\int_{-1}^0 x_0(\sigma k)d\sigma\rangle_{\mathbb{R}^n}. \tag{6.11}$$

From continuity of x at 0, we get

$$\lim_{k\to 0}\int_{-1}^0 x_0(\sigma k)d\sigma=\lim_{k\to 0}\frac{1}{k}\int_{-k}^0 x_0(\theta)d\theta\rangle_{\mathbb{R}^n}=x_0(0)=x_{00}, \tag{6.12}$$

and then

$$\lim_{k\to 0}\frac{\widehat{J_T}(k,0)-\widehat{J_T}(0,0)}{k}=-\langle A_1x_{00},\ P(0)x_{00}\rangle_{\mathbb{R}^n}, \tag{6.13}$$

which ends this proof. □

7. Application to sensitivity of $\mathcal{H}_\infty$ robust performance. As a typical example of $\mathcal{H}_\infty$ suboptimal control we consider the system

$$\begin{cases}\dot{x}(t)=Ax(t-k)+B_0u(t)+B_1u(t-h)+B_dw(t)\quad \text{a.e. } t\geq 0,\\ z(t)=Cx(t)+Du(t),\\ x(0)=x_{00}\in\mathbb{R}^n,\quad x=x_0 \text{ a.e. } t\in[-k,0],\quad u=u_0 \text{ a.e. } t\in[-h,0],\\ x_0\in L^2(-k,0;\mathbb{R}^n),\quad u\in L^2(-h,\infty;\mathbb{R}^m),\quad w\in L^2(0,\infty;\mathbb{R}^q),\end{cases} \tag{7.1}$$

where $(A, B_d, B_1) \in \mathcal{L}(\mathbb{R}^n \times \mathbb{R}^q \times \mathbb{R}^m, \mathbb{R}^n)$ and $(C, D) \in \mathcal{L}(\mathbb{R}^n \times \mathbb{R}^m, \mathbb{R}^p)$, with the objective of ensuring a given level of disturbance attenuation. We recall (see Tadmor [28]) that for $\gamma > 0$, the inequality $\|z\|_{L^2(0,\infty;\mathbb{R}^p)} < \gamma^2 \|w\|_{L^2(0,\infty;\mathbb{R}^q)}$ is achievable if and only if there exists a control $u_\gamma \in L^2(0,\infty;\mathbb{R}^m)$ that realizes

$$\widehat{J}_\gamma(k,h) = \sup_{w \in L^2(0,\infty;\mathbb{R}^q)} \inf_{u \in L^2(0,\infty;\mathbb{R}^m)} \frac{1}{2} \{\|z\|_{L^2(0,\infty;Z)} - \gamma^2 \|w\|_{L^2(0,\infty;\mathbb{R}^q)}\}. \tag{7.2}$$

Assume that the initial conditions u_0 and x_0 are continuous on the left side of 0, the triple (A, B_1, C) is controllable and observable, $D^T[C\ D] = [0\ R]$ with R positive definite, and moreover there exists some $\varepsilon > 0$ such that $\forall\ (\omega, x, u) \in \mathbb{R} \times \mathbb{R}^n \times \mathbb{R}^m$ satisfying $j\omega x = Ax + Bu$ we have $\|Cx + Du\|_{\mathbb{R}^p} \geq \varepsilon \|x\|_{\mathbb{R}^n}$. Then it is known from Glover and Doyle [17] that $\widehat{J}_\gamma(0,0)$ exists if and only if there exists a nonnegative solution to the Riccati equation

$$A^T P_\gamma + P_\gamma A + P_\gamma(\gamma^{-2} B_d B_d^T - B R^{-1} B^T) P_\gamma + C^T C = 0, \tag{7.3}$$

where $B = B_0 + B_1$. Then we have the following statement.

THEOREM 7.1. *Take a $\gamma > 0$ such that there exists a nonnegative solution to the Riccati equation* (7.3). *Assume that there exists some neighborhood of* $(0,0)$ *where* $(k,h) \longrightarrow J^\gamma(k,h)$ *is well defined. Then, with* x_0 *continuous on the left side of* 0, *we have*

$$\frac{\partial \widehat{J}_\gamma(0,0)}{\partial k} = \langle P_\gamma x_{00}, A x_{00} \rangle_{\mathbb{R}^n}, \tag{7.4}$$

$$\frac{\partial \widehat{J}_\gamma(0,0)}{\partial h} = -\langle P_\gamma x_{00}, B_1 R^{-1} B^T P_\gamma x_{00} \rangle_{\mathbb{R}^n}. \qquad \square \tag{7.5}$$

For instance, this theorem says that if $B_0 = 0$ and h is sufficiently small, then we have: $\frac{\partial \widehat{J}_\gamma(0,0)}{\partial h} = -\langle x_{00}, K_\gamma R K_\gamma x_{00} \rangle_{\mathbb{R}^n} \leq 0$, where $K_\gamma = -R^{-1} B_1^T P_\gamma$ is the nondelay $\mathcal{H}_\infty$ suboptimal gain, so that $\widehat{J}_\gamma(0,h) \leq \widehat{J}_\gamma(0,0) \leq 0$. This means that *even if there exists a γ-admissible feedback for a small $h > 0$*, the corresponding closed loop system is less robust than the closed loop system, ensuring the disturbance attenuation at the same level without delay. Moreover, for a given γ, the degradation of the robust performance due to the presence of a small input delay is proportional to the square of the nondelay $\mathcal{H}_\infty$ suboptimal gain.

Sketch of the proof. Briefly, we only recast the problem of sensitivity analysis in the framework previously introduced for the LQ optimal case and use a Riccati-type result given by Van Keulen [29] for $\mathcal{H}_\infty$ control in the Pritchard and Salamon class. Then we have analogues of Propositions 4.8 and 4.6 (likewise for Theorems 5.1(ii) and 5.2(ii)). In the following, we survey the main lines of the retarded input case.

First, we observe that with $\mathcal{B}_d = \binom{B_d}{0} \in \mathcal{L}(\mathbb{R}^q, \mathcal{H}^m)$, a product-space description of (7.1) is obtained in $\mathcal{H}^m$ by adding the input $\mathcal{B}_d w$ to the equation of evolution (3.4). That is, the solution x is the first component of

$$Y_h(t,\sigma) = \mathcal{S}_h(t) \begin{pmatrix} x_{00} \\ \varphi_h(\sigma) \end{pmatrix} + \int_0^t \mathcal{S}_h(t-s)\{\mathcal{B}_d w(s) + \mathcal{B}_h u(s)\} ds, \tag{7.6}$$

and setting $\mathcal{C}_h = (\,C \quad 0\,) \in \mathcal{L}(\mathcal{H}^m, \mathbb{R}^q)$ gives

$$z(t) = \mathcal{C}_h Y_h(t,\sigma) + Du(t), \tag{7.7}$$

with $(\mathcal{S}_h, \mathcal{B}_h, \mathcal{C}_h) \in \mathcal{C}_{PS}(\mathcal{H}^m, \mathcal{D}(\mathcal{A}_h^*)', \mathbb{R}^m, \mathbb{R}^p)$ (see Proposition 4.2(i)). Now we remark that if there exists some feedback $F \in \mathcal{L}(\mathcal{H}^m, \mathbb{R}^m)$ (an admissible output operator) such that the semigroup

$$\mathcal{S}_{\mathcal{B}_{h^F}}(t)x = \mathcal{S}_h(t)x + \int_0^t \mathcal{S}_h(t-s)\mathcal{B}_h F\, \mathcal{S}_{\mathcal{B}_{h^F}}(s)x ds, \quad t \geq 0, \quad x \in \mathcal{H}^m, \tag{7.8}$$

is exponentially stable on $\mathcal{D}(\mathcal{A}_h^*)'$, then with [29, Lemma 2.14], we get

$$\begin{cases} Y_h(t,\sigma) &= \mathcal{S}_{\mathcal{B}_{h^F}}(t)\begin{pmatrix} x_{00} \\ \varphi_h(\sigma)\end{pmatrix} + \displaystyle\int_0^t \mathcal{S}_{\mathcal{B}_{h^F}}(t-s)\mathcal{B}_d w(s) ds, \\ z(t) &= (\mathcal{C}_h + DF)Y_h(t,\sigma). \end{cases} \tag{7.9}$$

Moreover, we then have regular $(\mathcal{S}_{\mathcal{B}_{h^F}}, \mathcal{B}_d, (\mathcal{C}_h + DF)) \in \mathcal{C}_{PS}(\mathcal{H}^m, \mathcal{D}(\mathcal{A}_h^*)', \mathbb{R}^m, \mathbb{R}^p)$, and it follows from [29, Lemma 3.2] that the linear operator $G_{\mathcal{B}_{h^F}}$ defined by

$$G_{\mathcal{B}_{h^F}} w(t) = (\mathcal{C}_h + DF)\int_0^t \mathcal{S}_{\mathcal{B}_{h^F}}(t-s)\mathcal{B}_d w(s) ds \tag{7.10}$$

satisfies

$$G_{\mathcal{B}_{h^F}} \in \mathcal{L}(L^2(0,\infty;\mathbb{R}^q), L^2(0,\infty;\mathbb{R}^p)). \tag{7.11}$$

The feedback $F \in \mathcal{L}(\mathcal{H}^m, \mathbb{R}^m)$ will be said to be γ-admissible for system (7.6)–(7.7) if it is an admissible output operator for $\mathcal{H}^m$ and $\mathcal{D}(\mathcal{A}_h^*)'$ with respect to $\mathcal{S}_h$, such that the semigroup $\mathcal{S}_{\mathcal{B}_{h^F}}$ is exponentially stable on $\mathcal{D}(\mathcal{A}_h^*)'$ and $G_{\mathcal{B}_{h^F}}$ satisfies $\|G_{\mathcal{B}_{h^F}}\| \leq \gamma$.

Then the solution of problem (7.2) comes from Van Keulen [29, Theorem 4.4]. That is, assuming that there exists $\varepsilon > 0$ such that $\forall(\omega, X, u) \in \mathbb{R} \times \mathcal{D}(\mathcal{A}_h) \times \mathbb{R}^m$ satisfying $j\omega X = \mathcal{A}_h X + \mathcal{B}_h u$, we have $\|\mathcal{C}_h X + Du\|_{\mathbb{R}^p} \geq \varepsilon \|X\|_{\mathcal{H}^m}$ and that $D^T[C\ D] = [0\ R]$ with R positive; the two following items are equivalent.

(1) There exists a γ-admissible feedback for the system (7.6)–(7.7).

(2) There exists a unique operator $\mathcal{P}_{\gamma,h} \in \sigma^+[\mathcal{D}(\mathcal{A}_h^*)']$ solution of the Riccati equation (in $\mathcal{H}^m$)

$$\mathcal{A}_h^* \mathcal{P}_{\gamma,h} + \mathcal{P}_{\gamma,h}\mathcal{A}_h + \mathcal{P}_{\gamma,h}(\gamma^{-2}\mathcal{B}_d\mathcal{B}_d^* - \mathcal{B}_h R^{-1}\mathcal{B}_h^*)\mathcal{P}_{\gamma,h} + \mathcal{C}_h^*\mathcal{C}_h = 0. \tag{7.12}$$

Moreover, the feedback $F_\gamma = -R^{-1}\mathcal{B}_h^*\mathcal{P}_{\gamma,h}$ is γ-admissible and the input $u_\gamma(t,h) = F_\gamma Y_h(t,.)$ achieves

$$\widehat{J}_\gamma(0,h) = -\frac{1}{2}\langle Y_h(0,.), \mathcal{P}_{\gamma,h} Y_h(0,.)\rangle_{\mathcal{H}^m}. \tag{7.13}$$

The rest of the proof is easy. Observing that Proposition 4.8 still holds when taking $\mathcal{P}_{\gamma,h}$ as a substitute for $\mathcal{P}_h$, and setting

$$\mathcal{P}_{\gamma,h} = \begin{pmatrix} \mathcal{P}^1_{\gamma,h} & h\mathcal{P}^2_{\gamma,h} \\ h\mathcal{P}^{2*}_{\gamma,h} & h\mathcal{P}^3_{\gamma,h} \end{pmatrix}$$

in $\mathcal{L}(\mathcal{H}^m)$, we get the following equations:

$$\begin{cases} \forall x_1, x_2 \in \mathbb{R}^n & : \\ 0 &= -\langle (A - B_0 R^{-1}\delta_{\sigma=0}\mathcal{P}^{2*}_{\gamma,h})x_1, \mathcal{P}^1_{\gamma,h}x_2\rangle_{\mathbb{R}^n} \\ & \quad -\langle \mathcal{P}^1_{\gamma,h}(t)x_1, (A - B_0 R^{-1}\delta_{\sigma=0}\mathcal{P}^{2*}_{\gamma,h})x_2\rangle_{\mathbb{R}^n} \\ & \quad +\langle B_0^T\mathcal{P}^1_{\gamma,h}x_1, R^{-1}B_0^T\mathcal{P}^1_{\gamma,h}x_2\rangle_{\mathbb{R}^n} \\ & \quad +\gamma^{-2}\langle B_d^T\mathcal{P}^1_{\gamma,h}x_1, B_d^T\mathcal{P}^1_{\gamma,h}x_2\rangle_{\mathbb{R}^n} \\ & \quad +\langle \delta_{\sigma=0}\mathcal{P}^{2*}_{\gamma,h}x_1, R^{-1}\delta_{\sigma=0}\mathcal{P}^{2*}_{\gamma,h}x_2\rangle_{\mathbb{R}^m} + \langle Cx_1, Cx_2\rangle_{\mathbb{R}^p}, \end{cases} \tag{7.14}$$

$$
(7.15)\quad \left\{\begin{array}{lcl}
\forall x\in\mathbb{R}^n, & & \forall y\in H^1:\\
0 & = & -\langle \mathcal{P}^1_{\gamma,h}B_1\delta_{\sigma=-1}y,x\rangle_{\mathbb{R}^n}+\langle \mathcal{P}^1_{\gamma,h}B_0R^{-1}\delta_{\sigma=0}\mathcal{P}^3_{\gamma,h}y,x\rangle_{\mathbb{R}^n}\\
& & +\langle \delta_{\sigma=0}\mathcal{P}^3_{\gamma,h}y,R^{-1}\delta_{\sigma=0}\mathcal{P}^{2*}_{\gamma,h}x\rangle_{\mathbb{R}^m}-\langle Dy,\mathcal{P}^{2*}_{\gamma,h}x\rangle_{L^2}\\
& & +h\{\langle \mathcal{P}^2_{\gamma,h}y,Ax\rangle_{\mathbb{R}^n}-\langle B_0^T\mathcal{P}^2_h y,R^{-1}B_0^T\mathcal{P}^1_{\gamma,h}x\rangle_{\mathbb{R}^n}\\
& & -\langle B_0^T\mathcal{P}^2_{\gamma,h}y,R^{-1}\delta_{\sigma=0}\mathcal{P}^{2*}_{\gamma,h}x\rangle_{\mathbb{R}^m}\},\\
\mathcal{P}^{2*}_{\gamma,h}(-1) & = & B_1^T\mathcal{P}^1_{\gamma,h},
\end{array}\right.
$$

and

$$
(7.16)\quad \left\{\begin{array}{lcl}
\forall y\in H^1, & & \forall z\in H^1:\\
0 & = & -\langle Dy,\mathcal{P}^3_h(t)z\rangle_{L^2}-\langle \mathcal{P}^3_{\gamma,h}y,Dz\rangle_{L^2}\\
& & +\langle \delta_{\sigma=0}\mathcal{P}^3_h y,R^{-1}\delta_{\sigma=0}\mathcal{P}^3_{\gamma,h}z\rangle_{\mathbb{R}^m}\\
& & +h\{\langle B_0^T\mathcal{P}^2_{\gamma,h}y,R^{-1}\delta_{\sigma=0}\mathcal{P}^3_{\gamma,h}z\rangle_{\mathbb{R}^m}\\
& & +\langle B_0^T\mathcal{P}^2_{\gamma,h}z,R^{-1}\delta_{\sigma=0}\mathcal{P}^3_{\gamma,h}y\rangle_{\mathbb{R}^m}\\
& & -\langle B_1\delta_{\sigma=-1}y,\mathcal{P}^2_{\gamma,h}z\rangle_{\mathbb{R}^n}-\langle \mathcal{P}^2_{\gamma,h}y,B_1\delta_{\sigma=-1}z\rangle_{\mathbb{R}^n}\}\\
& & +h^2\langle B_0^T\mathcal{P}^2_{\gamma,h}z,R^{-1}B_0^T\mathcal{P}^2_{\gamma,h}y\rangle_{\mathbb{R}^m},\\
\mathcal{P}^3_{\gamma,h}y(-1) & = & hB_1^T\mathcal{P}^2_{\gamma,h}y.
\end{array}\right.
$$

Remark 7.1. Except for the term $\gamma^{-2}B_dB_d^T$ in (7.21), we have to deal with the set of equations satisfied by the components of the decomposition of $\mathcal{P}_h$ (cf. (4.24) to (4.26) and Remark 4.3). Therefore Theorem 5.1(i) remains true when setting $\mathcal{P}_{\gamma,h}$ as a substitute for $\mathcal{P}_h$ and P_γ for $\mathcal{P}_\infty$. Thus, in particular, we have

$$
(7.17)\qquad \mathcal{P}^1_{\gamma,h}=P_\gamma+o(h),\quad w-\lim_{h\to0}\mathcal{P}^3_{\gamma,h}=0,
$$

$$
(7.18)\qquad w-\lim_{h\to0}\mathcal{P}^2_{\gamma,h}y=-P_\gamma B_1\int_{-1}^0 y(\sigma)d\sigma\quad\forall y\in L^2(-1,0;\mathbb{R}^m).\qquad\square
$$

Therefore we get

$$
\lim_{h\to0}\frac{\widehat{J}_\gamma(0,h)-\widehat{J}_\gamma(0,0)}{h}=-\lim_{h\to0}\langle x_{00},\,\mathcal{P}^2_{\gamma,h}\varphi^h\rangle_{L^2}=\langle B_1^T\mathcal{P}_\gamma x_{00},\,\lim_{h\to0}\int_{-1}^0\varphi^h(\sigma)d\sigma\rangle_{L^2}.
$$

But $\varphi^h(\sigma)=u_0(\sigma h)$ so that $\int_{-1}^0\varphi^h(\sigma)d\sigma=\frac{1}{h}\int_{-h}^0 u_0(\theta)d\theta$ and then with continuity of u on the left side of 0 we have

$$
\lim_{h\to0}\int_{-1}^0\varphi^h(\sigma)d\sigma=\lim_{h\to0}\frac{1}{h}\int_{-h}^0 u_0(\theta)d\theta=u_\gamma(0)=-R^{-1}B^TP_\gamma x_{00}.
$$

So finally we get

$$
\lim_{h\to0}\frac{\widehat{J}_\gamma(0,h)-\widehat{J}_\gamma(0,0)}{h}=-\langle x_{00},P_\gamma B_1R^{-1}B^TP_\gamma x_{00}\rangle_{\mathbb{R}^n},
$$

which we aimed to prove. $\square$

8. Conclusion. In this paper, we have considered the sensitivity of the optimal cost for linear time invariant systems with respect to small delays. We have given a new approach to solving this problem and provided new results. Related topics of interest, such as providing a measure of degradation for when the delay free optimal feedback is implemented on the delayed model, are under investigation.

Appendix.

A.1. Proof of Proposition 3.1. Part (i). The characterization of the C_0-semigroup $\mathcal{S}_h$ and $\mathcal{A}_h$ and Part (i) follows easily from a rescaling in the standard form of the solution semigroup for the delay equation (see, e.g., [16]). This semigroup is given as $\mathcal{S}_h(t) = \left(\begin{smallmatrix} \mathcal{S}_{11}(t) & \mathcal{S}_{12}(t) \\ 0 & \mathcal{S}_{22}(t) \end{smallmatrix}\right)$, where $\mathcal{S}_{11}(t) \in \mathcal{L}(\mathbb{R}^n)$, $\mathcal{S}_{12}(t) \in \mathcal{L}((L^2(-1,0;\mathbb{R}^m),\mathbb{R}^n)$ and $\mathcal{S}_{22}(t) \in \mathcal{L}(L^2(-1,0;\mathbb{R}^m))$ are given by

$$\begin{array}{rcl} \mathcal{S}_{11}(t) & = & e^{At}, \quad \mathcal{S}_{12}(t)z = h\displaystyle\int_0^{\frac{t}{h}} e^{A(t-\sigma h)} B_1 z(\sigma-1)d\sigma, \\ \mathcal{S}_{22}(t) & = & T\left(\dfrac{t}{h}\right), \text{ with } T(t) \text{ semigroup of translation in } L^2(-1,0;\mathbb{R}^m). \end{array} \tag{A.1}$$

Now let us characterize the adjoint of $\mathcal{A}_h$. For this, taking some $\left(\begin{smallmatrix} x \\ z \end{smallmatrix}\right) \in \mathcal{D}(\mathcal{A}_h)$, consider $\left(\begin{smallmatrix} v \\ w \end{smallmatrix}\right) \in \mathcal{D}(\mathcal{A}_h^*)$ and $\left(\begin{smallmatrix} f \\ g \end{smallmatrix}\right) \in \mathcal{H}^m$ such that

$$\begin{aligned} \left\langle \mathcal{A}_h \begin{pmatrix} x \\ z \end{pmatrix}, \begin{pmatrix} v \\ w \end{pmatrix} \right\rangle_{\mathcal{H}} &= \langle Ax + B_1 z(-1), v\rangle_{\mathbb{R}^n} + \frac{1}{h}\int_{-1}^0 \langle Dz, w\rangle_{\mathbb{R}^m} d\sigma \\ &= \langle x, f\rangle_{\mathbb{R}^n} + \int_{-1}^0 \langle z, g\rangle_{\mathbb{R}^m} d\sigma. \end{aligned}$$

So choosing $z = 0$, we get $\langle Ax, v\rangle_{\mathbb{R}^n} = \langle x, f\rangle_{\mathbb{R}^n}$. So $v \in \mathbb{R}^n$ and $A^T v = f$.

Moreover, since $\int_{-1}^0 \langle z, g\rangle_{\mathbb{R}^m} d\sigma = -\int_{-1}^0 \langle Dz, \int_{-1}^\sigma g(s)ds\rangle_{\mathbb{R}^m} d\sigma$, we have $\int_{-1}^0 \langle Dz, \frac{1}{h}w - B_1^T v + \int_{-1}^\sigma g(s)ds\rangle_{\mathbb{R}^m} d\sigma = 0$. Therefore $w = h(B_1^T v - \int_{-1}^\sigma g(s)ds)$, which implies $w(-1) = hB_1^T v$ and $Dw = -hg$. So finally we get

$$\mathcal{A}_h^* \begin{pmatrix} y \\ w \end{pmatrix} = \begin{pmatrix} A^T y \\ -\dfrac{1}{h} Dw \end{pmatrix} \quad \text{and} \quad \mathcal{D}(\mathcal{A}_h^*) = \left\{ \begin{pmatrix} y \\ w \end{pmatrix} \in \mathcal{W}^m,\ w(-1) = hB_1^T y \right\},$$

which we intended to prove.

Part (ii). Now the semigroup $\{\mathcal{S}_h(t)\}_{t\geq 0}$ may be extended by transposition and duality as a semigroup on $\mathcal{D}(\mathcal{A}_h^*)'$ (see, e.g., [19]) as soon as we prove that $\{\mathcal{S}_h^*(t)\}_{t\geq 0}$ is a C_0-semigroup on $\mathcal{D}(\mathcal{A}_h^*)$. The adjoint semigroup $\mathcal{S}_h^*$ may be computed as $\mathcal{S}_h^*(t) = \left(\begin{smallmatrix} \mathcal{S}_{11}^*(t) & 0 \\ \mathcal{S}_{12}^*(t) & \mathcal{S}_{22}^*(t) \end{smallmatrix}\right) \in \mathcal{L}(\mathcal{H})$, where

$$\begin{aligned} \mathcal{S}_{11}^*(t) &= e^{A^T t},\ t \geq 0, \\ \mathcal{S}_{12}^*(t)v(\sigma) &= hB_1^T\, \overline{\mathcal{S}_{11}^*}(t - h(\sigma+1))v \ \text{ with } \ \overline{\mathcal{S}_{11}^*}(t) = \mathcal{S}_{11}^*(t)\chi_{\{t\geq 0\}}, \\ \mathcal{S}_{22}^*(t)z(\sigma) &= \begin{cases} z\left(\sigma - \dfrac{t}{h}\right) & \text{if } \ t - h < \sigma h \leq 0, \\ 0 \text{ if } \ -h \leq \sigma h \leq t - h. \end{cases} \end{aligned}$$

Then for the rest of the proof, it will be sufficient to check that the space $\mathcal{D}(\mathcal{A}_h^*)$ is stable by $\mathcal{S}_h^*(t)$. For that, take $\left(\begin{smallmatrix} v \\ w \end{smallmatrix}\right) \in \mathcal{D}(\mathcal{A}_h^*)$ and denote $\psi(\sigma)$ the second component of $\mathcal{S}_h^*(t)\left(\begin{smallmatrix} v \\ w \end{smallmatrix}\right)$. Then $\psi(\sigma) = hB_1^T\, \overline{\mathcal{S}}_{11}^*(t - h(\sigma+1))v + w(\sigma - \frac{t}{h})$, which implies :$\psi(-1) = hB_1^T\, \overline{\mathcal{S}_{11}^*}(t)v$.

Moreover,

$$\begin{cases} \sigma \leq \dfrac{t-h}{h} \Longrightarrow \psi(\sigma) = hB_1^T\, \mathcal{S}_{11}^*(t - h(\sigma+1))v, \\ \sigma > \dfrac{t-h}{h} \Longrightarrow \psi(\sigma) = w\left(\sigma - \dfrac{t}{h}\right), \end{cases}$$

with $\lim_{\sigma\to\frac{t-h}{h}} w(\sigma-\frac{t}{h}) = w(-1) = hB_1^T v = \psi(\frac{t-h}{h})$. Therefore, in order to prove that $D\psi \in L^2(-1,0;\mathbb{R}^m)$, it is enough to prove that $hB_1^T\ \mathcal{S}_{11}^*(t-h(.+1))v \in H^1(-1,\frac{t-h}{h};\mathbb{R}^m)\ \forall v\in\mathbb{R}^n$. But we may point out that $\{t\to\mathcal{S}_{11}^*(t)=e^{A^Tt}\}\in H^1(0,T;\mathcal{L}(\mathbb{R}^n))$, so when introducing F defined from $[-1,0]$ to $\mathbb{R}^n$ as $F(\sigma)=\mathcal{S}_{11}^*(t-h(\sigma+1))v$, we have the following smoothness property:

$$F\in H^1\left(\frac{t-T-h}{h},\frac{t-h}{h};\mathbb{R}^n\right)\subset H^1\left(-1,\frac{t-h}{h};\mathbb{R}^n\right)\quad\forall h>0.$$

So the final conclusion follows from the remark:

$$B_1^T\in\mathcal{L}(\mathbb{R}^n,\mathbb{R}^m)\implies hB_1^T\mathcal{S}_{11}^*(t-h(.+1))v=hB_1^TF(.)\in H^1\left(-1,\frac{t-h}{h};\mathbb{R}^m\right).\qquad\square$$

A.2. Proof of Lemma 3.2. In view of [1, Proposition 3.1, p. 172], this is equivalent to proving that there exists some $b>0$ such that $\forall\ t\in[0,T]$ and $u\in L^2(0,t\ ;\ \mathbb{R}^m)$, we get

$$\int_0^t\mathcal{S}_h(t-s)\mathcal{B}_hu(s)ds\in\mathcal{H}^m\ \text{ and }\ \|\int_0^t\mathcal{S}_h(t-s)\mathcal{B}_hu(s)ds\|_{\mathcal{H}^m}\le b\|u\|_{L^2(0,t;\mathbb{R}^m)}.$$

To this end, it will be sufficient to prove that $y\longrightarrow\langle\int_0^t\mathcal{S}_h(t-s)\mathcal{B}_hu(s)ds,y\rangle_{\mathcal{H}^m}$ defines a continuous linear form for the normed space topology of $\mathcal{H}^m$.

For this, let us take $z=\binom{y}{w}\in\mathcal{D}(\mathcal{A}_h^*)=\{\binom{y}{w}\in\mathcal{W}^m,\ w(-1)=hB_1^Ty\}$. Then

$$\left\langle\int_0^t\mathcal{S}_h(t-s)\mathcal{B}_hu(s)ds,z\right\rangle_{\mathcal{H}^m}=\int_0^t\langle\underbrace{\mathcal{B}_hu(s)}_{\in\mathcal{W}^{m\prime}},\ \underbrace{\mathcal{S}_h^*(t-s)z}_{\in\mathcal{D}(\mathcal{A}_h^*)\subset\mathcal{W}^m}\rangle_{\mathcal{H}^m}ds.$$

So denoting ψ the second component of $\mathcal{S}_h^*(t-s)z$, we get, with Appendix A.1,

$$\psi(\sigma)=hB_1^Te^{A^T(t-s-h(\sigma+1))}y\chi_{\{\sigma\le\frac{t-s-h}{h}\}}+w\left(\sigma-\frac{t-s}{h}\right)\chi_{\{\sigma\ge\frac{t-s-h}{h}\}}$$

in such a way that ψ defines a continuous function (see the end of the proof of Proposition 3.1 in appendix A.1 for details). Then applying the Dirac distribution $\delta_{\theta=0}$ to ψ, we get

$$\begin{aligned}\int_0^t\langle\mathcal{B}_hu(s),\mathcal{S}_h^*(t-s)z\rangle_{\mathcal{H}^m}ds&=\int_0^t\langle u(s),B_0e^{A^T(t-\sigma h)}y+{B_1}^Te^{A^T(t-h-s)}y\chi_{\{s\le t-h\}}\\&\qquad+\frac{1}{h}w\left(\frac{t-s}{h}\right)\chi_{\{s>t-h\}}\rangle_{\mathcal{H}_h^m}ds\\&=\left\langle\int_0^te^{A(t-s)}B_0u(s)ds+\int_0^te^{A(t-s)}B_1u(s-h)ds,y\right\rangle_{\mathbb{R}^n}\\&\qquad+\int_{-h}^0\langle u(s(h+1)+t),w(s)\rangle_{\mathbb{R}^m}ds,\end{aligned}$$

which proves the continuity of $z\to\langle\int_0^t\mathcal{S}_h(t-s)\mathcal{B}_hu(s)ds,z\rangle_{\mathcal{H}^m}$ in $\mathcal{D}(\mathcal{A}_h^*)$ endowed with the normed space topology of $\mathcal{H}^m$. By density argument, we conclude that this continuity result remains true in the normed space $\mathcal{H}_h^m$. $\square$

A.3. Proof of Theorem 3.6. In view of Remark 3.2 about identifying strong and weak solutions, we just need to prove that $\tilde{\mathcal{A}}_k$ defined by (3.17) coincides on $\mathcal{W}^n$ with $(\mathcal{A}_k^*)^*$, the transposed adjoint of $\mathcal{A}_k$. To this end, we shall use the method of transposition which simply reduces to a double integration by parts. Starting with the equation

$$\text{(A.2)} \qquad \left\{ \begin{array}{l} \begin{pmatrix} x(t) \\ y(t,.) \end{pmatrix} = \begin{pmatrix} A_0 & A_1\delta_{\sigma=-1} \\ 0 & \dfrac{1}{k}\dfrac{\partial}{\partial\sigma} \end{pmatrix} \begin{pmatrix} x(t) \\ y(t,.) \end{pmatrix} \\ \begin{pmatrix} x(0) \\ y(0,.) \end{pmatrix} = \begin{pmatrix} x_{00} \\ \psi^k(.) \end{pmatrix} \end{array} \right. \qquad \text{in} \quad \mathcal{H}^m,$$

we observe that any $\binom{x(t)}{y(t,.)} \in \mathcal{D}(\mathcal{A}_k)$ satisfies

$$\text{(A.3)} \qquad \left\{ \begin{array}{l} y(t,.) \in H^1(-1,0;\mathbb{R}^n) \ \text{ a.e. } t \in [0,T], \\ y(t,0) = x(t) \in L^2(0,T;\mathbb{R}^n), \end{array} \right.$$

and also that such a y is the solution to the equation of transport:

$$\text{(A.4)} \qquad \left\{ \begin{array}{l} k\dfrac{\partial y}{\partial t} = \dfrac{\partial y}{\partial \sigma} \\ y(0,.) = \psi^k(.) \in L^2(-1,0;\mathbb{R}^n) \\ y(t,0) = x(t) \ \in L^2(0,T,\mathbb{R}^n) \end{array} \right. \qquad \text{in } L^2(0,T;L^2(-1,0;\mathbb{R}^n)).$$

Now let us apply the method of transposition to the previous equation. Take $\varphi \in L^2(0,T;L^2(-1,0;\mathbb{R}^n)) = L^2(0,T;\mathbb{R}^n) \times L^2(-1,0;\mathbb{R}^n)$. Then (A.4) implies

$$\text{(A.5)} \quad k\int_{t=0}^{t=T}\int_{\sigma=-1}^{\sigma=0} \frac{\partial y}{\partial t}(t,\sigma)\varphi(t,\sigma)dtd\sigma = \int_{t=0}^{t=T}\int_{\sigma=-1}^{\sigma=0} \frac{\partial y}{\partial \sigma}(t,\sigma)\varphi(t,\sigma)dtd\sigma.$$

From an integration by parts, it follows that

$$\text{(A.6)} \quad k\int_{\sigma=-1}^{\sigma=0}\left(y\varphi|_{t=0}^{t=T} - \int_{t=0}^{t=T} y\frac{\partial\varphi}{\partial t}\right)d\sigma = \int_{t=0}^{t=T}\left(y\varphi|_{\sigma=-1}^{\sigma=0} - \int_{\sigma=-1}^{\sigma=0} y\frac{\partial\varphi}{\partial\sigma}d\sigma\right)dt.$$

Now, let us make the following assumptions on the test-functions:

$$\varphi(t,-1) = 0 \ \text{ a.e. } t \in [0,T] \quad \text{and} \quad \varphi(T,\sigma) = 0 \ \text{ a.e. } \sigma \in [-1,0].$$

Then in view of the initial and limiting conditions given in (A.4), (A.6) becomes

$$\text{(A.7)} \qquad k\int_{\sigma=-1}^{\sigma=0}\psi^k(\sigma)\varphi(0,\sigma)d\sigma - \int_{t=0}^{t=T}\int_{\sigma=-1}^{\sigma=0} y\left(k\frac{\partial\varphi}{\partial t} - \frac{\partial\varphi}{\partial\sigma}d\sigma\right)dt = \int_{t=0}^{t=T} x(t)\varphi(t,\sigma)dt.$$

Since a necessary condition to make such an equation meaningful is

$$k\frac{\partial\varphi}{\partial t} - \frac{\partial\varphi}{\partial\sigma} \in L^2(0,T;\mathbb{R}^n) \times L^2(-1,0;\mathbb{R}^n),$$

we choose the space

$$\mathcal{E}_{test} = \left\{ \varphi \in L^2(0,T;L^2(-1,0;\mathbb{R}^n)), \quad \left| \begin{array}{l} k\dfrac{\partial\varphi}{\partial t} - \dfrac{\partial\varphi}{\partial\sigma} \in L^2(0,T;L^2(-1,0;\mathbb{R}^n)), \\ \varphi(T,.) = 0, \quad \varphi(.,-1) = 0, \end{array} \right. \right\}$$

as a set of test-functions. From

$$\frac{\partial}{\partial t}^{*} = -\frac{\partial}{\partial t},\ \mathcal{D}\left(\frac{\partial}{\partial t}^{*}\right) = \{y \in H^1(0,T;\mathbb{R}^n),\ y(T) = 0\}$$

and

$$\frac{\partial}{\partial \sigma}^{*} = -\frac{\partial}{\partial \sigma},\ \mathcal{D}\left(\frac{\partial}{\partial \sigma}^{*}\right) = \{y \in H^1(-1,0;\mathbb{R}^n),\ y(-1) = 0\},$$

we see that the space $\mathcal{E}_{test}$ is a subset of the set of strong solutions of the adjoint equation of (A.4) and that

$$\varphi \in \mathcal{E}_{test} \quad \Leftrightarrow \quad \varphi(.,\sigma) \in H^1\left(0,T;\mathcal{D}\left(\frac{\partial}{\partial \sigma}^{*}\right)\right) \text{ and } \varphi(t,.) \in H^1\left(-1,0;\mathcal{D}\left(\frac{\partial}{\partial t}^{*}\right)\right).$$

Now for $\varphi \in \mathcal{E}_{test}$, we may write, with $L^2 \times L^2 = L^2(0,T;\mathbb{R}^n) \times L^2(-1,0;\mathbb{R}^n)$,

$$k\langle \psi^k, \delta_{t=0}\varphi\rangle_{L^2(-1,0;\mathbb{R}^q)} - \left\langle y, \left(k\frac{\partial}{\partial t} - \frac{\partial \varphi}{\partial \sigma}\right)\varphi\right\rangle_{L^2\times L^2} = \langle \delta^*_{\sigma=0}x(t), \varphi\rangle_{L^2\times L^2}. \tag{A.8}$$

Then with the operator $\frac{\partial}{\partial t} - \frac{\partial \varphi}{\partial \sigma}$, the formal adjoint of the operator $-(\frac{\partial}{\partial t} - \frac{\partial \varphi}{\partial \sigma})$, we get

$$k\langle \psi^k, \delta_{t=0}\varphi\rangle_{L^2(-1,0;\mathbb{R}^n)} + \left\langle \left(k\frac{\partial}{\partial t} - \frac{\partial \varphi}{\partial \sigma}\right)y, \varphi\right\rangle_{L^2\times L^2} = \langle \delta^*_{\sigma=0}x(t), \varphi\rangle_{L^2\times L^2}. \tag{A.9}$$

Hence

$$\left\{\begin{array}{lcll} k\dfrac{\partial y}{\partial t}(t,.) & = & \dfrac{\partial y}{\partial \sigma}(t,.) + \delta^*_{\sigma=0}x(t) & \\ & = & Dy(t,.) + \delta^*_{\sigma=0}x(t) & \text{in } H^1(-1,0;\mathbb{R}^n)', \\ y(0,.) & = & \psi^k(.) \in L^2(-1,0;\mathbb{R}^n) & \end{array}\right. \tag{A.10}$$

where $D = \frac{\partial}{\partial \sigma}$ with $\mathcal{D}(\frac{\partial}{\partial \sigma}) = \{y \in H^1(-1,0;\mathbb{R}^n),\ y(0) = 0\}$.

We may observe that the linear operator $(\frac{\partial}{\partial \sigma}^{*})^* \in \mathcal{L}(L^2(-1,0;\mathbb{R}^n), \mathcal{D}(\frac{\partial}{\partial \sigma}^{*})')$ has been constructed en route as a continuous extension of $D \in \mathcal{L}(\mathcal{D}(D), L^2(-1,0;\mathbb{R}^n))$. Hence D may be seen as an element of

$$\mathcal{L}(H^1(-1,0;\mathbb{R}^n), H^1(-1,0;\mathbb{R}^n)') \cap \mathcal{L}\left(L^2(-1,0;\mathbb{R}^n), \mathcal{D}\left(\frac{\partial}{\partial \sigma}^{*}\right)'\right).$$

Now (A.2) becomes

$$\left\{\begin{array}{l} \begin{pmatrix}\dot{x}\\ \dot{y}\end{pmatrix} = \widetilde{\mathcal{A}}_k \begin{pmatrix}x\\ y\end{pmatrix} \\ \begin{pmatrix}x(0)\\ y(0,.)\end{pmatrix} = \begin{pmatrix}x_{00}\\ \psi^k(.)\end{pmatrix} \in \mathcal{H}^n \end{array}\right. \text{ in } L^2(0,T,\mathcal{D}(\mathcal{A}^*_k)'), \tag{A.11}$$

where $\widetilde{\mathcal{A}}_k$ is the extension by transposition of $\mathcal{A}_k$, given on $\mathcal{W}^n$ by

$$\widetilde{\mathcal{A}}_k = \begin{pmatrix} A_0 & A_1\delta_{\sigma=-1} \\ \dfrac{1}{k}\delta^*_{\sigma=0} & \dfrac{1}{k}D \end{pmatrix}$$

and satisfies $\widetilde{\mathcal{A}}_k \in \mathcal{L}(\mathcal{D}(\mathcal{A}_k), \mathcal{H}^n) \cap \mathcal{L}(\mathcal{W}^n, \mathcal{W}^{n'}) \cap \mathcal{L}(\mathcal{H}^n, \mathcal{D}(\mathcal{A}^*_k)')$, which is what we wanted to prove. □

REFERENCES

[1] A. Bensoussan, G. Da Prato, M. C. Delfour, and S. K. Mitter, *Representation and Control of Infinite Dimensional Systems* Vol. 1, Birkhaüser, Basel, Boston, 1992.

[2] D. H. Chyung, *Controllability of linear systems with multiple delays in control*, IEEE Trans. Automat. Control, AC-15 (1970), pp. 694–695.

[3] F. H. Clarke and P. R. Wolenski, *The sensitivity of optimal control problems to time delay*, SIAM J. Control Optim., 29 (1991), pp. 1176–1215.

[4] R. F. Curtain, H. Logemann, S. Townley, and H. Zwart, *Well-posedness, stabilizability and admissibility for Pritchard–Salamon systems*, J. Math. Systems Estim. Control, 4 (1994), pp. 1–38.

[5] R. F. Curtain and A. J. Pritchard, *The infinite dimensional Riccati equation for systems defined by evolution operators*, SIAM J. Control Optim., 14 (1976), pp. 951–983.

[6] R. F. Curtain and A. J. Pritchard, *An abstract theory for unbounded control action for distributed parameter systems*, SIAM J. Control Optim., 15 (1977), pp. 566–611.

[7] R. F. Curtain and H. Zwart, *An Introduction to Infinite-Dimensional Linear Systems Theory*, Springer-Verlag, New York, 1995.

[8] M. C. Delfour, *The linear quadratic optimal control problem with delays in state and control variables: A state space approach*, SIAM J. Control Optim., 24 (1986), pp. 835–883.

[9] M. C. Delfour and J. Karrakchou, *State space theory of linear time invariant systems with delays in state, control and observation variables, Part* I, J. Math Anal. Appl., 125 (1987), pp. 361–399.

[10] M. C. Delfour and J. Karrakchou, *State space theory of linear time invariant systems with delays in state, control and observation variables, Part* II, J. Math Anal. Appl., 125 (1987), pp. 400–450.

[11] M. C. Delfour, C. McCalla, and S. K. Mitter, *Stability and the infinite time quadratic cost problem for linear hereditary differential systems*, SIAM J. Control, 13 (1975), pp. 48–88.

[12] M. C. Delfour and S. K. Mitter, *Controllability, observability, and optimal feedback control of affine hereditary differential systems*, SIAM J. Control, 10 (1972), pp. 298–328.

[13] A. L. Dontchev, *Sensitivity analysis of linear infinite-dimensional optimal control systems under changes of system order*, Control Cybernet., 3 (1974), pp. 21–35.

[14] A. L. Dontchev, *Sensitivity analysis of optimal control system with small time delay*, Control Cybernet., 4 (1975), pp. 91–104.

[15] A. L. Dontchev, *Perturbations, Approximations and Sensitivity Analysis of Optimal Control Systems*, Springer-Verlag, New York, 1983.

[16] A. Ichikawa, *Quadratic control of evolution equations with delays in control*, SIAM J. Control Optim., 20 (1982), pp. 645–668.

[17] K. Glover and J. C. Doyle, *State-space formulae for all stabilizing controllers that satisfy an $\mathcal{H}_\infty$-norm bound and relations to risk sensitivity*, Systems Control Lett., 11 (1988), pp. 167–172.

[18] P. V. Kokotović and R. A. Yackel, *Singular perturbation of linear regulators: Basic theorems*, IEEE Trans. Automat. Control, AC-17 (1972), pp. 29–37.

[19] A. Pazy, *Semigroups of Linear Operators and Application to Partial Differential Equations*, Springer-Verlag, New York, 1983.

[20] A. J. Pritchard, *Sensitivity analysis for evolution equations in Hilbert space*, in Proceedings of the 3rd IFAC Conference on Sensitivity, Adaptivity, and Optimality, Ischia, Italy, Instrument Soc. Amer., Pittsburgh, PA, 1973, pp. 107–113.

[21] A. J. Pritchard and D. Salamon, *The linear quadratic control problem for infinite dimensional systems with unbounded input and output operators*, SIAM J. Control Optim., 25 (1987), pp. 121–144.

[22] A. J. Pritchard and D. Salamon, *The linear quadratic control problem for retarded systems with delays in control and observation*, IMA J. Control Information, 2 (1985), pp. 335–362.

[23] D. Salamon, *Control and Observation of Neutral Systems*, Pitman, Boston, 1984.

[24] D. Salamon, *Infinite dimensional linear systems with unbounded control and observation: A functional analytic approach*, Trans. Amer. Math. Soc., 300 (1987), pp. 383–431.

[25] O. J. Staffans, *Extended initial and forcing function semigroups generated by a functional equation*, SIAM J. Math. Anal., 16 (1985), pp. 1034–1048.

[26] O. J. Staffans, *Semigroups generated by a neutral functional-differential equation*, SIAM J. Math. Anal., 17 (1986), pp. 46–57.

[27] O. J. Staffans, *Quadratic optimal control of stable systems through spectral factorization*, Math. Control Signals Systems, 8 (1995), pp. 167–197.

[28] G. Tadmor, *Worst-case design in the time domain. The maximum principle and the standard*

$\mathcal{H}_\infty$*-problem*, Math Control Signals Systems, 3 (1990), pp. 301–324.

[29] B. Van Keulen, $\mathcal{H}_\infty$ *Control for Infinite-Dimensional Systems*: *A State Space Approach*, Ph.D. thesis, Department of Mathematics and Computing Science, University of Groningen, Groningen, The Netherlands, 1993.

[30] R. B. Vinter and R. H. Kwong, *The infinite time quadratic control problem for linear systems with state and control delays: An evolution equation approach*, SIAM J. Control Optim., 19 (1981), pp. 139–153.

SIAM J. CONTROL OPTIM.
Vol. 38, No. 6, pp. 1683–1706

ADAPTIVE CONTROL FOR SEMILINEAR STOCHASTIC SYSTEMS*

T. E. DUNCAN†, B. MASLOWSKI‡, AND B. PASIK-DUNCAN†

Abstract. An adaptive, ergodic cost stochastic control problem for a partially known, semilinear, stochastic system in an infinite dimensional space is formulated and solved. The solutions of the Hamilton–Jacobi–Bellman equations for the discounted cost and the ergodic cost stochastic control problems require some special interpretations because they do not typically exist in the usual sense. The solutions of the parameter dependent ergodic Hamilton–Jacobi–Bellman equations are obtained from some corresponding discounted cost control problems as the discount rate tends to zero. The solutions of the ergodic Hamilton–Jacobi–Bellman equations are shown to depend continuously on the parameter. A certainty equivalence adaptive control is given that is based on the optimal controls from the solutions of the ergodic Hamilton–Jacobi–Bellman equations and a strongly consistent family of estimates of the unknown parameter. This adaptive control is shown to achieve the optimal ergodic cost for the known system.

Key words. stochastic adaptive control, ergodic control, stochastic semilinear systems, stochastic optimal control, distributed parameter systems

AMS subject classifications. 93C40, 93C20, 60J27, 60H15

PII. S0363012999351826

1. Introduction.

Ergodic cost stochastic control problems for finite dimensional, nonlinear stochastic systems have been investigated for more than two decades (e.g., [2], [9], [25], and [1], [28], and the references therein). There has been some work on the corresponding adaptive control problem for these partially known stochastic systems. In [3], [4], [5] some properties of the solution of the Hamilton–Jacobi–Bellman (HJB) equation of optimal control are used to verify self-optimality of an adaptive control using a strongly consistent family of estimates of the unknown parameters. In [17] an almost optimal adaptive control is constructed for a partially known nonlinear stochastic system. To obtain an optimal feedback control in an explicit form, the associated HJB equation must be solved, and for an adaptive control problem, a continuous dependence of the solution of the HJB equation must be verified. Some results for an infinite time horizon discounted cost control problem for a semilinear stochastic system in an infinite dimensional state space are given in [22], where the HJB equation is considered in the mild form, and in [23], where a viscosity solution of the HJB equation is used. Some results for an ergodic cost control problem for a semilinear stochastic system in an infinite dimensional state space are given in [20], where the HJB equation is solved. This is apparently the only result for ergodic cost control for semilinear stochastic systems using the HJB equation. Some other approaches and results for ergodic cost control are given in [16], [18]. For adaptive control, there are results for linear stochastic systems with an ergodic quadratic cost in [13], [14], [15]. Since the results for the discounted cost and the ergodic cost stochastic control problems for known semilinear stochastic systems are relatively recent, it appears that this is the first work on adaptive control for stochastic semilinear systems.

*Received by the editors February 11, 1999; accepted for publication (in revised form) November 29, 1999; published electronically June 15, 2000. This research was supported in part by NSF grant DMS 9623439 and GACR grant 201/98/1454.

http://www.siam.org/journals/sicon/38-6/35182.html

†Department of Mathematics, University of Kansas, Lawrence, KS 66045 (duncan@math.ukans.edu, bozenna@math.ukans.edu).

‡Institute of Mathematics, Academy of Sciences, Prague, Czech Republic (maslow@math.cas.cz).

In this paper an adaptive, ergodic cost control problem is solved for a semilinear stochastic system in an infinite dimensional state space using a solution of the HJB equation. An adaptive control is obtained from the certainty equivalence principle and the optimal control from the solution of the HJB equation. Continuity of the optimal cost with respect to the parameter is shown and the adaptive control is verified to be self-optimal.

A brief outline of the paper is given now. In section 2, the adaptive control problem is formulated and the assumptions are described. An example of a controlled stochastic parabolic partial differential equation is given for which the assumptions are satisfied. Some preparatory results for the analysis of the adaptive control problem are given. For example, the tightness of the probability laws of the solutions of the controlled semilinear systems and a uniform boundedness of their moments are verified. Some uniform bounds on the derivatives of the Markov transition semigroup for the uncontrolled system are also verified. In section 3, the parameter estimation problem is considered. In this section it is assumed that the unknown parameter appears affinely in the stochastic system. With an identifiability condition, it is shown that a family of least squares estimates is strongly consistent. In section 4, the self-optimality of an adaptive control is verified. Initially the parameter dependent, infinite dimensional HJB equations for the ergodic and discounted costs control problems are formally introduced. Their solutions are defined by the generator of an Ornstein–Uhlenbeck semigroup. A family of suitably normalized solutions of the HJB equations for the discounted costs are shown to be relatively compact for arbitrarily small discount rates and all values of the parameter in the Sobolev space $W^{1,2}(H,\mu)$, where H is the Hilbert space that is the state space for the system and μ is a limiting Gaussian measure for the solution of the associated uncontrolled linear system (Ornstein–Uhlenbeck process). This relative compactness property provides the ergodic control result as a suitable limit of the discounted control results as the discount rate tends to zero. The proof of the relative compactness uses a method in [20] for known systems in a space of continuous, polynomially bounded functions where an upper bound on the norm of the controls is required. The existence and the uniqueness of the solution of the ergodic HJB equation and its relation to an optimal control and the optimal cost are given. While the existence of the solution follows from a result in [20], the uniqueness of the solution in $W^{1,2}(H,\mu)$ is new. The continuous dependence of the solutions of the ergodic HJB equations on the parameter in the $W^{1,2}(H,\mu)$ is verified. A certainty equivalence adaptive control is defined that is based on the optimal controls from the solutions of the ergodic HJB equations and a strongly consistent family of estimates of the unknown parameter. The main results of section 4, Theorem 4.1 and Corollary 4.2, state the self-optimality of this adaptive control, that is, this adaptive control achieves the optimal ergodic cost for the true system.

2. Preliminaries. Let $(X(t),\, t \geq 0)$ be an H-valued, parameter dependent, controlled process that satisfies the stochastic differential equation

$$\begin{aligned} dX(t) &= (AX(t) + f(\alpha, X(t)) - u(t))dt + Q^{1/2}dW(t), \\ X(0) &= x, \end{aligned} \tag{2.1}$$

where H is a real, separable Hilbert space with inner product $\langle \cdot, \cdot \rangle$ and norm $|\cdot|$, $A : \mathrm{Dom}(A) \to H$ is a densely defined, unbounded linear operator on H, $f(\alpha, \cdot) : H \to H$ for each $\alpha \in \mathcal{A} \subset \mathbb{R}^q$ that is a compact set of parameters, $(W(t),\, t \geq 0)$ is a

standard, cylindrical H-valued Wiener process defined on a filtered probability space $(\Omega, \mathcal{F}, (\mathcal{F}_t), \mathbb{P})$ and $Q^{1/2} \in \mathcal{L}(H)$. The family of admissible controls is

$$\mathcal{U} = \{u : \mathbb{R}_+ \times \Omega \to B_R \mid u \text{ is measurable and } (\mathcal{F}_t) \text{ adapted}\}, \tag{2.2}$$

where $B_R = \{y \in H \mid |y| \le R\}$ and $R > 0$ is fixed. A family of Markov controls, e.g., $u(t) = \tilde{u}(X(t))$, is also considered where $\tilde{u} \in \tilde{\mathcal{U}}$ and

$$\tilde{\mathcal{U}} = \{\tilde{u} : H \to B_R \mid \tilde{u} \text{ is Borel measurable}\}. \tag{2.3}$$

The cost functionals $J(x, \lambda, u)$ and $\tilde{J}(x, u)$ are given as

$$J(x, \lambda, u) = \mathbb{E}_{x,u} \int_0^\infty e^{-\lambda t}(\psi(X(t)) + h(u(t)))dt \tag{2.4}$$

and

$$\tilde{J}(x, u) = \liminf_{T\to\infty} \mathbb{E}_{x,u} \frac{1}{T} \int_0^T (\psi(X(t)) + h(u(t)))dt, \tag{2.5}$$

where $\lambda > 0$, $h : B_R \to \mathbb{R}_+$, and $\psi : H \to \mathbb{R}$, that describe a discounted and an ergodic control problem, respectively.

The adaptive control problem is to find a family of strongly consistent estimates of the unknown parameter α and to determine an adaptive control from the family of admissible controls such that the optimal ergodic cost, $\inf_{u\in\mathcal{U}} \tilde{J}(x, u)$, is achieved.

The following assumptions are selectively used in the paper.

(A1) The linear operator $Q = Q^{1/2^*}Q^{1/2}$ is invertible, $Q^{-1} \in \mathcal{L}(H)$ and $(S(t), t \ge 0)$, where $S(t) = e^{tA}$, is an exponentially stable semigroup of contractions, that is,

$$\|S(t)\|_{\mathcal{L}(H)} \le e^{-\omega t}$$

for all $t \ge 0$ and some $\omega > 0$. Furthermore, the semigroup is Hilbert–Schmidt and there is a $\gamma > 0$ such that

$$\int_0^T t^{-\gamma} \|S(t)\|^2_{\mathrm{HS}} dt < \infty$$

for some $T > 0$, where $\|\cdot\|_{\mathrm{HS}}$ is the Hilbert–Schmidt norm.

(A2) The function $f(\alpha, \cdot) : H \to H$ is Lipschitz continuous and Gâteaux differentiable. The Gâteaux derivative $Df(\alpha, \cdot)h$ is continuous on H for each $h \in H$ and $\alpha \in \mathcal{A}$ and there is a $\beta \in \mathbb{R}$ such that

$$\langle Df(\alpha, x)h, h\rangle \le \beta|h|^2$$

for all $x \in H$, $h \in H$, and $\alpha \in \mathcal{A}$.

(A3) The function $f(\cdot, x) : \mathcal{A} \to H$ is continuous for each $x \in H$ and there are constants $p > 0$ and $C > 0$ such that

$$|f(\alpha, x)| \le C(1 + |x|^p)$$

for each $x \in H$ and $\alpha \in \mathcal{A}$.

(A4) $\psi \in C_b(H)$.

(A5) The function $h : H \to \mathbb{R}$ is convex and bounded on bounded sets and continuous. The function $\tilde{H} : H \to \mathbb{R}$ given by $\tilde{H}(x) = \sup_{|y|\le R}[\langle y, x\rangle - h(y)]$ is continuously Fréchet differentiable.

Some implications of the assumptions (A1)–(A5) are described now. Consider the linear stochastic differential equation obtained from (2.1) by choosing $f \equiv 0$ and $u \equiv 0$, that is,

$$dZ(t) = AZ(t)dt + Q^{1/2}dW(t), \quad Z(0) = x. \tag{2.6}$$

It is well known (e.g., [11]) that if (A1) is satisfied then (2.6) has a unique mild solution

$$Z(t) = S(t)x + \int_0^t S(t-r)Q^{1/2}dW(r) \tag{2.7}$$

which is an H-valued process with continuous sample paths, is ergodic, and has a unique invariant probability measure

$$\mu = N(0, Q_\infty), \tag{2.8}$$

where

$$Q_\infty = \int_0^\infty S(r)QS^*(r)dr$$

is a trace class operator on H. If (A2) is satisfied, then (2.1) has a unique mild solution

$$X(t) = S(t)x + \int_0^t S(t-r)(f(\alpha, X(r)) - u(r))dr + \int_0^t S(t-r)Q^{1/2}dW(r) \tag{2.9}$$

for each $u \in \mathcal{U}$ and $\alpha \in \mathcal{A}$. If the control in (2.1) has the feedback form $u(t) = \tilde{u}(X(t))$, where $\tilde{u} \in \tilde{\mathcal{U}}$, then the solution of (2.1) is obtained by an absolute continuity of measures as a weak solution in the probabilistic sense. More specifically, if $\mathbb{P}$ is the probability measure for the solution of (2.1) with $u \equiv 0$, then the probability measure $\mathbb{P}_{\tilde{u}}$ for the solution of

$$dX(t) = (AX(t) + f(\alpha, X(t)) - \tilde{u}(X(t)))dt + Q^{1/2}dW(t), \quad X(0) = x \tag{2.10}$$

is induced from $\mathbb{P}$ by the Radon–Nikodým derivative

$$\frac{d\mathbb{P}_{\tilde{u}}}{d\mathbb{P}} = \exp\left[-\int_0^T \langle Q^{-1/2}\tilde{u}(X(s)), dW(s)\rangle - \frac{1}{2}\int_0^T |Q^{-1/2}\tilde{u}(X(s))|^2 ds\right]. \tag{2.11}$$

The assumption (A3) is used to verify a suitable continuous dependence of the solutions of the ergodic HJB equations on the parameter which is important to verify the self-optimality of a certainty equivalence adaptive control. The assumptions (A4) and (A5) are standard conditions on a cost functional in the stochastic control of semilinear systems (e.g., [20], [22]). Note that (A5) is satisfied in the case where $h(x) = |x|^2$ so that $\tilde{H}(x) = \hat{H}(|x|)$, where

$$\hat{H}(r) = \begin{cases} \frac{r^2}{4} & \text{if } |r| \le 2R, \\ R|r| - R^2 & \text{if } |r| > 2R. \end{cases}$$

To provide some additional perspective for the adaptive control problem, an example of a stochastic partial differential equation that satisfies the assumptions is given. Consider the stochastic partial differential equation

$$\frac{\partial y}{\partial t}(t,\xi) = \frac{\partial^2 y}{\partial \xi^2}(t,\xi) + F(\alpha, y(t,\xi)) - u(t,\xi) + \eta(t,\xi) \tag{2.12}$$

for $(t,\xi) \in \mathbb{R}_+ \times (0,1)$ with the initial condition $y(0,\xi) = y_0(\xi)$, $\xi \in (0,1)$ and the Dirichlet boundary conditions

$$y(t,0) = y(t,1) = 0$$

for $t \geq 0$. The function $F : \mathcal{A} \times \mathbb{R} \to \mathbb{R}$ satisfies the following: $F(\cdot, y)$ is continuous for each $y \in \mathbb{R}$, $F(\alpha, \cdot)$ is (globally) Lipschitz continuous for each $\alpha \in \mathcal{A}$, $|F(\alpha, y)| \leq C(1 + |y|)$ for some $C > 0$ and all $\alpha \in \mathcal{A}$ and $y \in \mathbb{R}$, and $F'(\alpha, y) \leq \beta$ for some $\beta \in \mathbb{R}$, and the term η formally denotes a space time white noise. The control $(u(t),\, t \geq 0)$ is assumed to be adapted to the noise process and to take values in a ball, B_R, in $L^2(0,1)$. The formal equation (2.12) can be rigorously described in a standard way as an equation of the form (2.1) in the Hilbert space $H = L^2(0,1)$, $A = \partial^2/\partial\xi^2$, $\mathrm{Dom}(A) = \{\varphi \in L^2(0,1) \mid \varphi, \varphi' \text{ are absolutely continuous, } \varphi'' \in L^2(0,1), \varphi(0) = \varphi(1) = 0\}$, $f(\alpha, x)(\xi) = F(\alpha, x(\xi))$ for $x \in H$, $\alpha \in \mathcal{A}$, $\xi \in (0,1)$ and a cylindrical Wiener process with $Q = \delta I$ where $\delta > 0$ is a constant and I is the identity on I. For ψ and h in the cost functionals (2.4), (2.5) arbitrary $\psi \in C_b(L^2(0,1))$ and $h : L^2(0,1) \to \mathbb{R}_+$ satisfying (A5) can be chosen, e.g., $h(u) = |u|^2$. It is well known that all of the assumptions (A1)–(A5) are satisfied where $\gamma \in (0, 1/4)$ in (A1).

It is convenient to denote by $\mathbb{P}^\alpha_{x,u}$ the probability measure on Ω for the solution of (2.1) with $X(0) = x$, a control $u \in \mathcal{U}$, and a parameter $\alpha \in \mathcal{A}$. Let $\mathbb{E}^\alpha_{x,u}$ be the expectation for the probability $\mathbb{P}^\alpha_{x,u}$.

In the following proposition some stability, boundedness, and tightness properties are given.

PROPOSITION 2.1. *If* (A1) *and* (A2) *are satisfied, where* $\omega - \beta > 0$, *then the following apply.*

(i) *For each* $p' > 0$ *and* $T > 0$ *the following inequalities are satisfied:*

$$\sup_{\alpha \in \mathcal{A}} \sup_{u \in \mathcal{U}} \mathbb{E}^\alpha_{x,u} \left[\sup_{0 \leq t \leq T} |X(t)|^{2p'} \right] \leq C_T |x|^{2p'} + \tilde{C}_T \tag{2.13}$$

and

$$\sup_{t \in \mathbb{R}_+} \sup_{\alpha \in \mathcal{A}} \sup_{u \in \mathcal{U}} \mathbb{E}^\alpha_{x,u} [|X(t)|^{2p'}] \leq C_1 |x|^{2p'} + C_2. \tag{2.14}$$

(ii) *For each* $x \in H$, $\alpha \in \mathcal{A}$, *and* $u \in \mathcal{U}$ *the following inequality is satisfied:*

$$\limsup_{t \to \infty} \frac{1}{t} \int_0^t |X(s)|^2 ds < \infty \qquad \textit{almost surely (a.s.)} \ \mathbb{P}^\alpha_{x,u}. \tag{2.15}$$

(iii) *For each* $x \in H$, $\alpha \in \mathcal{A}$, *and* $u \in \mathcal{U}$ *there is a sequence* $(K_n,\, n \in \mathbb{N})$ *of compact sets in* H *such that*

$$\lim_{n \to \infty} \sup_{t \geq 1} \mathbb{P}^\alpha_{x,u}(X(t) \in H \setminus K_n) = 0, \tag{2.16}$$

so the family of measures $(\mu_{X(t)},\, t \geq 1)$ *is tight where* $\mu_{X(t)}$ *is the probability law for* $X(t)$.

Proof. To verify (i) the solution of (2.1) can be expressed as

$$X(t) = \mu(t, \tilde{Z}) + \tilde{Z}(t),$$

where

$$\tilde{Z}(t) = -\int_0^t S(t-r)u(r)dr + \int_0^t S(t-r)Q^{1/2}dW(r)$$

and $\mu(t,\varphi)$ is the solution of the deterministic integral equation

$$\mu(t,\varphi) = S(t)x + \int_0^t S(t-r)f(\alpha, \mu(r,\varphi)) + \varphi(r))dr,$$

where $\varphi \in C(\mathbb{R}_+, H)$ and $\varphi(0) = 0$. Since the semigroup $(S(t),\, t \geq 0)$ is exponentially stable, it easily follows that

$$\sup_{t\in\mathbb{R}_+} \mathbb{E}|\tilde{Z}(t)|^m < \infty \tag{2.17}$$

for each $m > 0$. Thus to verify (2.14), it suffices to show that

$$\sup \mathbb{E}^{\alpha}_{x,u}[|\mu(t,\tilde{Z})|^{2p'}] \leq C(1 + |x|^{2p'}) \tag{2.18}$$

for some constant C and the supremum is taken over all $t \in \mathbb{R}_+$, $u \in \mathcal{U}$, and $\alpha \in \mathcal{A}$. It can be assumed in (A2) that $\beta \leq 0$, for otherwise βI can be subtracted from f and added to A. Using the standard approximation of f by the sequence $(f_n,\, n \in \mathbb{N})$, where $f_n = (1/n)[(I - nf)^{-1} - I]$, and using the Gronwall lemma it follows that

$$|\mu(t,\varphi)|^{2p'} \leq e^{-2p'\omega t}|x|^{2p'} + 2p\int_0^t e^{-2p'\omega(t-s)}|f(\alpha,\varphi(s))|\ |\mu(s,\varphi)|^{2p'-1}ds \tag{2.19}$$

(cf. [20, Lemma 2.2] or [12] for a similar verification). By (A3) it follows that

$$|f(\alpha,x)| \leq \tilde{C}(1 + |x|^p) \tag{2.20}$$

for all $x \in H$ and $\alpha \in \mathcal{A}$ and some constant $\tilde{C}$. If $p' = 1/2$, then (2.17) and (2.19) imply (2.18). Otherwise a verification by induction is made as follows. Assume that

$$\sup_{t\in\mathbb{R}_+,\ \alpha\in\mathcal{A},\ u\in\mathcal{U}} \mathbb{E}^{\alpha}_{x,u}[|\mu(t,\tilde{Z})|^{2p'-(1/2)}] \leq C_3|x|^{2p'-(1/2)} + C_4$$

for all $x \in H$ and some constants C_3 and C_4. Using (2.19) and the Hölder inequality with exponents $q = [2p' - (1/2)]/(2p' - 1)$ and $q' = q/(q-1)$ it follows that

$$\begin{aligned}\mathbb{E}^{\alpha}_{x,u}[|\mu(t,\tilde{Z})|^{2p'}] &\leq e^{-2p'\omega t}|x|^{2p'} + C_5\int_0^t e^{-2p'\omega(t-s)}(\mathbb{E}^{\alpha}_{x,u}|f(\alpha,\tilde{Z}(s))|^{q'})^{1/q'}ds \\ &\qquad \cdot (\mathbb{E}^{\alpha}_{x,u}|\mu(t,\tilde{Z})|^{2p'-(1/2)})^{1/q} \\ &\leq C_6|x|^{2p'} + C_7\end{aligned}$$

for all $x \in H$ and some constants C_6 and C_7. This completes the verification of (2.14). It has been shown [30] that

$$\mathbb{E}\left[\sup_{0\leq t\leq T}|\tilde{Z}(t)|^q\right] < \infty$$

for each $q > 0$. Thus the inequality (2.13) follows from (2.19) in a similar way.

To verify (ii) the inequalities (2.19) and (2.20) imply that

$$\frac{1}{t}\int_0^t |\mu(r,\varphi)|^2 dr \le \frac{C_8(x)}{t} + C_9 \frac{1}{t}\int_0^t |\varphi(s)|^{2p} ds$$

for some $C_8(x)$ depending on $x \in H$ and constant C_9. Thus

$$\begin{aligned} \frac{1}{t}\int_0^t |X(s)|^2 ds &\le \frac{2}{t}\int_0^t |\mu(s,\tilde{Z})|^2 ds + \frac{2}{t}\int_0^t |\tilde{Z}(s)|^2 ds \\ &\le \frac{2C_8(x)}{t} + \frac{2C_9}{t}\int_0^t |\tilde{Z}(s)|^{2p} ds + \frac{2}{t}\int_0^t |\tilde{Z}(s)|^2 ds. \end{aligned} \tag{2.21}$$

Note that

$$\frac{1}{t}\int_0^t \left|\int_0^s S(s-r)u(r)dr\right|^{2q} ds \le \frac{1}{t}\int_o^t \left(\frac{R}{\omega}\right)^{2q} ds \le \left(\frac{R}{\omega}\right)^{2q} \tag{2.22}$$

and

$$\lim_{t\to\infty} \frac{1}{t}\int_0^t \left|\int_0^s S(s-r)Q^{1/2} dW(r)\right|^{2q} ds = C(q)(\operatorname{Tr} Q_\infty)^q \qquad \text{a.s. } \mathbb{P}^\alpha_{x,u} \tag{2.23}$$

for any $q \in \mathbb{N}$, and $C(q)$ depends only on q. This equality follows because the Ornstein–Uhlenbeck process $(\int_0^t S(t-r)dW(r),\ t \ge 0)$ is ergodic and by the strong law of large numbers (cf. [27]) the family of time averages in (2.23) converge almost surely to the $2q$th moment of the invariant measure, $\mu = N(0, Q_\infty)$. The inequalities (2.21) and (2.22) and the equality (2.23) verify the inequality (2.15) in (ii).

To verify (iii) the solution $(X(t),\ t \ge 0)$ of (2.1) is expressed as

$$\begin{aligned} X(t+1) &= S(1)X(t) + \int_t^{t+1} S(t+1-r)(f(\alpha, X(r)) - u(r))dr \\ &\quad + \int_t^{t+1} S(t+1-r)Q^{1/2} dW(r) \\ &= S(1)X(t) + \int_0^1 S(1-r)\lambda(\alpha, r+t)dr + \int_0^1 S(1-r)Q^{1/2} d\tilde{W}(r), \end{aligned} \tag{2.24}$$

where $\tilde{W}(r) = W(r+t)$ and

$$\lambda(\alpha, s) = f(\alpha, X(s)) - u(s).$$

Since it can be assumed that $\gamma \in (0,1)$ in (A1), let $q = 1/\gamma$ and $v \in (1/q, 1]$. Define the linear operator $J_v : L^2(0,1;H) \to H$ as

$$J_v h = \int_0^1 (1-s)^{v-1} S(1-s)h(s)ds.$$

It is well known (e.g., [11]) that J_v is a compact operator and

$$\int_0^1 S(1-r)Q^{1/2} dW(r) = \frac{\sin\frac{1}{2}\pi\gamma}{\pi} J_{\gamma/2}(Y),$$

where

$$Y(s) = \int_0^s (s-u)^{-\gamma/2} S(s-u) Q^{1/2} dW(u).$$

Thus (2.24) can be rewritten as

$$X(t+1) = S(1)X(t) + J_1\lambda(\alpha, \cdot + t) + \frac{\sin \frac{1}{2}\pi\gamma}{\pi} J_{\gamma/2}(Y)$$

for $t \geq 0$. Let $|\cdot|_q$ be the norm in $L^q(0,1;H)$. By (2.20) and (i) it follows that

$$\mathbb{E}^\alpha_{x,u} |\lambda(\alpha, \cdot + t)|_q^q \leq \mathbb{E}^\alpha_{x,u} k_1 \left(R^q + k_2 \int_0^1 |X(r+t)|^{qp} dt \right) \leq k_3 \tag{2.25}$$

and (A1) implies that

$$\mathbb{E}^\alpha_{x,u} \int_0^1 |Y(s)|^2 ds \leq k_4 \int_0^1 \left(\int_0^s (s-u)^{-\gamma} |S(s-u)|^2_{\mathrm{HS}} du \right)^2 ds \leq k_5 \tag{2.26}$$

for some constants $k_1 - k_5$ that do not depend on $t \in \mathbb{R}_+$, $\alpha \in \mathcal{A}$, and $u \in \mathcal{U}$. Define a sequence of compact sets $(K_n,\, n \in \mathbb{N})$ as

$$K_n = \{y \in H \mid y \in S(1)x + J_1 h + J_{\gamma/2} g,\ |x|^2 + |h|_q^q + |g|_q^q \leq n\}.$$

By the Chebyshev inequality it follows that

$$\mathbb{P}^\alpha_{x,u}(X(t+1) \notin K_n) \leq \frac{1}{n} \left[\mathbb{E}^\alpha_{x,u} \left(|X(t)|^2 + |\lambda(\alpha, \cdot + t)|_q^q + \left(\frac{\sin \frac{1}{2}\pi\gamma}{\pi} \right)^q |Y|_q^q \right) \right]$$

so that

$$\lim_{n\to\infty} \sup_{t \geq 1} \mathbb{P}^\alpha_{x,u}(X(t) \in H \setminus K_n) = 0$$

by (2.25), (2.26), and (i). □

In what follows let $(P_t^\alpha,\, t \geq 0)$ be the Markov transition semigroup induced by the solution of (2.1) with $u \equiv 0$. It is clear that for each $\alpha \in \mathcal{A}$, $(P_t^\alpha,\, t \geq 0)$ is a semigroup of bounded, linear operators on $C_b(H)$.

This section is concluded with some bounds on the Fréchet derivative of $P_t^\alpha \varphi$ for $t \geq 0$. For a bounded, Borel measurable function $\phi : H \to \mathbb{R}$, let $\|\phi\| = \sup_{x \in H} |\phi(x)|$ be the essential supremum.

PROPOSITION 2.2. *If* (A1) *and* (A2) *are satisfied with* $\omega - \beta > 0$, *then for each* $t > 0$, $\alpha \in \mathcal{A}$, *and* $\varphi \in C_b(H)$ *the function* $P_t^\alpha \varphi : H \to C_b(H)$ *is Fréchet differentiable and its Fréchet derivative* $DP_t^\alpha \varphi$ *satisfies the following inequalities:*

$$\|DP_t^\alpha \varphi\| \leq |Q^{-1/2}|_{\mathcal{L}(H)} t^{-1/2} \|\varphi\|, \qquad t > 0, \tag{2.27}$$

and

$$\|DP_t^\alpha \varphi\| \leq |Q^{-1/2}|_{\mathcal{L}(H)} e^{-(\omega-\beta)(t-1)} \|\varphi\|, \qquad t > 1, \tag{2.28}$$

where $\|\cdot\| = \sup_{x \in H} |\cdot|$. *From* (2.27) *and* (2.28) *the following inequality is satisfied:*

$$\|DP_t^\alpha \varphi\| \leq k(\omega_1) |Q^{-1/2}|_{\mathcal{L}(H)} t^{-1/2} e^{-\omega_1 t} \|\varphi\| \tag{2.29}$$

for $t > 0$ and each $\omega_1 \in (0, \omega - \beta)$ for a constant $k(\omega_1)$ that depends only on ω_1, ω, and β.

Proof. Let $(X^{\alpha,x}(t), t \geq 0)$ be the solution of (2.1) with $u \equiv 0$. By Proposition 7 in [10] it follows that the map $x \mapsto X^{\alpha,x}(t)$ is Gâteaux differentiable in the mean square for each $t \in \mathbb{R}_+$ and its directional derivative at x in the direction $h \in H$, $Y_h^{\alpha,x}(t)$, is a mild solution of the random linear differential equation

$$\frac{d}{dt}Y_h^{\alpha,x}(t) = (A + Df(\alpha, X^{\alpha,x}(t)))Y_h^{\alpha,x}(t),$$

$$Y_h^{\alpha,x}(0) = h.$$

Let $g_n(\alpha, x) = n(nI - A)^{-1}Df(\alpha, x)$ for $n \in \mathbb{N}$ and let $(y_n(t), t \geq 0)$ be the strong solution of the random linear differential equation

$$\frac{d}{dt}y_n(t) = (A + g_n(\alpha, X^{\alpha,x}(t)))y_n(t),$$
$$y_n(0) = n(nI - A)^{-1}h = h_n.$$

By (A1) it follows that

$$\begin{aligned}\frac{d}{dt}|y_n(t)|^2 &= 2\langle (A + g_n(\alpha, X^{\alpha,x}(t)))y_n(t), y_n(t)\rangle \\ &\leq -2\omega|y_n(t)|^2 + 2\langle g_n(\alpha, X^{\alpha,x}(t))y_n(t), y_n(t)\rangle\end{aligned}$$

so that

$$|y_n(t)|^2 \leq |h_n|^2 - 2\omega\int_0^t |y_n(s)|^2 ds + 2\int_0^t \langle g_n(\alpha, X^{\alpha,x}(s))y_n(s), y_n(s)\rangle ds.$$

Letting $n \to \infty$ and using standard properties of the Yoshida approximations and (A2) it follows that

$$|Y_k^{\alpha,x}(t)| \leq |h|^2 - 2(\omega - \beta)\int_0^t |Y_h^{\alpha,x}(s)|^2 ds$$

and

$$|Y_h^{\alpha,x}(t)|^2 \leq |h|^2 e^{-2(\omega-\beta)t}. \tag{2.30}$$

By Theorem 4.1 in [12] it follows that $DP_t^\alpha \varphi \in C_b(H)$ for $t > 0$ and

$$\langle DP_t^\alpha \varphi(x), h\rangle = \frac{1}{t}\mathbb{E}\varphi(X^{\alpha,x}(t))\int_0^t \langle Q^{-1/2}Y_h^{\alpha,x}(s), dW(s)\rangle$$

for $t > 0$ is satisfied for each $h \in H$, so by (2.30) it follows that

$$\|DP_t^\alpha \varphi\| \leq |Q^{-1/2}|_{\mathcal{L}(H)} t^{-1/2}\|\varphi\| \tag{2.31}$$

for $t > 0$ and (2.27) is verified. Furthermore, setting $\varphi_1^\alpha = P_1^\alpha \varphi$ and using the semigroup property of P_t^α it follows that

$$\begin{aligned}\langle DP_t^\alpha \varphi(x), h\rangle &= \langle DP_{t-1}^\alpha \varphi_1^\alpha(x), h\rangle \\ &= \langle D(\mathbb{E}\varphi_1^\alpha(X^{\alpha,x}(t-1)), h\rangle \\ &= \mathbb{E}\langle D\varphi_1^\alpha(X^{\alpha,u}(t-1), Y_h^{\alpha,x}(t-1)\rangle\end{aligned}$$

for $t > 1$, $h \in H$, and $x \in H$. Now using (2.30) and (2.31) with $t = 1$ it follows that

$$\|DP_t^\alpha \varphi\| \leq |Q^{-1/2}|_{\mathcal{L}(H)}\|\varphi\| e^{-(\omega-\beta)(t-1)}$$

for $t \geq 1$. ∎

3. Parameter estimation. In this section the estimation of the unknown parameter is considered where the parameter appears affinely in f, that is,

$$f(\alpha, x) = f_o(x) + \sum_{i=1}^{q} \alpha^i f_i(x), \tag{3.1}$$

where $\alpha = (\alpha^1, \dots, \alpha^q)^{\mathrm{T}}$. It is assumed that $f_0, f_1, \dots, f_q$ satisfy the relevant conditions on f in (A2) and (A3).

Let $(u(t),\, t \geq 0)$ be an admissible control and let $(X(t),\, t \geq 0)$ be the associated solution of (2.1). For notational simplicity, the dependence of X on u is suppressed. Let

$$\mathcal{A}(t) = (a_{ij}(t)) \tag{3.2}$$

and

$$\tilde{\mathcal{A}}(t) = (\tilde{a}_{ij}(t)) \tag{3.3}$$

for $t \geq 0$ be two $\mathcal{L}(\mathbb{R}^q, \mathbb{R}^q)$-valued processes, where

$$a_{ij}(t) = \int_0^t \langle Pf_i(X(s)), Pf_j(X(s))\rangle ds$$

and

$$\tilde{a}_{ij}(t) = \frac{a_{ij}(t)}{a_{ii}(t)},$$

and $P : H \to P(H)$ is a fixed finite dimensional projection on H with range in $\mathrm{Dom}(A^*)$ that is chosen to satisfy the subsequent assumptions (A6) and (A7).

For the verification of the strong consistency (e.g., [26]) of a family of least squares estimates of the unknown parameter vector, the following two assumptions are used.

(A6) For each admissible control law, the $\mathcal{L}(\mathbb{R}^q, \mathbb{R}^q)$-valued process $(\hat{\mathcal{A}}(t),\, t \geq 0)$ satisfies

$$\liminf_{t\to\infty} |\det \tilde{\mathcal{A}}(t)| > 0 \qquad \text{a.s.}$$

and

(A7) there is a $c > 0$ such that $|Pf_i(x)|^2 > c$ for $i \in \{1, \dots, q\}$ and all $x \in H$.

It is elementary to give examples where (A6) and (A7) are satisfied. For example, (A6) and (A7) are trivially satisfied if $(Pf_1, Pf_2, \dots, Pf_q)$ are nonzero, orthogonal elements for each $x \in H$ and their norms are uniformly bounded away from zero.

The estimate of the unknown parameter vector at time t, $\hat{\alpha}(t)$, is the minimizer of the quadratic functional of α, $L(t;\alpha)$, given by

$$\begin{aligned} L(t;\alpha) = &- \int_0^t \sum_{i=1}^{q} \langle P\alpha^i f_i(X(s)), dPX(s)\rangle \\ &+ \frac{1}{2}\int_0^t \sum_{i=1}^{q} |P\alpha^i f_i(X(s))|^2 ds. \end{aligned} \tag{3.4}$$

The minimizer of (3.4) is the solution of the family of linear equations

$$\mathcal{A}(t)\hat{\alpha}(t) = \mathcal{A}(t)\alpha_0 + b(t) \tag{3.5}$$

or equivalently

$$\tilde{\mathcal{A}}(t)\hat{\alpha}(t) = \tilde{\mathcal{A}}(t)\alpha_0 + \tilde{b}(t), \tag{3.6}$$

where $\mathcal{A}(t)$ and $\tilde{\mathcal{A}}(t)$ are given by (3.2) and (3.3), respectively, $b(t) = (b_1(t), \dots, b_q(t))^{\mathrm{T}}$, $\tilde{b}(t) = (\tilde{b}_1(t), \dots, \tilde{b}_q(t))^{\mathrm{T}}$,

$$b_j(t) = \int_0^t \langle Pf_i(X(s)), dPQ^{1/2}W(t)\rangle$$

and

$$\tilde{b}_j(t) = \frac{b_j(t)}{a_{jj}(t)}$$

for $j = \{1, \dots, q\}$, and α_0 is the true parameter vector.

The family of estimates $(\hat{\alpha}(t),\, t \geq 0)$ is strongly consistent as described in the following result.

THEOREM 3.1. *Let* $(u(t),\, t \geq 0)$ *be an admissible control law. If* (A1)–(A4) *and* (A6)–(A7) *are satisfied, then the family of least squares estimates* $(\hat{\alpha}(t),\, t \geq 0)$*, where* $\hat{\alpha}(t)$ *is the solution of* (3.6)*, is strongly consistent, that is,*

$$\lim_{t\to\infty} \hat{\alpha}(t) = \alpha_0 \qquad a.s., \tag{3.7}$$

where α_0 *is the true parameter vector.*

Proof. By (A7), a time change in the stochastic integrals in the components of $b(t)$, and the law of large numbers for Brownian motion, it follows that

$$\lim_{t\to\infty} \tilde{b}(t) = 0 \qquad \text{a.s.} \tag{3.8}$$

The assumption (A6) ensures that for $t \gg 0$ $\tilde{\mathcal{A}}^{-1}(t,\omega)$ exists and is bounded for almost all ω, so the equality (3.8) implies that $\hat{\alpha}(t) \to \alpha_0$ a.s. as $t \to \infty$. □

In [21], the parameter estimation of α in (2.1) is considered where f depends on α not necessarily affinely. With an identifiability condition and some other conditions, it is shown that a family of maximum likelihood estimates of α is strongly consistent. This result is a generalization of [5] to some infinite dimensional systems. The infinite dimensional setting presents some significant difficulties that either do not occur or are relatively easily overcome in finite dimensions, e.g., the application of an Itô formula without strong solutions, the tightness of a family of empirical measures, and some properties of Markov semigroups.

4. Adaptive control. In this section an adaptive control is constructed using a solution to the infinite dimensional HJB equation. This control is shown to be self-optimizing for a strongly consistent family of estimates of the unknown parameter. To verify the self-optimality property, a continuous dependence of the solutions of the HJB equations with respect to the parameter in a suitable function space is an important tool. Initially the HJB equations are introduced corresponding to the discounted and the ergodic cost functionals (2.4) and (2.5) and a summary and some modifications of some known results on these control problems are given.

The formal HJB equations corresponding to the control problems (2.1), (2.4) and (2.1), (2.5) are, respectively,

$$\begin{aligned}\frac{1}{2}\operatorname{Tr} QD^2 v_\alpha^\lambda(x) + \langle Ax, Dv_\alpha^\lambda(x)\rangle + \langle f(\alpha,x), Dv_\alpha^\lambda(x)\rangle \\ - \tilde{H}(Dv_\alpha^\lambda(x)) + \psi(x) = \lambda v_\alpha^\lambda(x),\end{aligned} \tag{4.1}$$

$$\begin{aligned}\frac{1}{2}\operatorname{Tr} QD^2 v_\alpha(x) + \langle Ax, Dv_\alpha(x)\rangle + \langle f(\alpha,x), Dv_\alpha(x)\rangle \\ - \tilde{H}(Dv_\alpha(x)) + \psi(x) = \rho(\alpha).\end{aligned} \tag{4.2}$$

In (4.2) it is necessary to solve for the pair $(v_\alpha, \rho(\alpha))$, $\rho(\alpha) \in \mathbb{R}$ for each $\alpha \in \mathcal{A}$.

The existence of strong solutions to (4.1) and (4.2) is unlikely because of the first two terms on the left-hand side of these equations, specifically because Q is not trace class and A is only densely defined in H. The approach in [20] is to replace the first two terms in (4.1) and (4.2) by the generator of an Ornstein–Uhlenbeck semigroup in a suitable function space. The results of [20], [22] are used but for simplicity the solutions of (4.1) and (4.2) are defined in a weaker sense which is suitable for the applications to adaptive control.

Let $\mu = N(0, Q_\infty)$ be the invariant measure and

$$(R_t\varphi)(x) = \mathbb{E}_x\varphi(Z(t))$$

be the Markov transition semigroup for the Ornstein–Uhlenbeck process $(Z(t),\, t \geq 0)$ that is the solution of (2.6). It is well known that $(R_t,\, t \geq 0)$ is a strongly continuous semigroup on the Hilbert space

$$\mathcal{H} = L^2(H, \mu).$$

Let $\mathcal{L}$ be the infinitesimal generator of the semigroup $(R_t,\, t \geq 0)$ in $\mathcal{H}$. Furthermore, let $\mathcal{L}_0$ be given by

$$\mathcal{L}_0\varphi(x) = \frac{1}{2}\operatorname{Tr} QD^2\varphi(x) + \langle x, A^*D\varphi(x)\rangle \tag{4.3}$$

for $x \in H$ and $\varphi \in \operatorname{Dom}(\mathcal{L}_0)$, where $\operatorname{Dom}(\mathcal{L}_0) = \{\varphi \in C_b^2(H) \mid (1/2)\operatorname{Tr} QD^2\varphi(\cdot) \in C_b(H),\ \langle \cdot, A^*D\varphi(\cdot)\rangle \in C_b(H)\}$.

Let $\varphi \in \operatorname{Dom}(\mathcal{L}_0)$ and use the Itô formula to obtain

$$\begin{aligned}(\mathcal{L}\varphi)(x) &= \lim_{t\downarrow 0}\frac{1}{t}(R_t\varphi(x) - \varphi(x)) \\ &= \lim_{t\downarrow 0}\frac{1}{t}(\mathbb{E}_x\varphi(Z(t)) - \varphi(x)) \\ &= (\mathcal{L}_0\varphi)(x)\end{aligned}$$

so $\mathcal{L}$ is a closed extension of the operator $\mathcal{L}_0$. This equality motivates the following definition of solution of (4.1) and (4.2).

DEFINITION 4.1. *A function $v_\alpha^\lambda \in \operatorname{Dom}(\mathcal{L})$ and a pair $(v_\alpha, \rho(\alpha)) \in \operatorname{Dom}(\mathcal{L}) \times \mathbb{R}$ are solutions to* (4.1) *and* (4.2)*, respectively, if*

$$\mathcal{L}v_\alpha^\lambda + \langle f(\alpha,\cdot), Dv_\alpha^\lambda\rangle - \tilde{H}(Dv_\alpha^\lambda) + \psi = \lambda v_\alpha^\lambda \tag{4.4}$$

and

$$\mathcal{L}v_\alpha + \langle f(\alpha,\cdot), Dv_\alpha\rangle - \tilde{H}(Dv_\alpha) + \psi = \rho(\alpha) \tag{4.5}$$

are satisfied.

This definition of solutions to (4.1) and (4.2) requires only that the solutions be in $\mathrm{Dom}(\mathcal{L}) \subset L^2(H,\mu)$ so (4.4) and (4.5) are understood in an $L^2(H,\mu)$ sense. This relatively weak notion of solution is used to avoid some technical complications. Some results on the solutions to (4.1) and (4.2) are given in [20] and [22]. It is shown that the solutions are more regular than that required in Definition 4.1. For the following two propositions the parameter $\alpha \in \mathcal{A}$ is fixed.

PROPOSITION 4.1. *If* (A1), (A2), (A4), *and* (A5) *are satisfied, then* (4.1) *has one and only one solution* v_α^λ *in* $\mathrm{Dom}(\mathcal{L}) \cap C_b^1(H)$. *Furthermore,*

$$v_\alpha^\lambda(x) = \inf_{u\in\mathcal{U}} J(x,\lambda,u) \tag{4.6}$$

so that v_α^λ *gives the optimal cost and an optimal control in feedback form is* $\hat{u}_\alpha^\lambda(x) = D\tilde{H}(Dv_\alpha^\lambda(x))$ *for the discounted cost control problem* (2.1), (2.4).

This proposition has been basically proven by Gozzi and Rouy [22] when $f(\alpha,\cdot)$ is bounded. The generalization in Proposition 4.2 has been done by Goldys and Maslowski [20].

The ergodic control problem is usually considered to be more difficult than the discounted control problem because the HJB equation (4.2) has an intrinsic degeneracy; that is, there is no uniqueness of the solution to (4.2) because if $(v_\alpha,\ \rho(\alpha))$ is a solution of (4.2), then $(v_\alpha + c,\ \rho(\alpha))$ for $c \in \mathbb{R}$ is also a solution. The following proposition describes results in [20] for a slightly more general problem.

PROPOSITION 4.2. *If* (A1), (A2), (A4), *and* (A5) *are satisfied, where* $\omega - \beta > 0$ *and*

$$R < \frac{\sqrt{\omega_1}}{|Q^{-1/2}|_{\mathcal{L}(H)} k(\omega_1)\sqrt{\pi}}, \tag{4.7}$$

where $\omega_1 \in (0, \omega-\beta)$ *and* $k(\omega_1) > 0$ *is the constant given in* (2.29), *then there is a unique solution* $(v_\alpha, \rho(\alpha)) \in \mathrm{Dom}(\mathcal{L}) \times \mathbb{R}_+$ *such that* $v_\alpha \in C^1(H)$, $Dv_\alpha \in C_b(H)$, *and* $v_\alpha(0) = 0$. *Furthermore,*

$$\rho(\alpha) = \inf_{u\in\mathcal{U}} \tilde{J}(x,u) \tag{4.8}$$

so that $\rho(\alpha)$ *is the optimal cost and an optimal control in feedback form is* $\hat{u}_\alpha(x) = D\tilde{H}(Dv_\alpha(x))$ *for the ergodic control problem* (2.1), (2.5)

The following result provides a relative compactness of some translates of (v_α^λ) that allows the ergodic solution v_α to be obtained as a "limit" of the discounted control problems.

PROPOSITION 4.3. *If* (A1), (A2), (A4), *and* (A5) *are satisfied, where* $\omega - \beta > 0$ *and the inequality* (4.7) *are satisfied, then the family of functions* $(\bar{v}_\alpha^\lambda;\ \alpha \in \mathcal{A},\ \lambda \in (0,1])$ *is relatively compact in the Sobolev space* $W^{1,2}(H,\mu)$, *where* $\bar{v}_\alpha^\lambda = v_\alpha^\lambda - c(\alpha,\lambda)$ *and*

$$c(\alpha,\lambda) = \frac{1}{\lambda}\int_H (\langle f(\alpha,x), Dv_\alpha^\lambda(x)\rangle - \tilde{H}(Dv_\alpha^\lambda(x)) + \psi(x))\mu(dx).$$

Proof. Let $\hat{P}_t^{\alpha,\lambda} : C_b(H) \to C_b(H)$ be the Markov transition semigroup corresponding to the solution of (2.1) using the optimal control $\hat{u}_\alpha^\lambda$, that is, $\hat{P}_t^{\alpha,\lambda}\varphi(x) = \mathbb{E}\varphi(X^{x,\alpha,\lambda}(t))$, where $(X^{x,\alpha,\lambda}(t),\, t \geq 0)$ satisfies

$$dX^{x,\alpha,\lambda}(t) = (AX^{x,\alpha,\lambda}(t) + f(\alpha, X^{x,\alpha,\lambda}(t)) - \hat{u}_\alpha^\lambda(X^{x,\alpha,\lambda}(t)))dt + Q^{1/2}dW(t),$$
$$X^{x,\alpha,\lambda}(0) = x.$$

Initially it is verified that there is a function $\gamma \in L^1(0,\infty)$ that does not depend on α or λ such that

$$\|D\hat{P}_t^{\alpha,\lambda}\varphi\| \leq \gamma(t)\|\varphi\| \tag{4.9}$$

for each $\varphi \in C_b(H)$. It is known that the function $\xi(t,x) = \hat{P}_t^{\alpha,\lambda}\varphi(x)$ is a solution to the backward Kolmogorov equation, which is defined as a mild solution, that is, $\xi(t,\cdot) \in C_b^1(H)$ for each $t > 0$ and it satisfies the integral equation

$$\xi(t,x) = P_t^\alpha\varphi(x) + \int_0^t P_{t-s}^\alpha\langle -\hat{u}_\alpha^\lambda(\cdot), D\xi(s,\cdot)\rangle(x)ds \tag{4.10}$$

for $t > 0$, where $(P_t^\alpha,\, t \geq 0)$ is the Markov transition semigroup of the solution of (2.1) with $u \equiv 0$. Using the differentiability of (4.10) in x and Proposition 2.2, it follows that

$$\begin{aligned}\|D\hat{P}_t^{\alpha,\lambda}\varphi\| &\leq k(\omega_1)|Q^{-1/2}|_{\mathcal{L}(H)}t^{-1/2}e^{-\omega_1 t}\|\varphi\| \\ &\quad + \int_0^t k(\omega_1)|Q^{-1/2}|_{\mathcal{L}(H)}(t-s)^{-1/2}e^{-\omega_1(t-s)}R\|D\hat{P}_s^{\alpha,\lambda}\varphi\|ds\end{aligned} \tag{4.11}$$

for $t > 0$, where $\omega_1 \in (0, \omega - \beta)$ and $k(\omega_1)$ is given in (2.29). Let $c = k(\omega_1)|Q^{-1/2}|_{\mathcal{L}(H)}\|\varphi\|$ and $C = k(\omega_1)|Q^{-1/2}|_{\mathcal{L}(H)}R$ so (4.11) can be written as

$$\|D\hat{P}_t^{\alpha,\lambda}\varphi\| \leq ct^{-1/2}e^{-\omega_1 t} + C\int_0^t (t-s)^{-1/2}e^{-\omega_1(t-s)}\|D\hat{P}_t^{\alpha,\lambda}\varphi\|ds.$$

By the generalized Gronwall lemma (Lemma 7.11 in [24]), if

$$\left(C\Gamma\left(\frac{1}{2}\right)\right)^2 < \theta < \omega_1, \tag{4.12}$$

where $\Gamma(\cdot)$ is the gamma function, then there is a universal constant $k > 0$ such that

$$\|D\hat{P}_t^{\alpha,\lambda}\varphi\| \leq \gamma(t)\|\varphi\|$$

for $t > 0$, where

$$\begin{aligned}\gamma(t) &= k(\omega_1)|Q^{-1/2}|_{\mathcal{L}(H)} \\ &\quad \times \left(e^{-\omega_1 t}t^{-1/2} + k\int_0^t e^{(\theta-\omega_1)(t-s)}(t-s)^{-1/2}e^{-\omega_1 s}s^{-1/2}ds\right)\end{aligned}$$

and $\gamma \in L^1(0,\infty)$ by a property of convolution from Young's inequality. The inequality (4.7) guarantees that (4.12) is satisfied for some θ so the inequality (4.9) is satisfied.

Next it is shown that there is a constant C_1 that does not depend on α or λ such that

$$\|D\bar{v}_\alpha^\lambda\| \le C_1$$

for $\lambda \in (0,1]$ and $\alpha \in \mathcal{A}$. Since v_α^λ is the optimal cost and $\hat{u}_\alpha^\lambda$ is the optimal control for the discounted control problem (2.1), (2.4) from Proposition 2.2, it follows that

$$v_\alpha^\lambda(x) = \int_0^\infty e^{-\lambda t}\hat{P}_t^{\alpha,\lambda}(\psi + h(\hat{u}_\lambda^\alpha))(x)dt$$

for each $x \in H$, $\lambda \in (0,1]$, and $\alpha \in \mathcal{A}$. Since v_α^λ and $\bar{v}_\alpha^\lambda$ differ only by a constant, it follows from (4.9) that

$$\begin{aligned}\|D\bar{v}_\alpha^\lambda\| &\le \int_0^\infty \|D\hat{P}_t^{\alpha,\lambda}(\psi + h(\hat{u}_\lambda^\alpha))\|dt \\ &\le \left(\|\varphi\| + \sup_{|y|\le R}|h(y)|\right)\int_0^\infty \gamma(t)dt := C_1. \end{aligned} \tag{4.13}$$

Let $Q_t = \int_0^t S(r)QS^*(r)dr$. Clearly Q_t is the covariance operator of the (Gaussian) probability law of $Z(t)$, where $(Z(t),\, t \ge 0)$ is the Ornstein–Uhlenbeck process that satisfies (2.6). It is known that $(Z(t),\, t \ge 0)$ is strongly Feller, that is, $\text{Range}(S(t)) \subset \text{Range}(Q_t^{1/2})$ for $t > 0$ so the operator $\Gamma(t) = Q_t^{-1/2}S(t) \in \mathcal{L}(H)$ [11]. Furthermore, it follows from [8] and [20] that

$$\int_0^\infty |\Gamma(t)|_{\mathcal{L}(H)}dt < \infty. \tag{4.14}$$

The resolvent of the infinitesimal generator $\mathcal{L}$ can be expressed as

$$(\lambda I - \mathcal{L})^{-1}\varphi = \int_0^\infty e^{-\lambda t}R_t\varphi dt$$

for $\varphi \in \mathcal{H}$ so the solution v_α^λ of (4.4) satisfies the integral equation

$$v_\alpha^\lambda = \int_0^\infty e^{-\lambda t}R_t(\langle f(\alpha,\cdot), Dv_\alpha^\lambda\rangle - \tilde{H}(Dv_\alpha^\lambda) + \psi)dt.$$

Since the inequality

$$|DR_t\varphi|_{L^2(H,\mu;H)} \le |\Gamma(t)|_{\mathcal{L}(H)}|\varphi|_{\mathcal{H}} \tag{4.15}$$

is satisfied for $t > 0$, by the Cameron–Martin formula (Lemma 3 in [7]) the integral operators

$$T_\lambda = \int_0^\infty e^{-\lambda t}DR_t dt$$

for $\lambda \in [0,1]$ converge in the $\mathcal{L}(\mathcal{H}, L^2(H,\mu;H))$ norm. Since for each $t > 0$ the linear operator $DR_t : \mathcal{H} \to L^2(H,\mu;H)$ is compact [7], the operators T_λ are also compact. Thus

$$D\bar{v}_\alpha^\lambda = Dv_\alpha^\lambda = T_\lambda\xi_\alpha^\lambda,$$

where $T_\cdot : [0,1] \to \mathcal{L}(\mathcal{H}, L^2(H,\mu;H))$ is continuous and

$$\xi_\alpha^\lambda := \langle f(\alpha,\cdot), Dv_\alpha^\lambda\rangle - \tilde{H}(Dv_\alpha^\lambda) + \psi$$

for $\alpha \in \mathcal{A}$ and $\lambda \in (0,1]$ are uniformly bounded in $\mathcal{H}$ by (4.13). It follows that the family $(D\bar{v}_\alpha^\lambda,\ \alpha \in \mathcal{A},\ \lambda \in (0,1])$ is relatively compact in $L^2(H,\mu;H)$. Note that

$$\bar{v}_\alpha^\lambda = \int_0^\infty e^{-\lambda t} R_t(\xi_\alpha^\lambda - \lambda c(\lambda,\alpha))dt, \tag{4.16}$$

where

$$\int_H (\xi_\alpha^\lambda - \lambda c(\lambda,\alpha))d\mu = 0. \tag{4.17}$$

Furthermore, by [7] the semigroup is compact in $\mathcal{H}$ and there is a $\lambda_1 > 0$ such that $|R_t\varphi|_{\mathcal{H}} \le e^{-\lambda_1 t}|\varphi|_{\mathcal{H}}$ for each $\varphi \in \mathcal{H}$ satisfying $\int \varphi d\mu = 0$. Using (4.16) and (4.17), v_α^λ can replace $D\bar{v}_\alpha^\lambda$ above to verify that $(v_\alpha^\lambda,\ \alpha \in \mathcal{A},\ \lambda \in (0,1])$ is relatively compact in $\mathcal{H}$. □

A result on the approximation of functions in $\mathrm{Dom}(\mathcal{L})$ by functions in $\mathrm{Dom}(\mathcal{L}_0)$ is given. Similar results in different spaces have been given (e.g., [6], [20], [22]). Let $\mathcal{W} \subset W^{1,2}(H,\mu)$ be given by

$$\begin{aligned}\mathcal{W} = \{\varphi \in W^{1,2}(H,\mu)\ |\varphi \in \mathrm{Dom}(\mathcal{L}),\ \|D\varphi\| < \infty,\\ |\varphi(x)| + |\mathcal{L}\varphi(x)| \le k(1+|x|^q) \text{ for all } x \in H\\ \text{and some real numbers } k \text{ and } q\}.\end{aligned}$$

LEMMA 4.1. *If $\varphi \in \mathcal{W}$, then there is a sequence $(\varphi_n, n \in \mathbb{N})$ such that $\varphi_n \in \mathrm{Dom}(\mathcal{L}_0)$, $\varphi_n \in \mathcal{W}$ for fixed k and q defining $\mathcal{W}$ for all $n \in \mathbb{N}$, $\sup_n \|D\varphi_n\| < \infty$, and*

$$\lim_{n\to\infty} |\varphi_n - \varphi|_{\mathcal{H}} = 0, \tag{4.18}$$

$$\lim_{n\to\infty} |\mathcal{L}_0\varphi_n - \mathcal{L}\varphi|_{\mathcal{H}} = 0, \tag{4.19}$$

$$\lim_{n\to\infty} |D\varphi_n - D\varphi|_{L^2(H,\mu;H)} = 0. \tag{4.20}$$

Proof. Choose $\lambda > 0$ and fix it. Let $\xi = \lambda\varphi - \mathcal{L}\varphi$. Since the inequality

$$|\xi(x)| \le k(1+|x|^q) \tag{4.21}$$

for all $x \in H$ is satisfied, there is a sequence $(\xi_n, n \in \mathbb{N})$ such that $\xi_n \in \mathrm{Dom}(\mathcal{L}_0)$ and ξ_n satisfies (4.21) for all $n \in \mathbb{N}$, $\xi_n \to \xi$ in $\mathcal{H}$ as $n \to \infty$, and $\sup_n \|D\xi_n\| < \infty$. A similar construction is given in Lemma 4.5 in [20]. Now let

$$\varphi_n = \int_0^\infty e^{-\lambda t} R_t\xi_n dt. \tag{4.22}$$

Since $(\lambda I - \mathcal{L})^{-1} : \mathrm{Dom}(\mathcal{L}_0) \to \mathrm{Dom}(\mathcal{L}_0)$ and $\varphi_n = (\lambda I - \mathcal{L})^{-1}\xi_n$, it follows that $\varphi_n \in \mathrm{Dom}(\mathcal{L}_0)$. Furthermore,

$$|\varphi_n - \varphi|_{\mathcal{H}} \le \int_0^\infty e^{-\lambda t}|R_t(\xi_n - \xi)|_{\mathcal{H}} dt \le |\xi_n - \xi|_{\mathcal{H}} \int_0^\infty e^{-\lambda t} dt$$

and

$$|D\varphi_n - D\varphi|_{L^2(H,\mu;H)} \le \int_0^\infty e^{-\lambda t}|DR_t(\xi_n - \xi)|_{L^2(H,\mu;H)}dt$$
$$\le |\xi_n - \xi|_{\mathcal{H}} \int_0^\infty e^{-\lambda t}|\Gamma(t)|_{\mathcal{L}(H)}dt.$$

This latter inequality follows from (4.15) and the convergence of the right-hand side to zero as $n \to \infty$ follows from (4.14) and the definition of $(\xi_n,\, n \in \mathbb{N})$. It can be verified directly that

$$\|DR_t\xi_n\| = \left\|D\mathbb{E}\xi_n\left(S(t)x + \int_0^t S(t-r)Q^{1/2}dW(r)\right)\right\| \le \|D\xi_n\|$$

for $t > 0$ so that

$$\sup_n \|D\varphi_n\| \le \frac{1}{\lambda}\sup_n \|D\xi_n\| < \infty.$$

Since $\mathcal{L}_0\varphi_n = \lambda\varphi_n - \xi_n$ it follows that $\mathcal{L}_0\varphi_n = \mathcal{L}\varphi_n \to y$ on $\mathcal{H}$ as $n \to \infty$ for some $y \in \mathcal{H}$ because $\mathcal{L}$ is a closed operator. Thus, $y = \mathcal{L}\varphi$. The uniform polynomial bounds on φ_n and thereby on $\mathcal{L}_0\varphi_n$ follow from the same bounds on ξ_n and (4.22). □

The following proposition is a "$W^{1,2}(H,\mu)$-version" of a result in [20] on the existence and the uniqueness of a solution to the ergodic HJB equation which is described in Proposition 4.2. While the existence result is weaker than the one in Proposition 4.2, the family of solutions for uniqueness is enlarged. The parameter $\alpha \in \mathcal{A}$ in the following proposition is fixed.

PROPOSITION 4.4. *If* (A1), (A2), (A4), (A5) *with* $\omega - \beta > 0$, *and* (4.7) *are satisfied, then there is a unique solution* $(v_\alpha, \rho(\alpha)) \in (\mathcal{W} \cap C(H)) \times \mathbb{R}$ *of* (4.2) *such that* $v_\alpha(0) = 0$. *Furthermore, the equality* (4.8) *is satisfied and an optimal control in feedback form is* $\hat{u}_\alpha(x) = D\tilde{H}(Dv_\alpha(x))$ *for the ergodic control problem* (2.1), (2.5).

Proof. The existence of the solution $(v_\alpha, \rho(\alpha))$ with the required properties including (4.8) and an optimal feedback control follow from a result in [20] that is given here as Proposition 4.2. However, it should be noted that the existence of a solution is a simple consequence of Proposition 4.3. Since $(|\lambda c(\alpha,\lambda)|,\, \lambda \in (0,1])$ is uniformly bounded, there is a sequence $(\lambda_n,\, n \in \mathbb{N})$ such that $\lambda_n \downarrow 0$ and $(\bar{v}_\alpha^{\lambda_n}, \lambda_n c(\alpha,\lambda_n)) \to (\bar{v}_\alpha, \delta)$ in $W^{1,2}(H,\mu) \times \mathbb{R}$ for some $(\bar{v}_\alpha, \delta) \in W^{1,2}(H,\mu) \times \mathbb{R}$. Letting $\lambda_n \downarrow 0$ in (4.1) and using the closedness of $\mathcal{L}$ in $\mathcal{H}$ it follows that $\bar{v}_\alpha \in \mathrm{Dom}(\mathcal{L})$ and

$$\mathcal{L}\bar{v}_\alpha + \langle f(\alpha,\cdot), D\bar{v}_\alpha\rangle - \tilde{H}(D\bar{v}_\alpha) + \psi = \delta$$

is satisfied.

Now let $(\bar{v}, \bar{\delta}) \in (\mathcal{W} \cap C(H)) \times \mathbb{R}$ be a solution of (4.2) satisfying $\bar{v}(0) = 0$. To verify uniqueness, it suffices to show that $(\bar{v}, \bar{\delta}) = (v_\alpha, \rho(\alpha))$, where $(v_\alpha, \rho(\alpha))$ is the solution of (4.2) whose existence is given by Proposition 4.2.

Initially it is shown that

$$\bar{\delta} = \rho(\alpha). \tag{4.23}$$

This verification is analogous to the corresponding part of the proof of Proposition 4.2 that is given in [20]. Let $(\bar{v}_n,\, n \in \mathbb{N})$ be a sequence such that $\bar{v}_n \in \mathrm{Dom}(\mathcal{L}_0)$ for each $n \in \mathbb{N}$ and

$$\bar{v}_n \to \bar{v}, \quad \mathcal{L}_0\bar{v}_n \to \mathcal{L}\bar{v} \qquad \text{in } \mathcal{H}, \tag{4.24}$$

$$\sup_n \|D\bar{v}_n\| < \infty, \quad D\bar{v}_n \to D\bar{v} \qquad \text{in } L^2(H,\mu;H), \tag{4.25}$$

and $\bar{v}_n$ and $\mathcal{L}_0\bar{v}_n$ are uniformly, polynomially bounded (cf. Lemma 4.1). Clearly, the pair $(\bar{v}_n, \bar{\delta})$ satisfies the equation

$$\mathcal{L}_0\bar{v}_n + \langle f(\alpha,\cdot), D\bar{v}_n\rangle - \tilde{H}(D\bar{v}_n) + \psi_n = \bar{\delta},$$

where

$$\psi_n = \bar{\delta} + \tilde{H}(D\bar{v}_n) - \langle f(\alpha,\cdot), D\bar{v}_n\rangle - \mathcal{L}_0\bar{v}_n.$$

Apply the Itô formula using the function $-\bar{\delta}t + \bar{v}_n(x)$ and the process that is the solution of (2.1) with the control $D\tilde{H}(D\bar{v}_n)$ to show that $\bar{\delta}$ is the optimal cost for the control problem (2.1) and (2.5), where ψ is replaced by ψ_n in (2.5). Since $\psi_n \to \psi$ (at least) pointwise by (4.24) and (4.25) and the sequence $(\psi_n,\, n \in \mathbb{N})$ is uniformly polynomially bounded, the limit as $n \to \infty$ can be taken to show that $\bar{\delta}$ is the optimal cost for the control problem (2.1) and (2.5) so that (4.14) is verified.

It remains to show that $\bar{v} = v_\alpha$. Let $(v_n,\, n \in \mathbb{N})$ be a sequence such that

$$v_n \to v_\alpha, \quad \mathcal{L}_0 v_n \to \mathcal{L} v_\alpha \qquad \text{in } \mathcal{H}, \tag{4.26}$$

$$\sup_n \|Dv_n\| < \infty. \quad Dv_n \to Dv_\alpha \qquad \text{in } L^2(H,\mu;H), \tag{4.27}$$

and v_n and $\mathcal{L}_0 v_n$ are uniformly, polynomially bounded using the notation of Lemma 4.1. Recall that $\alpha \in \mathcal{A}$ is fixed so the dependence of v_n on α is suppressed. Let $\bar{u}_n = D\tilde{H}(Dv_n)$ and $\bar{u} = D\tilde{H}(Dv_\alpha)$ be controls. For an arbitrary $\Phi \in C([0,T],H)$ it easily follows from (4.27) that

$$\begin{aligned}
&|\mathbb{E}_{\tau,x,\bar{u}_n}\Phi(X(\cdot)) - \mathbb{E}_{\tau,x,\bar{u}}\Phi(X(\cdot))| \\
&\quad \le \mathbb{E}_{\tau,x}\Phi(X(\cdot))\left|\exp\left(-\int_0^T \langle Q^{-1/2}\bar{u}_n(X(s)), dW(s)\rangle \right.\right. \\
&\qquad\qquad \left. - \frac{1}{2}\int_0^T |Q^{-1/2}\bar{u}_n(X(s))|^2 ds\right) \\
&\qquad \left. - \exp\left(-\int_0^T \langle Q^{-1/2}\bar{u}(X(s)), dW(s)\rangle - \frac{1}{2}\int_0^T |Q^{-1/2}\bar{u}(X(s))|^2 ds\right)\right| \\
&\quad \to 0 \qquad \text{as } n \to \infty
\end{aligned}$$

for $\tau \in [0,T)$, where $\mathbb{E}_{\tau,x,u}$ and $\mathbb{E}_{\tau,x}$ denote the expectations associated with the initial condition $X(\tau) = x$ and $(X(t),\, t \ge \tau)$ is the solution to (2.1) with $X(\tau) = x$ and $u \in \mathcal{U}$ and $u \equiv 0$, respectively. By the Skorokhod theorem there is a probability space $(\tilde{\Omega}, \tilde{\mathcal{F}}, \tilde{\mathbb{P}})$ and stochastic processes $(X_n(t), t \ge \tau)$ for $n \in \mathbb{N}$ and $(X_0(t),\, t \ge \tau)$ such that $X_n(\cdot)$ and $X_0(\cdot)$ have the probability laws on $C([\tau,\infty);H)$ that are identical with the solutions of (2.1) with the feedback controls $\bar{u}_n(X_n)$ and $\bar{u}(X_0)$, respectively, and

$$\lim_{n\to\infty} \sup_{s\in[\tau,T]} |X_n(s) - X_0(s)| = 0 \qquad \text{a.s. } \tilde{\mathbb{P}}$$

for each $T > \tau$. By the definition of $\tilde{H}$ in (A5) it follows that

$$\begin{aligned} 0 &= \tilde{H}(Dv_\alpha(x)) + h(\bar{u}(x)) - \langle \bar{u}(x), Dv_\alpha(x)\rangle \\ &\le \tilde{H}(D\bar{v}(x)) + h(\bar{u}(x)) - \langle \bar{u}(x), D\bar{v}(x)\rangle \end{aligned}$$

so that

$$\begin{aligned} &\mathcal{L}v_\alpha(x) + \langle f(\alpha,x), Dv_\alpha(x)\rangle - \langle \bar{u}(x), Dv_\alpha(x)\rangle \\ &\quad \le \mathcal{L}\bar{v}(x) + \langle f(\alpha,x), D\bar{v}(x)\rangle - \langle \bar{u}(x), D\bar{v}(x)\rangle, \end{aligned}$$

and therefore

$$\begin{aligned} &\mathcal{L}v_n(x) + \langle f(\alpha,x), Dv_n(x)\rangle - \langle \bar{u}_n(x), Dv_n(x)\rangle \\ &\quad \le \mathcal{L}_0\bar{v}_n(x) + \langle f(\alpha,x), D\bar{v}_n(x)\rangle - \langle \bar{u}_n(x), D\bar{v}_n(x)\rangle + \delta_n(x) \end{aligned}$$

for all $x \in H$, where $\delta_n : H \to \mathbb{R}$ converges to 0 as $n \to \infty$ in $L^2(H,\mu)$ and $(\delta_n,\ n \in \mathbb{N})$ is uniformly, polynomially bounded. Using the Itô formula it follows that the processes

$$\Psi_n(t) = v_n(X_n(t)) - \bar{v}_n(X_n(t)) - \int_\tau^t \delta_n(X_n(s))ds$$

for $t \ge \tau$ satisfy the inequality

$$\tilde{\mathbb{E}}\Psi_n(t) \le \Psi_n(\tau) = v_n(x) - \bar{v}_n(x),$$

where $\tilde{\mathbb{E}}$ is the expectation in $(\tilde{\Omega}, \tilde{\mathcal{F}}, \tilde{\mathbb{P}})$. Note that the probability laws for $X_n(s)$ for each $s \in \mathbb{R}_+$ and $n \in \mathbb{N} \cup \{0\}$ are mutually absolutely continuous with μ because by the Girsanov theorem any of these measures is equivalent to the Gaussian measure $N(S(t)x, Q_t)$ (the law for $Z(t)$ from the solution of (2.6)) and $N(S(t)x, Q_t)$ is equivalent to μ from the strong Feller property for each $t > 0$ and $x \in H$ (e.g., [27]). It follows that

$$\lim_{n\to\infty} \tilde{\mathbb{E}}\Psi_n(t) = \tilde{\mathbb{E}}(v_\alpha(X_0(t)) - \bar{v}(X_0(t)))$$

and thus

$$\mathbb{E}_{\tau,x,\bar{u}}(v_\alpha(X(t)) - \bar{v}(X(t))) \le v_\alpha(x) - \bar{v}(x)$$

for $t \ge \tau$ and $x \in H$. Since $\tau \in [0,t]$ and $x \in H$ are arbitrary, it follows that the process $(v_\alpha(X(t)) - \bar{v}(X(t)),\ t \ge 0)$, where $(X(t),\ t \ge 0)$ satisfies (2.1) with $u(t) = \bar{u}(X(t))$, is a supermartingale.

Furthermore,

$$\sup_{t\ge 0} \mathbb{E}_{x,\bar{u}} |v_\alpha(X(t)) - \bar{v}(X(t))| \le k \sup_{t\ge 0} \mathbb{E}(1 + |X(t)|^q) < \infty$$

for some positive constants k and q by Proposition 2.1 (i). Thus, there is a limit of $(v_\alpha(X(t)) - \bar{v}(X(t)),\ t \ge 0)$ a.s. $\mathbb{P}^\alpha_{x,\bar{u}}$ as $t \to \infty$. Since the solution of (2.1) is ergodic for these feedback controls, for each ball B in H there is a sequence of random times $(\sigma_n,\ n \in \mathbb{N})$ that increase to infinity such that $X(\sigma_n) \in B$ (cf. [29]). Since $v_\alpha(0) - \bar{v}(0) = 0$ and $v_\alpha - \bar{v} : H \to \mathbb{R}$ is a continuous function, letting $B = B_{1/n}$ for $n \in \mathbb{N}$ it follows that

$$\liminf_{t\to\infty} (v_\alpha(X(t)) - \bar{v}(X(t))) = 0 \qquad \text{a.s. } \mathbb{P}^\alpha_{x,\bar{u}}$$

so that

$$\lim_{t\to\infty}(v_\alpha(X(t)) - \bar{v}(X(t))) = 0 \qquad \text{a.s. } \mathbb{P}^\alpha_{x,\bar{u}}. \tag{4.28}$$

If there is a $y \in H$ such that $v_\alpha(y) - \bar{v}(y) > 0$, then $v_\alpha - \bar{v} > \Delta > 0$ in an open ball B_y centered at y. Choosing $B = B_y$ above gives a contradiction to (4.28) so that $v_\alpha \equiv \bar{v}$. □

COROLLARY 4.1. *If* (A1)–(A5) *where* $\omega - \beta > 0$ *are satisfied and* (4.7) *is satisfied, then the function* $v_\cdot : \mathcal{A} \to W^{1,2}(H,\mu)$ *given by* $\alpha \mapsto v_\alpha$ *is continuous and*

$$\|Dv_\alpha\| < C_1 \tag{4.29}$$

for all $\alpha \in \mathcal{A}$ *and some constant* $C_1 > 0$ *that does not depend on* $\alpha \in \mathcal{A}$ *where* v_α *is the unique solution of the ergodic HJB equation given in Proposition* 4.4.

Proof. Let $\bar{v}_\alpha$ be the first component of the solution of the ergodic HJB which is given in the proof of Proposition 4.4 as a limit of $\bar{v}^\lambda_\alpha$ in $W^{1,2}(H,\mu)$ as $\lambda \downarrow 0$. By (4.13) it follows that

$$\|Dv_\alpha\| = \|D\bar{v}_\alpha\| \le C_1$$

because by the uniqueness part of Proposition 4.4 v_α and $\bar{v}_\alpha$ differ only by a constant. Clearly the family $(\bar{v}_\alpha,\ \alpha \in \mathcal{A})$ is relatively compact in $W^{1,2}(H,\mu)$ by Proposition 4.3. Using this relative compactness and the closedness of the derivative operator for any sequence $(\bar{v}_{\alpha_n},\ n \in \mathbb{N})$ such that $\alpha_n \to \alpha_0$ as $n \to \infty$ there is a subsequence again denoted $(\bar{v}_{\alpha_n},\ n \in \mathbb{N})$ such that $\bar{v}_{\alpha_n} \to \tilde{v}$ and $D\bar{v}_{\alpha_n} \to D\tilde{v}$ a.e. μ for some $\tilde{v}$. Furthermore, the sequence $(\rho(\alpha_n),\ n \in \mathbb{N})$ is uniformly bounded by (4.13) and Proposition 4.4 and there is a subsequence again denoted by $(\rho(\alpha_n),\ n \in \mathbb{N})$ such that $\rho(\alpha_n) \to \tilde{\rho}$ for some $\tilde{\rho} \in \mathbb{R}$. Since $\mathcal{L}\bar{v}_{\alpha_n} + \langle f(\alpha_n,\cdot), D\bar{v}_n\rangle - \tilde{H}(D\bar{v}_{\alpha_n}) + \psi = \rho(\alpha_n)$ and the operator $\mathcal{L}$ is closed in $\mathcal{H}$, it follows by (A3) that the limit as $n \to \infty$ is

$$\mathcal{L}\tilde{v} + \langle f(\alpha,\cdot), D\tilde{v}\rangle - \tilde{H}(D\tilde{v}) + \psi = \rho \qquad \text{a.e. } \mu.$$

By (4.29), $\tilde{v}$ is continuous and $D\tilde{v}$ is bounded so that $\tilde{v} \in \mathcal{W}$. By the uniqueness part of Proposition 4.4 it follows that $\tilde{v} - \tilde{v}(0) = v_{\alpha_0}$. By (4.28) and the (μ) almost sure convergence $\bar{v}_{\alpha_n} \to \tilde{v}$ as $n \to \infty$ it follows that $\bar{v}_{\alpha_n}(0) \to \tilde{v}(0)$. Therefore, $v_{\alpha_n} = \bar{v}_{\alpha_n} - \tilde{v}_{\alpha_n}(0) \to \tilde{v} - \tilde{v}(0) = v_{\alpha_0}$ as $n \to \infty$. □

For a certainty equivalence adaptive control and a consistent family of estimators of the unknown parameter vector, it is shown that this control is self optimizing. Consider (2.1) with the true parameter value $\alpha_0 \in \mathcal{A}$, that is,

$$\begin{aligned} dX(t) &= (AX(t) + f(\alpha_0, X(t)) - \tilde{u}(t))dt + Q^{1/2}dW(t), \\ X(0) &= x, \end{aligned}$$

where

$$\tilde{u}(t) = D\tilde{H}(Dv_{\alpha(t)}(X(t))), \tag{4.30}$$

where $(\alpha(t),\ t \ge 0)$ is an adapted, measurable, $\mathcal{A}$-valued process satisfying

$$\lim_{t\to\infty}\alpha(t) = \alpha_0 \tag{4.31}$$

in probability $\mathbb{P}$ and v_α is given in Proposition 4.4.

The following theorem and its corollary provide solutions to an adaptive control problem described by (2.1) and (2.5).

THEOREM 4.1. *Given the adaptive control problem described by* (2.1) *and* (2.5), *where* $\alpha_0 \in \mathcal{A}$ *is the true parameter vector, if* (A1)–(A5) *are satisfied,* $\omega - \beta > 0$, *the inequality* (4.7), *is satisfied, and the* $\mathcal{A}$*-valued family of estimates of* α_0 $(\alpha(t),\, t \geq 0)$, *satisfies* (4.31), *then*

$$\tilde{J}(x, \tilde{u}) = \rho(\alpha_0) \tag{4.32}$$

for each $x \in H$, *where* $\tilde{u}$ *is the adaptive control given by* (4.30).

Proof. For notational simplicity, let $v = v_{\alpha_0}$ and $\rho = \rho(\alpha_0)$. Let $v_n \in \mathrm{Dom}(\mathcal{L}_0)$ for $n \in \mathbb{N}$ be a sequence that converges to v as in (4.26), (4.27). The Itô formula applied to $(v_n(X(t)) - \rho t,\, t \geq 0)$ yields the equation

$$\begin{aligned} &- \rho t + \mathbb{E}_{x,\tilde{u}} v_n(X(t)) \\ &\quad = v_n(x) + \mathbb{E}_{x,\tilde{u}} \int_0^t [-\rho + \mathcal{L}_0 v_n(X(s)) + \langle f(\alpha_0, X(s)), Dv_n(X(s))\rangle \\ &\qquad\qquad - \langle \tilde{u}(s), Dv_n(X(s))\rangle] ds. \end{aligned}$$

Since v_n satisfies

$$\mathcal{L}_0 v_n + \langle f(\alpha_0, \cdot), Dv_n\rangle - \tilde{H}(Dv_n) + \psi_n = \rho,$$

where

$$\psi_n = \rho + \tilde{H}(Dv_n) - \langle f(\alpha_0, \cdot), Dv_n\rangle - \mathcal{L}_0 v_n,$$

it follows that

$$\begin{aligned} &- \rho t + \mathbb{E}_{x,\tilde{u}} v_n(X(t)) \\ &\quad = v_n(x) + \mathbb{E}_{x,\tilde{u}} \int_0^t [-\psi_n(X(s)) + \tilde{H}(Dv_n(X(s))) - \langle \tilde{u}(s), Dv_n(X(s))\rangle] ds. \end{aligned}$$

By the definition of $\tilde{H}$ in (A5) it follows that

$$-\tilde{H}(Dv_n(X(s))) - h(\tilde{u}(X(s))) + \langle \tilde{u}_n(X(s)), Dv_n(X(s))\rangle = 0,$$

where

$$\tilde{u}_n(x) = D\tilde{H}(Dv_n(x)).$$

Therefore,

$$\begin{aligned} \rho t &= \mathbb{E}_{x,\tilde{u}} v_n(X(t)) - v_n(x) \\ &\quad + \mathbb{E}_{x,\tilde{u}} \int_0^t [\psi_n(X(s)) + h(\tilde{u}_n(X(s))) + \langle \tilde{u}(s) - \tilde{u}_n(X(s)), Dv_n(X(s))\rangle] ds. \end{aligned}$$

Using the properties of convergence $v_n \to v$ in (4.26), (4.27), the uniform polynomial bound on the sequence $(v_n,\, n \in \mathbb{N})$ and Proposition 2.1 (i), the passage to the limit $(n \to \infty)$ yields the equation

$$\begin{aligned} \rho t &= \mathbb{E}_{x,\tilde{u}} v(X(t)) - v(x) \\ &\quad + \mathbb{E}_{x,\tilde{u}} \int_0^t [\psi(X(s)) + h(\tilde{u}_0(X(s))) + \langle \tilde{u}(s) - \tilde{u}_0(X(s)), Dv(X(s))\rangle] ds, \end{aligned} \tag{4.33}$$

where $\tilde{u}_0(x) = D\tilde{H}(Dv(x))$. By Proposition 2.1 (i) it follows that

$$\lim_{t\to\infty} \frac{1}{t}\mathbb{E}_{x,\tilde{u}} v(X(t)) = 0$$

for each $x \in H$. By an interpolation result (Lemma 4.6 in [22]) it follows that

$$\|Dv_\alpha^\lambda\|_{r,\theta} \le c_1 + c_2\|Dv_\alpha^\lambda\| + c_3 \sup_{x\in B_r} |f(\alpha,x)|\|Dv_\alpha^\lambda\| \tag{4.34}$$

for some universal constants c_1, c_2, c_3 and a suitable constant $\theta > 0$, where $\|\cdot\|_{r,\theta}$ denotes the norm of Hölder continuous functions with exponent θ on the ball B_r. The result in [22] is stated for $r = \infty$ and f bounded but (4.34) follows directly if f is truncated outside of B_r and using the fact that $\mathcal{L}$ is a local operator so the truncation does not affect the solution v_α^λ in B_r. By (A3) and (4.13) it follows that $\|Dv_\alpha^\lambda\|_{r,\theta} \le c_3(1 + r^p)$ for some constant c_3 that does not depend on α or λ. Thus by the proof of Corollary 4.1 it follows that

$$\lim_{\alpha\to\alpha_0} Dv_\alpha = Dv_{\alpha_0} \tag{4.35}$$

uniformly on compact sets. For a compact set $K \subset H$ if follows that

$$\begin{aligned}
&\mathbb{E}_{x,\tilde{u}}|D\tilde{H}(Dv_{\alpha_0}(X(s))) - D\tilde{H}(Dv_{\alpha(t)}(X(s)))| \\
&\quad = \mathbb{E}_{x,\tilde{u}}|D\tilde{H}(Dv_{\alpha_0}(X(s))) - D\tilde{H}(Dv_{\alpha(t)}(X(s)))|1_{\{X(s)\in K\}} \\
&\quad\quad + \mathbb{E}_{x,\tilde{u}}|D\tilde{H}(Dv_{\alpha_0}(X(s))) - D\tilde{H}(Dv_{\alpha(t)}(X(s)))|1_{\{X(s)\notin K\}}.
\end{aligned} \tag{4.36}$$

The second term on the right-hand side of (4.36) is bounded above by

$$2R\mathbb{P}_{x,\tilde{u}}(X(s) \in H \setminus K)$$

which can be made arbitrarily small for a compact set K sufficiently large by Proposition 2.1 (iii) for $s \ge 1$ and R is given in (2.2). The first term on the right-hand side of (4.36) tends to zero as $s \to \infty$ by the locally uniform convergence in (4.35). These two facts imply that

$$\lim_{s\to\infty} \mathbb{E}_{x,\tilde{u}}|D\tilde{H}(Dv_{\alpha_0}(X(s))) - D\tilde{H}(Dv_{\alpha(s)}(X(s)))| = 0. \tag{4.37}$$

By a similar argument it follows that

$$\lim_{s\to\infty} \mathbb{E}_{x,\tilde{u}}|h(D\tilde{H}(Dv_{\alpha_0}(X(s)))) - h(D\tilde{H}(Dv_{\alpha(s)}(X(s))))| = 0. \tag{4.38}$$

Dividing (4.33) by t and letting $t \to \infty$, the equality (4.32) follows by (4.37) and (4.38). This completes the proof. □

COROLLARY 4.2. *Given the adaptive control problem described by* (2.1) *and* (2.5), *where* $\alpha_0 \in \mathcal{A}^0$, *is the true parameter vector and* f *satisfies* (3.1). *Let* $(\alpha(t), t \ge 0)$ *be the family of estimates of* α_0 *given by*

$$\alpha(t) = 1_{\{\hat{\alpha}(t)\in\mathcal{A}\}}\hat{\alpha}(t) + 1_{\{\hat{\alpha}(t)\notin\mathcal{A}\}}\alpha^*,$$

where $\hat{\alpha}(t)$ *is the solution of* (3.5) *and* α^* *is a fixed element of* $\mathcal{A}$. *If* (A1)–(A7) *are*

satisfied, $\omega - \beta > 0$*, and the inequality* (4.7) *is satisfied, then*

$$\tilde{J}(x, \tilde{u}) = \rho(\alpha_0)$$

for each $x \in H$*, where* $\tilde{u}$ *is the adaptive control given by* (4.30)*.*

REFERENCES

[1] A. Bensoussan, *Perturbation Methods in Optimal Control*, John Wiley, New York, 1988.

[2] J. M. Bismut, *Theorie probabiliste du controle des diffusion*, Mem. Amer. Math. Soc., 167 (1976), pp. 1–130.

[3] V. S. Borkar, *Optimal Control of Diffusion Processes*, Longman Press, London, UK, 1989.

[4] V. S. Borkar and M. K. Ghosh, *Ergodic control of multidimensional diffusions.* I: *The existence results*, SIAM J. Control Optim., 26 (1988), pp. 112–126.

[5] V. S. Borkar and M. K. Ghosh, *Ergodic control of multidimensional diffusions.* II: *Adaptive control*, Appl. Math. Optim., 21 (1990), pp. 191–220.

[6] S. Cerrai and F. Gozzi, *Strong solutions of Cauchy problems associated to weakly continuous semigroups*, Differential Integral Equations, 8 (1995), pp. 465–486.

[7] A. Chojnowska-Michalik and B. Goldys, *Existence, uniqueness and invariant measures for stochastic semilinear equations in Hilbert spaces*, Probab. Theory Related Fields, 102 (1995), pp. 331–356.

[8] A. Chojnowska-Michalik and B. Goldys, *On regularity properties of nonsymmetric Ornstein-Uhlenbeck semigroup in* L^p*-spaces*, Stochastics Stochastics Rep., 59 (1996), pp. 183–209.

[9] R. M. Cox, *Stationary and Discounted Control Diffusion Processes*, Ph.D. thesis, Columbia University, New York, 1984.

[10] G. Da Prato, K. D. Elworthy, and J. Zabczyk, *Strong Feller property for stochastic semilinear equations*, Stochastic Anal. Appl., 13 (1995), pp. 35–45.

[11] G. Da Prato and J. Zabczyk, *Stochastic Equations in Infinite Dimensions*, Cambridge University Press, Cambridge, UK, 1992.

[12] G. Da Prato and J. Zabczyk, *Ergodicity for Infinite Dimensional Systems*, Cambridge University Press, Cambridge, UK, 1996.

[13] T. E. Duncan, B. Goldys, and B. Pasik-Duncan, *Adaptive control of linear stochastic evolution system*, Stochastics Stochastics Rep., 36 (1991), pp. 71–90.

[14] T. E. Duncan, B. Maslowski, and B. Pasik-Duncan, *Adaptive boundary and point control of linear stochastic distributed parameter systems*, SIAM J. Control Optim., 32 (1994), pp. 648–672.

[15] T. E. Duncan, B. Maslowski, and B. Pasik-Duncan, *Adaptive boundary control of linear distributed parameter systems described by analytic semigroups*, Appl. Math. Optim., 33 (1996), pp. 107–138.

[16] T. E. Duncan, B. Maslowski, and B. Pasik-Duncan, *Ergodic boundary/point control of stochastic semilinear systems*, SIAM J. Control Optim., 36 (1998), pp. 1020–1047.

[17] T. E. Duncan, B. Pasik-Duncan, and L. Stettner, *Almost self-optimizing strategies for the adaptive control of diffusion processes*, J. Optim. Theory Appl., 81 (1994), pp. 479–507.

[18] T. E. Duncan, B. Pasik-Duncan, and L. Stettner, *On ergodic control of stochastic evolution equations*, Stochastic Process. Appl., 15 (1997), pp. 723–750.

[19] K. D. Elworthy, *Stochastic Differential Equations on Manifolds*, London Math. Soc. Lecture Note Ser. 70, Cambridge University Press, Cambridge, UK 1982.

[20] B. Goldys and B. Maslowski, *Ergodic control of semilinear stochastic equations and Hamilton-Jacobi equations*, J. Math. Anal. Appl., 234 (1999), pp. 592–631.

[21] B. Goldys and B. Maslowski, *Parameter estimation for controlled semilinear stochastic systems: Identifiability and consistency*, J. Multivariate Anal., submitted.

[22] F. Gozzi and E. Rouy, *Regular solutions to second-order stationary Hamilton-Jacobi equations*, J. Differential Equations, 130 (1996), pp. 201–234.

[23] F. Gozzi, E. Rouy, and A. Swiech, *Second-order Hamilton–Jacobi equations in Hilbert spaces and stochastic boundary control*, SIAM J. Control Optim., 38 (2000), pp. 400–430.

[24] D. Henry, *Geometric Theory of Semilinear Parabolic Equations*, Lecture Notes in Math. 840, Springer-Verlag, New York, 1981.

[25] H. J. Kushner, *Optimality conditions for the average cost per unit time problem with a diffusion model*, SIAM J. Control Optim., 16 (1978), pp. 330–346.

[26] R. S. Liptser and A. N. Shiryayev, *Statistics of Random Processes.* I: *General Theory*, Springer-Verlag, New York, 1977.

[27] B. Maslowski, *On probability distributions of solutions of semilinear stochastic evolution equations*, Stochastics Stochastics Rep., 45 (1993), pp. 17–44.
[28] M. Robin, *Long-term average cost problems for continuous time Markov processes—a survey*, Acta Appl. Math., 1 (1983), pp. 281–300.
[29] J. Seidler, *Ergodic behaviour of stochastic parabolic equations*, Czechoslovak Math. J., 47 (1992), pp. 277–316.
[30] J. Seidler, *Da Prato-Zabczyk maximal inequality revisited* I, Math. Bohem., 118 (1993), pp. 67–106.

SIAM J. CONTROL OPTIM.
Vol. 38, No. 6, pp. 1707–1728

MOMENT APPROACH TO NONLINEAR TIME OPTIMALITY*

G. M. SKLYAR† AND S. YU. IGNATOVICH‡

Abstract. The time-optimal control problem to a stationary point for an affine analytic nonlinear system is considered as a certain nonlinear power Markov moment min-problem. Conditions for a solution to be asymptotically close to a solution of a linear time-optimal control problem are obtained in terms of Lie brackets of the vector fields appearing in the system.

Key words. time-optimal control problem, Markov moment min-problem, series of nonlinear power moments, local equivalence

AMS subject classifications. 93B28, 93B17, 41A58

PII. S0363012997329767

1. Introduction and the main result. In the mathematical theory of linear control problems a remarkable place is occupied by the moment method. This method allows us to consider a number of such problems as problems from functional analysis. That gives a deep and well-developed technique for investigation.

The initial idea of the approach relies on the interpretation of the transfer conditions given by the Cauchy formula as moment equalities with respect to the control function. More specifically, let a linear system of the form

$$\dot{x} = A(t)x + B(t)u \tag{1.1}$$

be given, $x \in \mathbb{R}^n$, $u \in \mathbb{R}^1$, and the control $u(t)$, $t \in [0,T]$ transfers the system from an initial state $x(0) = x^0$ to the final state $x(T) = 0$. Then the function $u(t)$ satisfies the moment equalities

$$x_k^0 = \langle g_k, u \rangle = \int_0^T g_k(t)u(t)dt, \qquad k = 1, \dots, n, \tag{1.2}$$

where $g_k(t) = -\phi^{-1}(t)B(t)$ and $\phi(t)$ is the fundamental matrix of the system $\dot{x} = A(t)x$ such that $\phi(0) = I$.

If admissible controls are taken from a ball in a certain functional space H, then the control problem becomes the abstract Krein moment problem [15] in H and the controllability conditions are interpreted as conditions for solvability of the moment problem.

One of the most important and natural statements of a control problem is the problem with geometric restrictions on control. In this case $H = L_\infty[0,T]$, and we obtain the famous $(-1,1)$-Markov moment problem [17],

$$s_k = \int_0^T g_k(t)u(t)dt, \qquad k = 1, \dots, n, \qquad -1 \le u(t) \le 1. \tag{1.3}$$

*Received by the editors November 7, 1997; accepted for publication (in revised form) December 9, 1999; published electronically June 15, 2000.
http://www.siam.org/journals/sicon/38-6/32976.html

†Department of Mathematical Analysis, Kharkov National University, Svoboda sqr.4, Kharkov, 61077, Ukraine; Uniwersytet Szczeciński, Instytut Matematyki, ul. Wielkopolska 15, 70-451, Szczecin, Poland (sklar@sus.univ.szczecin.pl). This work was supported in part by the International Soros Science Education Program of the International Renáissance Foundation under grant APU 061096.

‡Department of Differential Equations and Control, Kharkov National University, Svoboda sqr.4, Kharkov, 61077, Ukraine (bob@online.kharkiv.com).

This problem has *a unique* solution if and only if s is such that $[0,T]$ is *the smallest* interval for which (1.3) holds. Moreover, if the functions $g_1(t),\dots,g_n(t)$ form a Tchebycheff system on $(0,T)$, then in the mentioned case $u(t)$ equals ± 1 and has no more than $n-1$ points of discontinuity.

That leads to the following statement of the moment problem (Markov moment problem on the smallest possible interval or min-problem [11]): for a given sequence of functions $\{g_k(t)\}_{k=1}^n$, $t\in[0,T]$, and a vector s, find the smallest possible interval $[0,\theta_s]\subset[0,T]$ such that for $\theta=\theta_s$ the following representation holds:

$$s_k=\int_0^\theta g_k(t)u(t)dt,\qquad k=1,\dots,n,\qquad |u(t)|\le 1; \tag{1.4}$$

and construct the function $u(t)=u_s(t)$ corresponding to this representation. The pair $(\theta_s,u_s(t))$ is called the solution of the min-problem. It follows from the above that the function $u_s(t)$ is unique, equals ± 1 on $(0,\theta_s)$, and has no more than $n-1$ points of discontinuity.

From the point of view of the optimal control, the Markov moment min-problem (1.4) is equivalent to the time-optimal control problem for linear system (1.1). In this relation note that the uniqueness and cited properties of the function $u_s(t)$ (which is interpreted as the time-optimal control) are well known in optimal control theory.

Nevertheless, the statement of the Markov min-problem gives a new, more precise tool for investigation of time optimality. On the one hand, it allows us to obtain an analytic solution of the time-optimal control problem in a number of important special cases [12, 13, 14] by employing deep techniques from classical moment theory. On the other hand, it suggests an approach to study the optimal solutions $(\theta_s,u_s(t))$, in particular, their behavior depending on the sequence $\{g_k(t)\}_{k=1}^n$ defined, in turn, by the system parameters.

Suppose the functions $g_1(t),\dots,g_n(t)$ are real analytic in $t\in[0,T]$. Then we represent the moment equalities (1.4) in the following form:

$$s_k=\sum_{i=0}^{\infty}\frac{1}{i!}g_k^{(i)}(0)\int_0^\theta t^i u(t)dt,\qquad k=1,\dots,n. \tag{1.5}$$

Let the functions $g_1(t),\dots,g_n(t)$ be linearly independent, as corresponds to the condition of null-controllability of (1.1). Let $\ell_1<\dots<\ell_n$ be indices of the first n linearly independent vectors from the sequence $\{g^{(j)}(0)\}_{j=0}^\infty$, and $G=(\frac{1}{\ell_1!}g^{(\ell_1)}(0),\dots,\frac{1}{\ell_n!}g^{(\ell_n)}(0))^{-1}$. Then (1.5) leads to

$$\tilde{s}_k=(Gs)_k=\int_0^\theta t^{\ell_k}u(t)dt+\sum_{j=\ell_k+1}^{\infty}r_{kj}\int_0^\theta t^j u(t)dt,\qquad k=1,\dots,n.$$

Note that in the right-hand side of the equalities the second term has a higher order of smallness than the first one as $\theta\to 0$. So, these equalities suggest asymptotic closeness of the solution of the min-problem (1.4) and the solution of the power min-problem with gaps [14]

$$\tilde{s}_k=\int_0^\theta t^{\ell_k}u(t)dt,\qquad k=1,\dots,n,\qquad |u(t)|\le 1, \tag{1.6}$$

as θ is small. Define this closeness more precisely following [18] as follows.

DEFINITION 1.1. *Two moment min-problems of the form* (1.4) *(with respect to sequences* $\{g_k(t)\}_{k=1}^n$ *and* $\{\tilde{g}_k(t)\}_{k=1}^n$ *of linearly independent functions), with solutions* $(\theta_s, u_s(t))$ *and* $(\tilde{\theta}_s, \tilde{u}_s(t))$*, respectively, are called locally equivalent to each other in a neighborhood of the origin if there exists a linear nonsingular operator* $L : \mathbb{R}^n \to \mathbb{R}^n$ *such that*

$$\frac{\tilde{\theta}_{Ls}}{\theta_s} \to 1, \qquad \frac{1}{\theta}\int_0^{\theta} |\tilde{u}_{Ls}(t) - u_s(t)|dt \to 0 \quad \text{as} \quad s \to 0,$$

where $\theta = \min\{\tilde{\theta}_{Ls}, \theta_s\}$.

Two time-optimal control problems (for controllable systems $\dot{x} = A(t)x + B(t)u$ *and* $\dot{x} = \tilde{A}(t)x + \tilde{B}(t)u$*) are called locally equivalent to each other in a neighborhood of the origin if the corresponding moment min-problems are locally equivalent.*

This approach allows us to construct a complete local classification of moment min-problems of the form (1.4) with real analytic functions $g_k(t)$ (and time-optimal control problems for systems of the form (1.1) with real analytic coefficients). This classification given in [18] is based on the following theorem.

THEOREM 1.2 (on local equivalence of linear problems). *Two time-optimal control problems (for controllable systems* $\dot{x} = A(t)x + B(t)u$ *and* $\dot{x} = \tilde{A}(t)x + \tilde{B}(t)u$ *with real analytic coefficients) are locally equivalent in a neighborhood of the origin if and only if the indices of the first n linearly independent vectors from the sequences* $\{(-A(t) + d/dt)^i B(t)\big|_{t=0}\}_{i=0}^{\infty}$ *and* $\{(-\tilde{A}(t) + d/dt)^i \tilde{B}(t)\big|_{t=0}\}_{i=0}^{\infty}$ *coincide.*

This fact means in essence that every min-problem (1.4) is locally equivalent to a certain power min-problem with gaps (1.6). Thus, in the sense of asymptotic behavior of solutions, the set of Markov moment min-problems of the form (1.4) splits into equivalence classes, and power min-problems with gaps (1.6) are representatives of these classes.

In the present work we intend to develop the moment approach to nonlinear systems of the form

$$\dot{x} = a(t,x) + b(t,x)u, \qquad a(t,0) \equiv 0, \tag{1.7}$$

in order to study the asymptotic behavior of solutions of the nonlinear time-optimal control problem to the stationary point $x = 0$ in a sense close to Definition 1.1. Namely, we are interested in the special subclass of such systems which are "close" to linear ones in the sense of the following definition.

DEFINITION 1.3. *Consider two time-optimal control problems*

$$\dot{x} = a(t,x) + b(t,x)u(t), \quad |u(t)| \le 1, \quad x(0) = x^0, \quad x(\theta) = 0, \quad \theta \to \min, \tag{1.8}$$

and

$$\dot{x} = A(t)x + B(t)u(t), \quad |u(t)| \le 1, \quad x(0) = x^0, \quad x(\theta) = 0, \quad \theta \to \min, \tag{1.9}$$

in a neighborhood of the origin. Suppose $a(t,0) \equiv 0$ *and both of systems are null-controllable. Denote by* $(\theta_{x^0}^{Lin}, u_{x^0}^{Lin}(t))$ *the solution of linear problem* (1.9) *and by* $\{(\theta_{x^0}, \tilde{u}(t)) : \tilde{u}(t) \in U_{x^0}\}$ *the set of solutions of nonlinear problem* (1.8) *(it is nonempty due to null-controllability of the system; see* [3]*).*

The nonlinear time-optimal control problem (1.8) *is said to be locally equivalent to the linear problem* (1.9) *if there exists a nonsingular transformation* Φ *of a neighborhood of the origin in* $\mathbb{R}^n$, $\Phi(0) = 0$, *such that*

$$\frac{\theta_{\Phi(x^0)}}{\theta_{x^0}^{Lin}} \to 1, \qquad \sup_{\tilde{u}(t) \in U_{\Phi(x^0)}} \frac{1}{\theta}\int_0^{\theta} |u_{x^0}^{Lin}(t) - \tilde{u}(t)|dt \to 0 \quad \text{as} \quad x^0 \to 0, \tag{1.10}$$

where $\theta = \min\{\theta_{\Phi(x^0)}, \theta_{x^0}^{Lin}\}$.

In other words, the local equivalence means that there exists such a change of variables $z = \Phi^{-1}(x)$ in the nonlinear system that solutions $\{(\theta_{x^0}^z, \tilde{u}(t)) : \tilde{u}(t) \in U_{x^0}^z\}$ of the time-optimal control problem for the system in the new variables z are "asymptotically close" to the solution of the linear problem, i.e.,

$$\frac{\theta_{x^0}^z}{\theta_{x^0}^{Lin}} \to 1, \qquad \sup_{\tilde{u}(t)\in U_{x^0}^z} \frac{1}{\theta}\int_0^\theta |u_{x^0}^{Lin}(t) - \tilde{u}(t)|dt \to 0 \quad \text{as} \quad x^0 \to 0; \quad \theta = \min\{\theta_{x^0}^z, \theta_{x^0}^{Lin}\}.$$

In order to formulate the main result we introduce mappings R_a, R_b (corresponding to $a(t,x)$, $b(t,x)$) acting on a real analytic vector function $d(t,x)$ by the rule

$$R_a d(t,x) = d_t(t,x) + d_x(t,x)\cdot a(t,x),$$
$$R_b d(t,x) = d_x(t,x)\cdot b(t,x).$$

Here $d_t(t,x) = \partial d(t,x)/\partial t$, and $d_x(t,x)$ is the matrix $\{\partial d_i(t,x)/\partial x_j\}_{i=1,\dots,n}^{j=1,\dots,n}$. Further, let $[R_a, R_b]$ denote the operator commutator, $[R_a, R_b] = R_a \circ R_b - R_b \circ R_a$, and $\operatorname{ad}_{R_a}^0 R_b = R_b$, $\operatorname{ad}_{R_a}^{m+1} R_b = [R_a, \operatorname{ad}_{R_a}^m R_b]$, $m \ge 0$. Let also $E(x) \equiv x$. Note that the condition $a(t,0) \equiv 0$ implies

$$\operatorname{ad}_{R_a}^i R_b E(x)\Big|_{\substack{t=0\\x=0}} = (-a_x(t,0) + d/dt)^i b(t,0)\Big|_{t=0}, \qquad i \ge 0. \tag{1.11}$$

The main result of the paper is the following theorem.

THEOREM 1.4 (on local equivalence of nonlinear problems to linear ones). *Consider a time-optimal control problem for control affine real analytic system* (1.7). *Let its linearization* $\dot{x} = a_x(t,0)x + b(t,0)u$ *be controllable.*

The original nonlinear time-optimal control problem is locally equivalent to the problem for a certain linear system (and then, in particular, for its own linearization) in the sense of Definition 1.3 *if and only if the system satisfies condition* (E),

$$\text{(E)}\quad \left[\operatorname{ad}_{R_a}^{m_1} R_b, \cdots \left[\operatorname{ad}_{R_a}^{m_{k-1}} R_b, \operatorname{ad}_{R_a}^{m_k} R_b\right]\cdots\right] E(x)\Big|_{\substack{t=0\\x=0}} \in \operatorname{Lin}\left\{\operatorname{ad}_{R_a}^i R_b E(x)\Big|_{\substack{t=0\\x=0}}\right\}_{i=0}^{m-2},$$

where $m = m_1 + \cdots + m_k + k$ *for all* $k \ge 2$, $m_1, \dots, m_k \ge 0$.

In its turn the time-optimal control problem for the linearization is reduced to the Markov moment min-problem which is equivalent to the power moment min-problem with gaps in the sense of Definition 1.1. *The solution of this latter problem (which can be computed analytically in a number of cases) may serve as an approximation of the optimal time and control for the original problem in a neighborhood of the origin (after a change of variables* $z = L\Phi^{-1}(x)$, *where* Φ *is from Definition* 1.3 *and* L *is a certain matrix). Moreover, this change of variables may be chosen in polynomial form.*

REMARK 1.5. *If* $\operatorname{rank}\{\operatorname{ad}_{R_a}^i R_b E(x)_{|\,\substack{t=0\\x=0}}\}_{i=0}^N = n$, *then condition* (E) *is substantial for* $k \ge 2$ *and* $m_1, \dots, m_k \ge 0$ *such that* $m = m_1 + \cdots + m_k + k \le N + 1$ *only. That means that only a finite number of relations determines the property of local equivalence to a linear time-optimal control problem.*

REMARK 1.6. *The proof of the sufficiency of the theorem is constructive, namely, we explicitly build a polynomial change of variables* $z = \Phi^{-1}(x)$ (*section* 5).

In order to prove Theorem 1.4, we propose a development of the moment approach to the nonlinear case. We suggest considering the following:

(i) series of nonlinear power moments instead of systems of the form (1.7);

(ii) transformations of such series instead of changes of variables in systems;

(iii) nonlinear moment min-problems instead of time-optimal control problems.

In this way the local equivalence of time-optimal control problems turns into the local equivalence of moment min-problems.

Elements of the moment approach are given in sections 2–5.

In Theorem 2.1 the representation of systems of the form (1.7) as series of nonlinear power moments is given. In essence, we transform M. Fliess' series representation of trajectories of system (1.7) to the case when the final point is fixed.

The concept of a nonlinear power Markov moment min-problem is introduced in section 3 as well as the definition of local equivalence of nonlinear and linear problems.

The basic tool in our approach is the concept of an essentially linearizable (ess-linearizable) series of nonlinear power moments given in section 4. Two substantial steps of the proof of Theorem 1.4 are presented in Theorems 4.2 and 5.1. Theorem 4.2 shows that a nonlinear time-optimal control problem is locally equivalent to a linear one if and only if its series is ess-linearizable. On the other hand, Theorem 5.1 states that the series is ess-linearizable if and only if the system satisfies condition (E).

As a result, the proof of Theorem 1.4 following from Theorems 4.2 and 5.1 can be found in section 6 as well as a number of special classes of systems satisfying the conditions of the theorem.

Section 7 contains a description of some results developing the moment approach and announces some future results.

2. Series of nonlinear power moments. The basic point for the present section is conditions which the control function $u(t)$ transferring the initial state to the origin by virtue of system (1.7) satisfies.

For systems of the form (1.7), it is natural to consider the Volterra series expansion of a solution [1, 2, 8, 9, 16], as an analogue of the Cauchy formula for the linear case. Taking into account that the final point $x(\theta) = 0$ is stationary for system (1.7) ($a(t, 0) \equiv 0$), one can obtain the transform conditions in the following form:

$$x_q^0 = \sum_{k=1}^{\infty} \int_0^{\theta} \int_0^{\tau_1} \dots \int_0^{\tau_{k-1}} w_q(\tau_1, \dots, \tau_k) u(\tau_1) \cdots u(\tau_k) d\tau_k \cdots d\tau_2 d\tau_1, \tag{2.1}$$

$q = 1, \dots, n$, where the kernels $w_q(\tau_1, \dots, \tau_k)$ depend on system parameters and the control $u(t)$ transfers system (1.7) from x^0 to the origin in time θ. Equalities (2.1) are supposed to be considered as a generalization of moment equalities (1.2).

In order to generalize the representation (1.5), it seems to be natural to pass in (2.1) to the expansion of kernels $w_q(\tau_1, \dots, \tau_k)$ into Taylor series. Note, however, that the series we obtain does not possess an important property: its remainder is not of higher order of smallness than the partial sum as $\theta \to 0$. To satisfy this condition one has to change the manner of summation of the series. This leads us to the necessity of finding another form of transfer conditions called *a series of nonlinear power moments.*

THEOREM 2.1. *Let functions $a(t, x)$, $b(t, x)$ be analytic on a neighborhood of the origin in $\mathbb{R}^{n+1}$. Denote by $U_{x^0}(\theta)$ the set of all controls $u(t) \in L_\infty[0, \theta]$, $\|u\| \le 1$, transferring x^0 to the origin in time θ by virtue of system* (1.7).

Then there exists a number $T_0 > 0$ such that for any $\theta \in [0, T_0]$ a point x^0 satisfying the condition $U_{x^0}(\theta) \neq \emptyset$ admits the representation in the form of a series

of nonlinear power moments,

$$x^0 = \sum_{m=1}^{\infty} \sum_{\substack{m_1+\dots+m_k+k=m \\ k\geq 1, m_j\geq 0}} v_{m_1\dots m_k}\xi_{m_1\dots m_k}(\theta,u), \qquad u(t)\in U_{x^0}(\theta), \tag{2.2}$$

where $\xi_{m_1\dots m_k}(\theta,u)$ *denote nonlinear power moments of the function* $u(t)$,

$$\xi_{m_1\dots m_k}(\theta,u) = \int_0^\theta \int_0^{\tau_1} \dots \int_0^{\tau_{k-1}} \tau_1^{m_1}\tau_2^{m_2}\dots\tau_k^{m_k} \prod_{j=1}^{k} u(\tau_j)d\tau_k\dots d\tau_2 d\tau_1, \tag{2.3}$$

and constant vector coefficients $v_{m_1\dots m_k}$ *are of the form*

$$v_{m_1\dots m_k} = \frac{(-1)^k}{m_1!\dots m_k!} \operatorname{ad}_{R_a}^{m_1} R_b \circ \dots \circ \operatorname{ad}_{R_a}^{m_k} R_b E(x)\Big|_{\substack{t=0\\x=0}}. \tag{2.4}$$

Moreover, there exists a constant $K>0$ *such that for any* $\theta\in[0,T_0]$ *and* $u(t)\in L_\infty[0,\theta]$, $\|u\|\leq 1$, *the following estimate holds:*

$$\sum_{m=1}^{\infty} \left\| \sum_{\substack{m_1+\dots+m_k+k=m \\ k\geq 1, m_j\geq 0}} v_{m_1\dots m_k}\xi_{m_1\dots m_k}(\theta,u) \right\| < K. \tag{2.5}$$

Let us explain the place which this result occupies with respect to Fleiss' well-known expansion of a solution of the Cauchy problem for system (1.7) into a series of iterated integrals [4, 5, 6, 7]. Suppose $u(t)\in U_{x^0}(\theta)$ and consider the trajectory $x(t)$ of (1.7). Then the vector function $y(t)=x(\theta-t)$ is a solution of the Cauchy problem

$$\dot y = -\hat a(t,y) - \hat b(t,y)\hat u(t), \qquad y(0)=0,$$

where $\hat u(t)=u(\theta-t)$, $\hat a(t,y)=a(\theta-t,y)$, and $\hat b(t,y)=b(\theta-t,y)$, and satisfies the end condition $y(\theta)=x^0$. Applying Fliess' expansion to $y(t)$, we obtain

$$\begin{aligned} y(t) = y(0) + \sum_{m=1}^{\infty} \sum_{M\subset\{1,\dots,m\}} (-1)^m R_{\hat c_m}\circ R_{\hat c_{m-1}}\circ\dots \\ \circ R_{\hat c_1}E(y)\Big|_{\substack{t=0\\y=0}} \int_0^t \int_0^{\tau_1} \dots \int_0^{\tau_{m-1}} \prod_{j\in M} u(\theta-\tau_j)d\tau_m\dots d\tau_2 d\tau_1, \end{aligned}$$

where $\hat c_j(t,y)=\hat b(t,y)$ if $j\in M$, and $\hat c_j(t,y)=\hat a(t,y)$ otherwise; $E(y)\equiv y$. Obviously, this representation implies

$$\begin{aligned} x(\theta-t) = \sum_{m=1}^{\infty} \sum_{M\subset\{1,\dots,m\}} (-1)^m R_{c_m}\circ R_{c_{m-1}}\circ\dots \\ \circ R_{c_1}E(x)\Big|_{\substack{t=\theta\\x=0}} \int_{\theta-t}^{\theta} \int_{\theta-t}^{\tau_1} \dots \int_{\theta-t}^{\tau_{m-1}} \prod_{j\in M} u(\tau_{m+1-j})d\tau_m\dots d\tau_2 d\tau_1, \end{aligned} \tag{2.6}$$

where $c_j(t,x)=b(t,x)$ if $j\in M$, and $c_j(t,x)=a(t,x)$ otherwise. One can show that there exists $T_0>0$ such that (2.6) holds for $t=\theta\in[0,T_0]$.

It is easy to see that in the case of an autonomous system ($a(t,x) \equiv a(x)$, $b(t,x) \equiv b(x)$), coefficients of the expansion do not depend on θ. So, taking into account the condition $a(0) \equiv 0$ which implies $R_a^m d(0) = 0$, $m \geq 1$, we obtain representation (2.2)–(2.4) directly by simplification of integrals

$$\int_0^\theta \int_0^{\tau_1} \cdots \int_0^{\tau_{m-1}} \prod_{j \in M} u(\tau_{m+1-j}) d\tau_m \dots d\tau_2 d\tau_1,$$

which leads to nonlinear power moments.

However, in the nonautonomous case the coefficients of integrals in (2.6) are functions on θ. Hence, to obtain representation (2.2)–(2.4) from (2.6), one has to transform the sum expanding each coefficient into the Taylor series at $t = 0$. This shows a difference between the expansion of the Cauchy problem solution and transfer conditions which coefficients are calculated at the final point with respect to the state ($x = 0$) and at the "initial point" with respect to the time ($t = 0$). Therefore, the direct reduction of a nonautonomous case to an autonomous one which is generally accepted for the Cauchy problem is impossible for the transfer problem. This fact explains the meaning of Theorem 2.1.

We emphasize especially that representation (2.2)–(2.4) holds only if $x = 0$ is a stationary point for the system.

DEFINITION 2.2. *We say that a series S of the form*

$$S = \sum_{m=1}^{\infty} \sum_{\substack{m_1+\dots+m_k+k=m \\ k \geq 1, m_j \geq 0}} v_{m_1 \dots m_k} \xi_{m_1 \dots m_k}, \tag{2.7}$$

$v_{m_1 \dots m_k} \in \mathbb{R}^n$, represents a system of the form (1.7) *if there exist two real analytic vector functions $a(t,x)$, $b(t,x)$ such that coefficients $v_{m_1 \dots m_k}$ satisfy* (2.4).

Note that each such series represents not one system but a family of systems (connected by certain time-dependent changes of variables).

An essential advantage of representation (2.2)–(2.4) is a possibility to study analytic transformations of system (1.7) as analytic transformations over its series of nonlinear power moments and, therefore, as certain transformations over the vector coefficients $v_{m_1 \dots m_k}$. The following lemma states the exact expression for coefficients of the transformed series.

LEMMA 2.3. *Let a series S represent a system of the form* (1.7)*, and let F be an analytic transformation of a neighborhood of the origin in $\mathbb{R}^n$, $F(0) = 0$. Then the series $\tilde{S} = F(S)$ represents the system $\dot{z} = \tilde{a}(t,z) + \tilde{b}(t,z)u$, where $z = F(x)$ and the coefficients $\tilde{v}_{m_1 \dots m_k}$ of $\tilde{S}$ can be found by the following formulas:*

$$\tilde{v}_{m_1 \dots m_k} = \sum_{r=1}^{k} \frac{1}{r!} D^{(r)} F(0) \sum{}' v_{m_{i_{11}} \dots m_{i_{1\ell_1}}} \cdots v_{m_{i_{r1}} \dots m_{i_{r\ell_r}}}, \tag{2.8}$$

where $D^{(r)}$ is the rth derivative operator and the sum $\sum'$ is taken on all collections of nonempty nonintersecting sets $\{i_{j1}, \dots, i_{j\ell_j}\}$, $j = 1, \dots, r$, such that $i_{j1} < \dots < i_{j\ell_j}$ and $\cup_{j=1}^{r} \{i_{j1}, \dots, i_{j\ell_j}\} = \{1, \dots, k\}$.

Further, we use the following definition.

DEFINITION 2.4. *For given $v_{m_1 \dots m_k}$, we refer to the number $m = m_1 + \dots + m_k + k$ as the order of $v_{m_1 \dots m_k}$ and to the number k as its multiplicity.*

REMARK 2.5. *As follows from Lemma* 2.3*, all terms of multiplicity k in the series $\tilde{S} = F(S)$ are completely determined by the first k derivatives of the transformation F.*

3. Nonlinear moment min-problem. In [11, 12], the new approach to a time-optimal control problem has been introduced on the base of the concept of a Markov moment min-problem to which the time-optimal control problem can be reduced. Developing this idea, we formulate *the nonlinear Markov moment problem on the smallest possible interval (nonlinear moment min-problem)* for a series S of the form (2.7) representing a system as follows.

For any vector $s \in \mathbb{R}^n$, indicate (if possible) the minimal number $\theta_s \geq 0$ for which there exists a function $u_s(t) \in L_\infty[0, \theta_s]$ such that $\|u_s\| \leq 1$ and the following equality holds:

$$s = \sum_{m=1}^{\infty} \sum_{\substack{m_1+\cdots+m_k+k=m \\ k\geq 1, m_j \geq 0}} v_{m_1 \ldots m_k} \xi_{m_1 \ldots m_k}(\theta, u), \tag{3.1}$$

as $\theta = \theta_s$ and $u(t) = u_s(t)$. If such a pair $(\theta_s, u_s(t))$ exists, then it is called *a solution* of the min-problem.

Note that the nonlinear Markov moment min-problem is naturally interpreted as the time-optimal control problem for system (1.7) (from the point $x(0) = s$). Moreover, if the system is locally null-controllable, then a solution of the time-optimal control problem exists for any initial point from a neighborhood of the origin due to [3].

In [11, 18] the question on local equivalence of linear moment min-problems was considered. It proved to be that power moment min-problems with gaps [14] play the main role in this investigation. Now we introduce the concept of local equivalence of a nonlinear moment min-problem and a linear power moment min-problem.

DEFINITION 3.1. *Let the nonlinear Markov moment min-problem* (3.1) *have a solution for any s from a neighborhood of the origin. Denote by $\{(\theta_s, \tilde{u}(t)) \colon \tilde{u}(t) \in U_s\}$ the set of all its solutions.*

This problem is called locally equivalent to the power moment min-problem with gaps

$$s_q = \int_0^\theta t^{\ell_q} u(t) dt, \qquad q = 1, \ldots, n, \qquad 0 \leq \ell_1 < \cdots < \ell_n, \tag{3.2}$$

if there exists a nonsingular analytic transformation Φ of a neighborhood of the origin, $\Phi(0) = 0$, such that

$$\frac{\theta_{\Phi(s)}}{\theta_s^{Lin}} \to 1, \tag{3.3}$$

$$\sup_{\tilde{u}(t) \in U_{\Phi(s)}} \frac{1}{\theta} \int_0^\theta |u_s^{Lin}(t) - \tilde{u}(t)| dt \to 0, \tag{3.4}$$

as $s \to 0$, where $\theta = \min\{\theta_{\Phi(s)}, \theta_s^{Lin}\}$ and $(\theta_s^{Lin}, u_s^{Lin}(t))$ is a solution of (3.2).

Note that this definition in essence coincides with the definition of local equivalence between nonlinear and linear time-optimal control problems (Definition 1.3).

4. Essentially linearizable series and local equivalence. Within our approach, we introduce the following definition.

DEFINITION 4.1. *A series S of the form* (2.7) *is called linearly nonsingular if* $\operatorname{rank}\{v_i\}_{i=0}^{\infty} = n$.

A series S of the form (2.7) *is called essentially linear (ess-linear) if it is linearly nonsingular and*

$$v_{m_1\dots m_k} \in \operatorname{Lin}\{v_i\}_{i=0}^{m-2}, \qquad \text{where} \quad m = m_1 + \cdots + m_k + k$$

for any $k \geq 2$, $m_1, \dots, m_k \geq 0$. We denote by $\mathcal{L}_e$ the set of all ess-linear series.

A series S is called ess-linearizable if there exists an analytic transformation (ess-linearizing transformation) F of a neighborhood of the origin, $F(0) = 0$, such that $F(S) \in \mathcal{L}_e$.

In other words, a linearly nonsingular series is ess-linear if and only if each coefficient of multiplicity greater than 1 is included in the linear span of coefficients of *smaller order* and multiplicity 1.

The following theorem shows the significance of this concept.

THEOREM 4.2. *Consider a time-optimal control problem for control affine real analytic system* (1.7). *Let its linearization at the origin be controllable.*

The original nonlinear time-optimal control problem is locally equivalent to the problem for a certain linear system in the sense of Definition 1.3 *if and only if the series S representing the system* (1.7) *is ess-linearizable.*

Proof of sufficiency. Suppose the ess-linearizable series S of the form (2.7) represents the system (1.7). Let F be such a transformation that $\tilde{S} = F(S) \in \mathcal{L}_e$. Without loss of generality, we may assume that $\tilde{S}$ is of the form $\tilde{S} = \Xi + \tilde{\rho}$, where

$$\Xi = (\xi_{\ell_1}, \dots, \xi_{\ell_n}), \qquad \tilde{\rho}_q = \sum_{m=\ell_q+2}^{\infty} \sum_{\substack{m_1+\cdots+m_k+k=m \\ k\geq 1, m_j \geq 0}} (\tilde{v}_{m_1\dots m_k})_q \xi_{m_1\dots m_k}, \qquad q = 1, \dots, n,$$

where coefficients $\tilde{v}_{m_1\dots m_k}$ and $v_{m_1\dots m_k}$ are connected by equalities (2.8).

We prove that nonlinear Markov moment min-problem (3.1) for the series S is locally equivalent to linear problem (3.2) in the sense of Definition 3.1 with transformation $\Phi = F^{-1}$; therefore, the original nonlinear time-optimal control problem is locally equivalent to the problem for a certain linear system in the sense of Definition 1.3.

Put $\Phi = F^{-1}$; then the set of solutions $\{(\theta_{\Phi(s)}, \tilde{u}(t)) : \tilde{u}(t) \in U_{\Phi(s)}\}$ of the problem (3.1) coincides with the set of solutions of the problem

$$s = \Xi(\theta, u) + \tilde{\rho}(\theta, u); \tag{4.1}$$

that is,

$$s = \Xi(\theta_{\Phi(s)}, \tilde{u}) + \tilde{\rho}(\theta_{\Phi(s)}, \tilde{u}) \quad \text{for any} \quad \tilde{u}(t) \in U_{\Phi(s)}.$$

Denote by $(\theta_s^{Lin}, u_s^{Lin}(t))$ the solution of the linear problem (3.2). Then

$$\theta_{s^0}^{Lin} \leq \theta_{\Phi(s)}, \qquad s^0 = s - \tilde{\rho}(\theta_{\Phi(s)}, \tilde{u}), \qquad \tilde{u}(t) \in U_{\Phi(s)}. \tag{4.2}$$

Estimate $\theta_{\Phi(s)}$ from above. Following [11] we introduce the operator $D : \mathbb{R}^n \to \mathbb{R} \times L_\infty[0, \infty]$, associating a pair $(\theta_x^{Lin}, u_x^{Lin}(t))$ to a vector x. Obviously, the operator $G_s(x) = s - \tilde{\rho}(D(x))$ is defined in a neighborhood of the origin. Next we show that the operator G_s has a stationary point in a certain subneighborhood of the origin.

Since $\xi_{m_1\dots m_k}(\theta, u) = \theta^m \xi_{m_1\dots m_k}(1, \hat{u})$, where $m = m_1 + \cdots + m_k + k$, $\hat{u}(t) = u(t\theta)$, $t \in [0, 1]$, we have from (2.5)

$$\left\| \sum_{\substack{m_1+\cdots+m_k+k=m \\ k\geq 1, m_j \geq 0}} \tilde{v}_{m_1\dots m_k} \xi_{m_1\dots m_k}(\theta, u) \right\| \leq (K_1\theta)^m, \qquad m \geq 1,$$

where the constant K_1 does not depend on θ and $u(t) \in L_\infty[0,\theta]$, $\|u\| \le 1$. Hence, for $\theta \to 0$,

$$|\tilde\rho_q(\theta,u)| \le \sum_{m=\ell_q+2}^{\infty} (K_1\theta)^m \le K_2\theta^{\ell_q+2}, \qquad q=1,\dots,n. \tag{4.3}$$

Consider now the closed neighborhood of the origin $V_\varepsilon = \{x : |x_q| \le \varepsilon^{\ell_q+1}, q = 1,\dots,n\}$, $\varepsilon > 0$, and note that $\theta_x^{Lin} \le C_1\varepsilon$ for any $x \in V_\varepsilon$, where $C_1 = \max\{\theta_x^{Lin}, x \in V_1\} \ge 1$. Put $\varepsilon(s) = \max_{1\le q\le n}\{(2|s_q|)^{\frac{1}{\ell_q+1}}\}$. It is easy to see that if $\varepsilon(s) \le (2K_2C_1^{\ell_n+2})^{-1}$, then the operator G_s maps the closed set $V_{\varepsilon(s)}$ to itself. At the same time, G_s is continuous, and hence it has a stationary point $s^1 \in V_{\varepsilon(s)}$, $G_s(s^1) = s^1$, that is,

$$s = \Xi(\theta_{s^1}^{Lin}, u_{s^1}^{Lin}) + \tilde\rho(\theta_{s^1}^{Lin}, u_{s^1}^{Lin}).$$

Hence

$$\theta_{\Phi(s)} \le \theta_{s^1}^{Lin}, \qquad s^1 = s - \tilde\rho(\theta_{s^1}^{Lin}, u_{s^1}^{Lin}) \to 0 \quad \text{as} \quad s \to 0. \tag{4.4}$$

Thus from (4.2)–(4.4), we have

$$\theta_{s^0}^{Lin} \le \theta_{\Phi(s)} \le \theta_{s^1}^{Lin}, \tag{4.5}$$

$$\left|\left(s^1 - s^0\right)_q\right| \le 2K_2(\theta_{s^1}^{Lin})^{\ell_q+2}, \qquad \left|\left(s - s^1\right)_q\right| \le K_2(\theta_{s^1}^{Lin})^{\ell_q+2}, \qquad q=1,\dots,n.$$

Following [18], we introduce an operator $H_\delta : \mathbb{R}^n \to \mathbb{R}^n$ acting by the rule

$$H_\delta(x) = \left(\frac{x_1}{\delta^{\ell_1+1}}, \dots, \frac{x_n}{\delta^{\ell_n+1}}\right),$$

and we note that

$$\theta_x^{Lin} = \delta\theta_{H_\delta(x)}^{Lin}, \qquad u_x^{Lin}(t) = u_{H_\delta(x)}^{Lin}(t/\delta), \qquad t \in [0, \theta_x^{Lin}].$$

Putting $\delta = \theta_{s^1}^{Lin}$, we obtain

$$\theta_{H_\delta(s^1)}^{Lin} = 1, \qquad \left\|H_\delta(s^1 - s^0)\right\| \le 2K_2\delta, \qquad \left\|H_\delta(s - s^1)\right\| \le K_2\delta,$$

so $\operatorname{dist}(H_\delta(s^0), \{x : \theta_x^{Lin} = 1\}) \le 2K_2\delta$, $\operatorname{dist}(H_\delta(s), \{x : \theta_x^{Lin} = 1\}) \le K_2\delta$. Since θ_x^{Lin} is uniformly continuous as a function of x and $\delta = \theta_{s^1}^{Lin} \to 0$ as $s \to 0$, we have

$$\frac{\theta_s^{Lin}}{\theta_{s^1}^{Lin}} = \frac{\theta_{H_\delta(s)}^{Lin}}{\theta_{H_\delta(s^1)}^{Lin}} = \theta_{H_\delta(s)}^{Lin} \to 1, \qquad \frac{\theta_{s^0}^{Lin}}{\theta_{s^1}^{Lin}} = \frac{\theta_{H_\delta(s^0)}^{Lin}}{\theta_{H_\delta(s^1)}^{Lin}} = \theta_{H_\delta(s^0)}^{Lin} \to 1 \qquad \text{as} \quad s \to 0,$$

which proves (3.3) due to (4.5).

In order to prove (3.4), let us make use of the fact that the function $u_s^{Lin}(t)$ is piecewise constant, $|u_s^{Lin}(t)| = 1$, and has no more than $n-1$ points of discontinuity on the interval $(0, \theta_s^{Lin})$. For any s consider the polynomial $p_s(t) = \sum_{q=1}^n \alpha_q^{(s)} t^{\ell_q}$, $t \in [0,1]$, $\max_{1\le q\le n} |\alpha_q^{(s)}| = 1$, which has its roots in the points of discontinuity

of $u_s^{Lin}(t\theta_s^{Lin})$ and $\operatorname{sign} p_s(t) = u_s^{Lin}(t\theta_s^{Lin})$ as $u_s^{Lin}(t\theta_s^{Lin}) \neq 0$. Let us consider any $\tilde{u}(t) \in U_{\Phi(s)}$ and put $\tilde{u}(t) = 0$ as $t > \theta_{\Phi(s)}$. We have

$$\int_0^{\theta_s^{Lin}} p_s(t/\theta_s^{Lin}) u_s^{Lin}(t)dt - \int_0^{\theta_{\Phi(s)}} p_s(t/\theta_s^{Lin})\tilde{u}(t)dt$$

$$= \sum_{q=1}^{n} \alpha_q^{(s)} \frac{\xi_{\ell_q}(\theta_s^{Lin}, u_s^{Lin}) - \xi_{\ell_q}(\theta_{\Phi(s)}, \tilde{u})}{(\theta_s^{Lin})^{\ell_q}} = \sum_{q=1}^{n} \frac{\alpha_q^{(s)}}{(\theta_s^{Lin})^{\ell_q}} \tilde{\rho}_q(\theta_{\Phi(s)}, \tilde{u}),$$

which yields

$$\left| \int_0^{\theta_s^{Lin}} p_s(t/\theta_s^{Lin}) u_s^{Lin}(t)dt - \int_0^{\theta_{\Phi(s)}} p_s(t/\theta_s^{Lin})\tilde{u}(t)dt \right| \leq \sum_{q=1}^{n} \frac{K_2 \theta_{\Phi(s)}^{\ell_q+2}}{(\theta_s^{Lin})^{\ell_q}}.$$

On the other hand,

$$\int_0^{\theta_s^{Lin}} p_s(t/\theta_s^{Lin}) u_s^{Lin}(t)dt - \int_0^{\theta_{\Phi(s)}} p_s(t/\theta_s^{Lin})\tilde{u}(t)dt$$

$$= \int_0^{\theta_s^{Lin}} p_s(t/\theta_s^{Lin}) \left(u_s^{Lin}(t) - \tilde{u}(t)\right) dt + \int_{\theta_{\Phi(s)}}^{\theta_s^{Lin}} p_s(t/\theta_s^{Lin})\tilde{u}(t)dt.$$

Thus we get

$$\frac{1}{\theta_s^{Lin}} \int_0^{\theta_s^{Lin}} \left|p_s(t/\theta_s^{Lin})\right| \left|u_s^{Lin}(t) - \tilde{u}(t)\right| dt = \frac{1}{\theta_s^{Lin}} \left| \int_0^{\theta_s^{Lin}} p_s(t/\theta_s^{Lin}) \left(u_s^{Lin}(t) - \tilde{u}(t)\right) dt \right|$$

$$\leq \theta_{\Phi(s)} \sum_{q=1}^{n} K_2 \left(\frac{\theta_{\Phi(s)}}{\theta_s^{Lin}}\right)^{\ell_q+1} + \sum_{q=1}^{n} \frac{1}{\ell_q + 1} \left| 1 - \left(\frac{\theta_{\Phi(s)}}{\theta_s^{Lin}}\right)^{\ell_q+1} \right|$$

for any $\tilde{u}(t) \in U_{\Phi(s)}$. The latter relation and (3.3) yield

$$\sup_{\tilde{u}(t)\in U_{\Phi(s)}} \frac{1}{\theta_s^{Lin}} \int_0^{\theta_s^{Lin}} \left|p_s(t/\theta_s^{Lin})\right| \left|u_s^{Lin}(t) - \tilde{u}(t)\right| dt \to 0 \quad \text{as} \quad s \to 0. \tag{4.6}$$

Now we prove that (4.6) implies (3.4). Consider any sequence $\{s_k\}_{k=1}^{\infty}$, $s_k \to 0$ as $k \to \infty$, and any $\tilde{u}_k(t) \in U_{\Phi(s_k)}$. Without loss of generality, we assume that $p_{s_k}(t) \to p(t)$ as $k \to \infty$ pointwise, where $p(t) = \sum_{q=1}^{n} \alpha_q t^{\ell_q}$, $t \in [0,1]$, $\max_{1\leq q\leq n} |\alpha_q| = 1$. Further, for any $\varepsilon > 0$, we put $F_\varepsilon = \{t \in [0,1] : |p(t)| \geq \varepsilon\}$. Note that $\mu F_\varepsilon \to 1$ as $\varepsilon \to 0$. Then we get

$$\int_0^1 \left|u_{s_k}^{Lin}(t\theta_{s_k}^{Lin}) - \tilde{u}_k(t\theta_{s_k}^{Lin})\right| dt \leq \frac{1}{\varepsilon} \int_{F_\varepsilon} |p(t)| \left|u_{s_k}^{Lin}(t\theta_{s_k}^{Lin}) - \tilde{u}_k(t\theta_{s_k}^{Lin})\right| dt + 2(1 - \mu F_\varepsilon)$$

$$\leq \frac{2}{\varepsilon} \int_0^1 |p(t) - p_{s_k}(t)|dt + \frac{1}{\varepsilon} \int_0^1 |p_{s_k}(t)| \left|u_{s_k}^{Lin}(t\theta_{s_k}^{Lin}) - \tilde{u}_k(t\theta_{s_k}^{Lin})\right| dt + 2(1 - \mu F_\varepsilon).$$

Using the arbitrariness of ε, we obtain from this relation and (4.6)

$$\lim_{k\to\infty} \sup_{\tilde{u}_k(t)\in U_{\Phi(s_k)}} \int_0^1 \left|u_{s_k}^{Lin}(t\theta_{s_k}^{Lin}) - \tilde{u}_k(t\theta_{s_k}^{Lin})\right| dt = 0,$$

and, therefore, (3.4) is valid due to (3.3). Sufficiency is proved.

Necessity is based essentially on Lemma 8.3.

Let the original nonlinear time-optimal control problem be locally equivalent to the problem for the linear system $\dot{x} = A(t)x + B(t)u$, and let $\ell_1 < \cdots < \ell_n$ be indices of the first n linearly independent vectors from the sequence $\{(-A(t) + d/dt)^j B(t)|_{t=0}\}_{j=0}^{\infty}$. Then the corresponding linear moment min-problem is locally equivalent to the power moment min-problem with gaps (3.2) [18]. Hence, under the conditions of the theorem, nonlinear moment min-problem (3.1) for a series S representing the system is locally equivalent to the power moment min-problem with gaps (3.2) as well. Without loss of generality, we may assume that the transformation Φ satysfying (3.3), (3.4) is the identity one. Our aim is to show that S is ess-linear and $\{v_{\ell_i}\}_{i=1}^n$ are the first n linearly independent elements from the sequence $\{v_j\}_{j=0}^{\infty}$.

Let $(\theta_s, u_s(t))$ and $(\theta_s^{Lin}, u_s^{Lin}(t))$ be solutions of min-problems (3.1) and (3.2), respectively. Then for any s from a neighborhood of the origin we have

$$s = \sum_{i=1}^{n} e_i \xi_{\ell_i}(\theta_s^{Lin}, u_s^{Lin}) = \sum_{m=1}^{\infty} \sum_{\substack{m_1+\cdots+m_k+k=m \\ k\ge 1, m_j\ge 0}} v_{m_1\ldots m_k} \xi_{m_1\ldots m_k}(\theta_s, u_s), \tag{4.7}$$

where $e_i = (0, \ldots, 1, \ldots, 0)$ with 1 on the ith place. Further, (3.3), (3.4) give

$$\xi_{m_1\ldots m_k}(\theta_s, u_s) - \xi_{m_1\ldots m_k}(\theta_s^{Lin}, u_s^{Lin}) = \bar{o}((\theta_s^{Lin})^m) \quad \text{as} \quad s \to 0, \tag{4.8}$$

$m = m_1 + \cdots + m_k + k$. Besides, since $\xi_{m_1\ldots m_k}(\theta, u) = \theta^m \xi_{m_1\ldots m_k}(1, \hat{u})$, where $\hat{u}(t) = u(t\theta)$, $t \in [0, 1]$, then estimate (2.5) for the series S implies

$$\left\| \sum_{\substack{m_1+\cdots+m_k+k=m \\ k\ge 1, m_j\ge 0}} v_{m_1\ldots m_k} \xi_{m_1\ldots m_k}(\theta_s, u_s) \right\| \le (K_1\theta_s)^m \tag{4.9}$$

as $\theta_s \in [0, T_0)$, $T_0 > 0$, where the constant K_1 does not depend on s.

Denote by U^j, $j \ge 0$, the set of all functions $u_j(t) \in L_\infty[0, 1]$, $\|u_j\| = 1$, having no more than j points of discontinuity. When s runs through the domain $\{s \in \mathbb{R}^n : \theta_s^{Lin} < T_0\}$, the function $u_s^{Lin}(t\theta_s^{Lin})$ runs through the set U^{n-1}. Hence, we infer from (4.7)–(4.9)

$$\begin{aligned} \sum_{m=1}^{\ell_n+1} \sum_{\substack{m_1+\cdots+m_k+k=m \\ k\ge 1, m_j\ge 0}} \theta^m v_{m_1\ldots m_k} \left(\xi_{m_1\ldots m_k}(1, u_{n-1}) + \bar{o}(1)\right) + \bar{o}(\theta^{\ell_n+1}) \\ = \sum_{i=1}^{n} \theta^{\ell_i+1} e_i \xi_{\ell_i}(1, u_{n-1}) \end{aligned} \tag{4.10}$$

as $\theta \to 0$, which is valid for any function $u_{n-1}(t) \in U^{n-1}$.

Further, we prove that $S \in \mathcal{L}_e$ by induction on *the order* of $v_{m_1\ldots m_k}$ using the following notation.

Notation. We say that two vectors $z_1, z_2 \in \mathbb{R}^n$ are equivalent up to terms of order $m \geq 1$ with respect to the series S and write $z_1 \overset{m,S}{\cong} z_2$ if $z_1 - z_2 \in \operatorname{Lin}\{v_i\}_{i=0}^{m-2}$ as $m \geq 2$ and $z_1 = z_2$ as $m = 1$.

Now let $q = 1$ or $q \geq 2$, $\operatorname{rank}\{v_j\}_{j=0}^{q-2} = r$, $\{v_{\ell_i}\}_{i=1}^{r}$ be the first r linearly independent vectors from the sequence $\{v_j\}_{j=0}^{q-2}$ (if $r > 0$), and

$$v_{\ell_i} \overset{\ell_i+1,S}{\cong} e_i, \quad i = 1, \dots, r, \quad v_{m_1 \dots m_k} \overset{m,S}{\cong} 0, \quad k \geq 2, \quad m_1 + \dots + m_k + k = m \leq q-1.$$

Then (4.10) implies in particular

$$(v_{q-1} - w)\xi_{q-1}(1, u_r) + \sum_{\substack{m_1+\dots+m_k+k=q \\ k \geq 2, m_j \geq 0}} v_{m_1 \dots m_k} \xi_{m_1 \dots m_k}(1, u_r) \overset{q,S}{\cong} 0$$

for any $u_r(t) \in U^r$, where $w = e_{r+1}$ as $q = \ell_{r+1} + 1$ and $w = 0$ otherwise. Due to Lemma 8.3, we get $v_{m_1 \dots m_k} \overset{q,S}{\cong} 0$ for any $k \geq 2$, $m_1 + \dots + m_k + k = q$ and $v_{q-1} \overset{q,S}{\cong} e_{r+1}$ if $q = \ell_{r+1} + 1$. The induction arguments complete the proof. □

REMARK 4.3. *As follows from the proof of Theorem* 4.2*, one can simplify Definition* 3.1*. Namely, in* (3.4) *it is sufficient to require the existence of a function* $u_{\Phi(s)}(t) \in U_{\Phi(s)}$ *such that* $\frac{1}{\theta} \int_0^\theta |u_{\Phi(s)}(t) - u_s^{Lin}(t)| dt \to 0$.

5. Essentially linearizable series and condition (E). The result of this section is the following theorem.

THEOREM 5.1. *A series* S *representing a system of the form* (1.7) *is ess-linearizable if and only if this system is null-controllable with respect to the first approximation and satisfies condition* (E).

Proof of necessity. Let a series representing a systems of the form (1.7) be ess-linearizable. Then, in particular, it is linearly nonsingular, and hence, due to (1.11) the system is null-controllable with respect to the first approximation.

Note that condition (E) is satisfied for systems represented by ess-linear series. On the other hand, the property of a system to satisfy condition (E) is invariant with respect to nonsingular analytic substitutions of variables. In fact, if F is an analytic transformation and the system (1.7) in the new variables $z = F(x)$ takes the form $\dot{z} = \tilde{a}(t, z) + \tilde{b}(t, z)u$, then

$$\operatorname{ad}^m_{R_{\tilde{a}}} R_{\tilde{b}} E(z)\Big|_{\substack{t=0 \\ z=0}} = D^{(1)}F(0) \operatorname{ad}^m_{R_a} R_b E(x)\Big|_{\substack{t=0 \\ x=0}}, \qquad m \geq 0,$$

$$\begin{aligned} &\left[\operatorname{ad}^{m_1}_{R_{\tilde{a}}} R_{\tilde{b}}, \dots, \left[\operatorname{ad}^{m_{k-1}}_{R_{\tilde{a}}} R_{\tilde{b}}, \operatorname{ad}^{m_k}_{R_{\tilde{a}}} R_{\tilde{b}}\right] \cdots\right] E(z)\Big|_{\substack{t=0 \\ z=0}} \\ &= D^{(1)}F(0)\left[\operatorname{ad}^{m_1}_{R_a} R_b, \cdots \left[\operatorname{ad}^{m_{k-1}}_{R_a} R_b, \operatorname{ad}^{m_k}_{R_a} R_b\right] \cdots\right] E(x)\Big|_{\substack{t=0 \\ x=0}} \end{aligned}$$

for any $k \geq 2$, $m_1, \dots, m_k \geq 0$. Hence condition (E) is satisfied for systems represented by ess-linearizable series as well.

To prove sufficiency we need the following notation.

Notation. We say that a linearly nonsingular series $S^{(p)}$ with coefficients $v^{(p)}_{m_1 \dots m_k}$ is partially ess-linear up to terms of multiplicity p and write $S^{(p)} \in \mathcal{L}^p_e$ if $p = 1$ or $v^{(p)}_{m_1 \dots m_k} \overset{m,S^{(p)}}{\cong} 0$ for k such that $2 \leq k \leq p$, where $m = m_1 + \dots + m_k + k$ (in terms of the notation introduced in the proof of Theorem 4.2).

Sufficiency. Let a series S represent a system of the form (1.7) which is null-controllable with respect to the first approximation and satisfies condition (E). Then S is linearly nonsingular. Let $\{v_{\ell_1},\dots,v_{\ell_n}\}$, $\ell_1<\dots<\ell_n$, be the first n linearly independent vectors from the sequence $\{v_i\}_{i=0}^{\infty}$. Due to Remark 2.5, one can find an ess-linearizing transformation in the form of a polynomial of power ℓ_n+1. We describe such a transformation in the form of *the ess-linearizing algorithm for the series* S.

The 1*st step.* The linear transformation $F_1=(v_{\ell_1},\dots,v_{\ell_n})^{-1}$ maps S to the series $S^{(1)}=F_1(S)=\sum v^{(1)}_{m_1\dots m_k}\xi_{m_1\dots m_k}\in\mathcal{L}_e^1$ such that $v^{(1)}_{\ell_q}=e_q$, $q=1,\dots,n$.

The $(p+1)$*th step.* Assume that after p steps the series S is transformed to the series $S^{(p)}\in\mathcal{L}_e^p$ of the form

$$\begin{aligned}S^{(p)}_q&=\xi_{\ell_q}+\sum_{m=\ell_q+2}^{\ell_n+1}\sum_{k=1}^{\min\{m,p\}}\sum_{m_1+\dots+m_k+k=m}(v^{(p)}_{m_1\dots m_k})_q\xi_{m_1\dots m_k}\\&\quad+\sum_{m=p+1}^{\ell_n+1}\sum_{k=p+1}^{m}\sum_{m_1+\dots+m_k+k=m}(v^{(p)}_{m_1\dots m_k})_q\xi_{m_1\dots m_k}+\rho^{(p)}_q,\end{aligned}\tag{5.1}$$

where $v^{(p)}_{\ell_r}=e_r$, $r=1,\dots,n$, and $\rho^{(p)}_q$ contain terms of order greater than ℓ_n+1,

$$\rho^{(p)}_q=\sum_{m=\ell_n+2}^{\infty}\sum_{m_1+\dots+m_k+k=m}(v^{(p)}_{m_1\dots m_k})_q\xi_{m_1\dots m_k},\qquad q=1,\dots,n.$$

Then the series $S^{(p)}$ represents a system $\dot x=a^{(p)}(t,x)+b^{(p)}(t,x)u$, and this system satisfies condition (E).

Introduce the polynomial transformation F_{p+1} defined as

$$\begin{aligned}&D^{(1)}F_{p+1}(0)=I,\qquad D^{(k)}F_{p+1}(0)=0,\qquad k\ne1,\qquad k\ne p+1,\\&\big(D^{(p+1)}F_{p+1}(0)e_{r_1}\cdots e_{r_{p+1}}\big)_q\\&\qquad=\begin{cases}0 & \text{as}\ \ \ell_q+1<\ell_{r_1}+\dots+\ell_{r_{p+1}}+p+1,\\ -\big(v^{(p)}_{\ell_{r_1}\dots\ell_{r_{p+1}}}\big)_q & \text{as}\ \ \ell_q+1\ge\ell_{r_1}+\dots+\ell_{r_{p+1}}+p+1\end{cases}\\&\quad\text{for}\quad q=1,\dots,n,\qquad 1\le r_1\le\dots\le r_{p+1}\le n,\end{aligned}\tag{5.2}$$

and consider the series $S^{(p+1)}=F_{p+1}(S^{(p)})$. Due to Lemmas 8.4 and 8.5 we get $S^{(p+1)}\in\mathcal{L}_e^{p+1}$.

In other words, for any $q=1,\dots,n$ we subtract the product of the lines of (5.1) with the indices $r_1,\dots,r_{p+1}$ multiplied by $(v^{(p)}_{\ell_{r_1}\dots\ell_{r_{p+1}}})_q$ from the qth line if and only if $\ell_q+1\ge\ell_{r_1}+\dots+\ell_{r_{p+1}}+p+1$, $1\le r_1\le\dots\le r_{p+1}\le n$, $q=1,\dots,n$. As $v^{(p)}_{\ell_r}=e_r$, $r=1,\dots,n$; then this allows us to exclude from the qth equality of (5.1) the moments $\xi_{\ell_{r_1}\dots\ell_{r_{p+1}}}$ such that $1\le r_1\le\dots\le r_{p+1}\le n$, where $\ell_q+1\ge\ell_{r_1}+\dots+\ell_{r_{p+1}}+p+1$, $q=1,\dots,n$. Conditions (i) and (ii) of Lemma 8.4 (satisfied due to Lemma 8.5) obviously yield that all moments of the form $\xi_{m_1\dots m_{p+1}}$ such that $m_1+\dots+m_{p+1}+p+1\le\ell_q+1$, $m_1,\dots,m_{p+1}\ge0$, become excluded from the qth equality too; therefore, $S^{(p+1)}\in\mathcal{L}_e^{p+1}$.

After ℓ_n+1 *steps* of the algorithm, one obtains the ess-linearizing transformation $F=F_{\ell_n+1}\circ\dots\circ F_1$ since $S^{(\ell_n+1)}\in\mathcal{L}_e$, and, therefore, S is ess-linearizable. □

6. Proof of Theorem 1.4 and examples. The proof of Theorem 1.4 follows from Theorems 4.2 and 5.1 immediately. Really, let a system of the form (1.7) be null-controllable with respect to the first approximation. Due to Theorem 4.2, the nonlinear time-optimal control problem for this system is equivalent to a certain linear problem if and only if the series representing the system is ess-linearizable. At the same time, Theorem 5.1 yields that this series is ess-linearizable if and only if the system satisfies condition (E). The construction of the polynomial transformation $F = \Phi^{-1}$ may be found in the proof of Theorem 5.1.

Moreover, as follows from the proof of Theorem 4.2, the equivalent linear time-optimal control problem corresponds to the moment min-problem (3.2), where $\ell_1, \dots, \ell_n$ are indices of the first n linearly independent vectors from the sequence $\{v_i\}_{i=0}^{\infty}$. Taking into account (1.11), we get that the time-optimal control problem for the linearization corresponds to the moment min-problem which is equivalent to (3.2) as well. Hence, if the nonlinear time-optimal control problem is equivalent to a certain linear problem, then it is equivalent to the problem for the linearization as well.

The fact that the solution of the problem (3.2) may serve as an approximation of the optimal time and control for the original time-optimal control problem in a neighborhood of the origin (after the change of variables) immediately follows now from Definitions 1.1 and 1.3 and Theorem 1.2. □

Consider now special classes of systems satisfying conditions of Theorem 1.4.

(i) Consider a system of the form

$$\dot{x} = A(t)x + f(t)(u + g(t,x)), \qquad g(t,0) \equiv 0, \tag{6.1}$$

where $x \in \mathbb{R}^n$, $u \in \mathbb{R}^1$, $A(t)$, $f(t)$, and $g(t,x)$ are a matrix, an n-dimensional vector, and a function real analytic in neighborhoods of the origin in $\mathbb{R}^1$ and $\mathbb{R}^{n+1}$, respectively. The system (6.1) is of the form (1.7) where $a(t,x) = A(t)x + f(t)g(t,x)$, $b(t,x) = f(t)$. Obviously,

$$\operatorname{ad}_{R_a} R_b E(x) = \mu_1^1(t,x)\Delta f(t) + \mu_0^1(t,x) f(t),$$

where $\Delta = \frac{d}{dt} - A(t)$, $\mu_1^1(t,x) \equiv 1$, $\mu_0^1(t,x) = -\sum_{i=1}^{n} \frac{\partial}{\partial x_i} g(t,x) f_i(t)$. It is easy to see that for any $m \geq 0$,

$$\operatorname{ad}_{R_a}^m R_b E(x) = \sum_{j=0}^{m} \mu_j^m(t,x)\Delta^j f(t),$$

where $\mu_m^m(t,x) \equiv 1$ and $\mu_j^m(t,x)$ are real analytic functions, $j = 0, \dots, m-1$.

Suppose the system $\dot{x} = A(t)x + f(t)u$ is locally controllable, which yields $\operatorname{rank}\{\Delta^m f(0)\}_{m=0}^{\infty} = n$. Then $\operatorname{rank}\{\operatorname{ad}_{R_a}^m R_b E(x)_{|\substack{t=0\\x=0}}\}_{m=0}^{\infty} = n$, so the series of nonlinear power moments representing system (6.1) is linearly nonsingular. Further,

$$\left[\operatorname{ad}_{R_a}^{m_1} R_b, \cdots \left[\operatorname{ad}_{R_a}^{m_{k-1}} R_b, \operatorname{ad}_{R_a}^{m_k} R_b\right] \cdots\right] E(x) = \sum_{j=0}^{M} \mu_j^{m_1 \dots m_k}(t,x)\Delta^j f(t),$$

where $M = \max\{m_1, \dots, m_k\} \leq m_1 + \cdots + m_k + k - 2$ as $k \geq 2$ and $\mu_j^{m_1 \dots m_k}(t,x)$ are real analytic functions, $j = 0, \dots, M$, $k \geq 2$, $m_1, \dots, m_k \geq 0$. Hence condition (E) is satisfied and the time-optimal control problem for system (6.1) is locally equivalent to the linear time-optimal problem for the system $\dot{x} = A(t)x + f(t)u$.

(ii) Consider an autonomous "shifted" bilinear system of the form

$$\dot{x} = Ax + (Bx + f)u, \tag{6.2}$$

where $x \in \mathbb{R}^n$, $u \in \mathbb{R}^1$, $f \in \mathbb{R}^n$. We have $a(x) = Ax$, $b(x) = Bx + f$. Denote by $[A, B]$ the matrix commutator, $[A, B] = B \cdot A - A \cdot B$, and let $\operatorname{ad}_A^0 B = B$, $\operatorname{ad}_A^{m+1} B = [A, \operatorname{ad}_A^m B]$, $m \geq 1$. Then

$$\operatorname{ad}_{R_a}^m R_b E(x) = \operatorname{ad}_A^m Bx + (-1)^m A^m f, \qquad m \geq 0,$$

$$\left[\operatorname{ad}_{R_a}^{m_1} R_b, \cdots \left[\operatorname{ad}_{R_a}^{m_{k-1}} R_b, \operatorname{ad}_{R_a}^{m_k} R_b\right] \cdots\right] E(x)$$

$$= \left[\operatorname{ad}_A^{m_1} B, \cdots \left[\operatorname{ad}_A^{m_{k-1}} B, \operatorname{ad}_A^{m_k} B\right] \cdots\right] x + \sum_{(p_1,\ldots,p_k)\in Z} \mu^{p_1\cdots p_k} \operatorname{ad}_A^{p_1} B \cdots \operatorname{ad}_A^{p_{k-1}} B \cdot A^{p_k} f,$$

where Z denotes the set of all permutations of numbers $m_1, \ldots, m_k$ and $\mu^{p_1\cdots p_k} \in \mathbb{R}$, $k \geq 2$, $m_1, \ldots, m_k \geq 0$.

Now let the system (6.2) be of the form

$$\begin{aligned} \dot{x}_1 &= u(1 + B_{11}x_1 + \cdots + B_{1n}x_n), \\ \dot{x}_q &= x_{q-1} + u(B_{qq}x_q + \cdots + B_{qn}x_n), \qquad 2 \leq q \leq n. \end{aligned} \tag{6.3}$$

Then $\operatorname{ad}_{R_a}^i R_b E(0) = (-1)^i e_{i+1}$, $i = 0, \ldots, n-1$. Further, elements of matrices $\operatorname{ad}_A^p B$, $p \geq 0$, satisfy the condition $(\operatorname{ad}_A^p B)_{ij} = 0$ as $i > j + p$, $1 \leq i, j \leq n$. Hence $\operatorname{ad}_A^p B e_q \in \operatorname{Lin}\{e_j\}_{j=1}^{q+p}$, $q = 1, \ldots, n-p$, $0 \leq p \leq n-1$, and therefore

$$\operatorname{ad}_A^{p_1} B \cdots \operatorname{ad}_A^{p_{k-1}} B \cdot A^{p_k} f \in \operatorname{Lin}\{\operatorname{ad}_{R_a}^i R_b E(0)\}_{i=0}^M,$$

where $M = p_1 + \cdots + p_k = m_1 + \cdots + m_k \leq m_1 + \cdots + m_k + k - 2$ as $k \geq 2$, which proves (E). Thus the time-optimal control problem for system (6.3) is locally equivalent to the linear problem for the system $\dot{x}_1 = u$, $\dot{x}_q = x_{q-1}$, $q = 2, \ldots, n$.

(iii) Consider the nonlinear triangular system

$$\begin{aligned} \dot{x}_q &= \varphi_q(t, x_1, \ldots, x_{q+1}), \qquad 1 \leq q \leq n-1, \\ \dot{x}_n &= \varphi_n(t, x) + \psi(t, x)u, \end{aligned} \tag{6.4}$$

where $\varphi_q(t, x_1, \ldots, x_{q+1})$, $q = 1, \ldots, n-1$, $\varphi_n(t, x)$, and $\psi(t, x)$ are real analytic functions in neighborhoods of the origin in $\mathbb{R}^{q+2}$ and $\mathbb{R}^{n+1}$, respectively, $\varphi_q(t, 0) \equiv 0$, $q = 1, \ldots, n$. Suppose also $\frac{\partial}{\partial x_{q+1}}\varphi_q(0, 0) \neq 0$, $q = 1, \ldots, n-1$, $\psi(0, 0) \neq 0$. Under these conditions there exists a nonsingular transformation which reduces system (6.4) to the system $\dot{x}_q = x_{q+1}$, $q = 1, \ldots, n-1$, $\dot{x}_n = \alpha(t, x) + \beta(t, x)u$; $\alpha(t, 0) \equiv 0$, $\beta(0, 0) \neq 0$, and can be found constructively [10]. Let us prove that the time-optimal control problem for system (6.4) is locally equivalent to the problem for the system $\dot{x}_q = x_{q+1}$, $q = 1, \ldots, n-1$, $\dot{x}_n = u$.

We have $a(t, x) = (\varphi_1(t, x_1, x_2), \ldots, \varphi_n(t, x))$; hence

$$a_x(t, x) = \begin{pmatrix} * & \frac{\partial\varphi_1(t,x)}{\partial x_2} & 0 & \ldots & 0 \\ * & * & \frac{\partial\varphi_2(t,x)}{\partial x_3} & \ldots & 0 \\ \ldots & \ldots & \ldots & \ldots & \ldots \\ * & * & * & \ldots & \frac{\partial\varphi_{n-1}(t,x)}{\partial x_n} \\ * & * & * & \ldots & \frac{\partial\varphi_n(t,x)}{\partial x_n} \end{pmatrix},$$

and therefore $\mathrm{ad}_{R_a} R_b E(x) = \frac{\partial \varphi_{n-1}(t,x)}{\partial x_n}\psi(t,x)e_{n-1} + \mu_n^1(t,x)e_n$. One easily proves that

$$\mathrm{ad}_{R_a}^m R_b E(x) = \prod_{j=n-m}^{n-1} \frac{\partial \varphi_j(t,x)}{\partial x_{j+1}}\psi(t,x)e_{n-m} + \sum_{j=1}^{m} \mu_{n+1-j}^m(t,x)e_{n+1-j}, \quad 1 \le m \le n-1.$$

Thus, $\mathrm{rank}\{\mathrm{ad}_{R_a}^m R_b E(x)_{|\substack{t=0\\x=0}}\}_{m=0}^{n-1} = n$ and $\mathrm{Lin}\{\mathrm{ad}_{R_a}^m R_b E(x)_{|\substack{t=0\\x=0}}\}_{m=0}^{M} = \mathrm{Lin}\{e_j\}_{j=n-M}^{n}$, $0 \le M \le n-1$. Further,

$$\left[\mathrm{ad}_{R_a}^{m_1} R_b, \cdots \left[\mathrm{ad}_{R_a}^{m_{k-1}} R_b, \mathrm{ad}_{R_a}^{m_k} R_b\right] \cdots\right] E(x) = \sum_{j=0}^{\min\{M,n-1\}} \mu_{n-j}^{m_1 \ldots m_k}(t,x)e_{n-j},$$

where $M = \max\{m_1, \ldots, m_k\} \le m_1 + \cdots + m_k + k - 2$ as $k \ge 2$, which proves the validity of (E) and, therefore, the desired result.

7. Some ways of generalization and arising problems.

(A) Within the moment approach, the further step which seems to be natural is to consider series of nonlinear power moments, transformations of series, and nonlinear moment min-problems irrespective to systems and the time-optimal control problem, as it has been done for the linear case [11]. In this way a natural object is a series of the form (2.7) with *arbitrary* constant vector coefficients $v_{m_1 \ldots m_k}$ satisfying the requirement (2.5) (which implies the convergence) and the Markov moment min-problem (3.1) for this series. In this relation it is extremely important to emphasize that, in contrast to the linear case, not every such series represents a system of the form (1.7). The first difficulty to emerge is the question on the existence of a solution of the min-problem. To avoid it one can modify the definition of the local equivalence of the nonlinear moment min-problem to a linear one. We require:

(i) the "local controllability" property: for any s from a neighborhood of the origin there exists θ such that the set $U_s(\theta)$ of all $u(t) \in L_\infty[0,\theta]$, $\|u\| \le 1$, satisfying (3.1) is not empty;

(ii) the existence of an approximation $\tilde{\theta}_{\Phi(s)}$ of $\theta_{\Phi(s)}^{inf} = \inf\{\theta : U_{\Phi(s)}(\theta) \ne \emptyset\}$ such that $U_{\Phi(s)}(\tilde{\theta}_{\Phi(s)}) \ne \emptyset$ and

$$\frac{\theta_s^{Lin}}{\tilde{\theta}_{\Phi(s)}} \to 1, \qquad \frac{\tilde{\theta}_{\Phi(s)}}{\theta_{\Phi(s)}^{inf}} \to 1, \qquad \sup_{\tilde{u}(t) \in U_{\Phi(s)}(\tilde{\theta}_{\Phi(s)})} \frac{1}{\theta}\int_0^\theta \left|u_s^{Lin}(t) - \tilde{u}(t)\right| dt \to 0 \quad \text{as} \quad s \to 0,$$

where $\theta = \min\{\theta_s^{Lin}, \theta_{\Phi(s)}^{inf}\}$.

With this definition the sufficiency in Theorem 4.2 remains true:

(i) If the series S is ess-linearizable (Definition 4.1) and $\{v_{\ell_i}\}_{i=1}^n$ are the first n linearly independent vectors from the sequence $\{v_i\}_{i=0}^\infty$, then the nonlinear moment min-problem is locally equivalent to the linear one of the form (3.2). The necessity in Theorem 4.2 can be proved in the following particular case.

(ii) If the nonlinear moment min-problem is locally equivalent to the linear min-problem of the form

$$s_q = \int_0^\theta t^{q-1}u(t)dt, \qquad q = 1, \ldots, n,$$

then the series S is ess-linearizable and $\mathrm{rank}\{v_i\}_{i=0}^{n-1} = n$.

However, in the general case the necessity in Theorem 4.2 is false. For example, the one-dimensional moment min-problems $s = \xi_{000} + \xi_2$ and $s = \xi_2$ are equivalent although the series $S = \xi_{000} + \xi_2$ is not ess-linearizable.

(B) The representation of systems in the form of series of nonlinear power moments is easily extended to the case of multi-input systems,

$$\dot{x} = a(t,x) + \sum_{i=1}^{h} b_i(t,x)u_i. \tag{7.1}$$

For such systems a representation of the form (2.2) becomes an expansion with respect to nonlinear power "multimoments,"

$$\xi^{i_1\ \dots\ i_k}_{m_1 \dots m_k} = \int_0^\theta \int_0^{\tau_1} \dots \int_0^{\tau_{k-1}} \tau_1^{m_1}\tau_2^{m_2}\dots\tau_k^{m_k} \prod_{j=1}^{k} u_{i_j}(\tau_j)\,d\tau_k \dots d\tau_2 d\tau_1,$$

with coefficients analogous to (2.4) but including $\operatorname{ad}^m_{R_a} R_{b_i}$, $m \geq 0$, $i = 1, \dots, h$. Condition (E) is generalized in a natural way as well.

Finally, one may consider series of nonlinear power "multimoments" with arbitrary constant coefficients satisfying an estimate analogous to (2.5). We call the series ess-linearizable if it can be transformed to the form

$$F(S)_q = \xi^{i_q}_{\ell_q} + \sum_{j=i_q+1}^{h} \alpha_{qj}\xi^{j}_{\ell_q} + \rho_q, \qquad q = 1, \dots, n,$$

where ρ_q includes terms of order greater than $\ell_q + 1$. Here $\ell_q \geq 0$, $i_q \in \{1, \dots, h\}$, and $(\ell_1, i_1) < \dots < (\ell_n, i_n)$ in the lexicographic sense.

For systems of the form (7.1), the result analogous to Theorem 5.1 is valid.

(i) A series of nonlinear power "multi-input" moments representing a system of the form (7.1) is ess-linearizable if and only if the system satisfies condition (E).

For series with arbitrary coefficients, we have the following assertions.

(ii) For a linearly nonsingular series of nonlinear power "multi-input" moments with arbitrary coefficients, the ess-linearizing algorithm can be applied; it gives an ess-linear series after a finite number of steps if and only if the series is ess-linearizable.

(iii) The series S is ess-linearizable if and only if each series $S^{(p)} \in \mathcal{L}^p_e$, $p \geq 1$, constructed by the ess-linearizing algorithm satisfies the conditions of Lemma 8.4 and, therefore, can be transformed to $S^{(p+1)} \in \mathcal{L}^{p+1}_e$ using a transformation analogous to (5.2).

(C) One of our main ideas is to interpret "the closeness" of a given nonlinear system to a linear one in terms of the reducibility of its series to the form $S = \Xi + R$, where Ξ is a vector of linear moments and $R_q/\Xi_q \to 0$ as $\theta \to 0$, $\|u\| \leq 1$. This suggests considering the much more general problem on classification of systems of the form (7.1). Namely, one can regard two systems as equivalent if their series are reducible to the same "canonical form" $S = C + R$ with the "principal part" C, where the components C_q are linear combinations of certain moments and $R_q/C_q \to 0$ as $\theta \to 0$, $\|u\| \leq 1$. The question is how to choose "principal parts" which represent classes of equivalence and systems corresponding to them. In the forthcoming paper, we give such a classification in terms of some structures induced by systems in *the algebra of nonlinear power "multimoments."*

Another important question is to investigate the realizability of series of nonlinear power "multimoments" as systems of the form (7.1):

(i) under what conditions on coefficients of the series there exist such vector functions $a(t,x), b_1(t,x), \dots, b_h(t,x)$ that the series represents (7.1);

(ii) under what conditions these vector functions are time-independent.

This question will be a subject of a forthcoming paper as well.

8. Auxiliary results. This section contains the technical results used in sections 4 and 5.

Denote by U^j, $j \geq 0$, the set of all functions $u_j(t) \in L_\infty[0,1]$, $\|u_j\| = 1$, having no more than j points of discontinuity.

LEMMA 8.1. *Let $r \geq 2$, $0 = \alpha_0 < \alpha_1 < \alpha_2 \leq \cdots \leq \alpha_{r-1} \leq \alpha_r = 1$, and*

$$u_r^1(t) = \sigma(-1)^{j-1}, \qquad t \in [\alpha_{j-1}, \alpha_j), \qquad j = 1, \dots, r, \qquad \sigma = \pm 1$$

(that is the function u_r^1 has at least one point of discontinuity). Then any moment $\xi_{m_1 \dots m_k}(1, u_r^1)$, $k \geq 2$, can be represented as $\xi_{m_1 \dots m_k}(1, u_r^1) = \sum_{j=0}^{m} \alpha_1^j \varphi_j(\alpha_2, \dots, \alpha_{r-1})$, $m = m_1 + \cdots + m_k + k$, where, in particular,

$$\varphi_j(\alpha_2, \dots, \alpha_{r-1}) = \begin{cases} 0 & \text{as } 0 < j \leq m_k, \\ \frac{2\sigma}{m_k+1} \xi_{m_1 \dots m_{k-1}}(1, u_{r-1}) & \text{as } j = m_k + 1, \end{cases}$$

and the function $u_{r-1}(t) \in U^{r-1}$ is defined as follows:

$$u_{r-1}(t) = -\sigma, \quad t \in [0, \alpha_1), \quad u_{r-1}(t) = \sigma(-1)^{j-1}, \quad t \in [\alpha_{j-1}, \alpha_j), \quad j = 2, \dots, r.$$

To formulate the following lemma, we introduce the notations

$$\begin{aligned} \prod_{j=\ell}^{k} \operatorname{ad}_{R_a}^{m_j} R_b &= \operatorname{ad}_{R_a}^{m_\ell} R_b \circ \cdots \circ \operatorname{ad}_{R_a}^{m_k} R_b, \ 1 \leq \ell \leq k, \\ \left[\operatorname{ad}_{R_a}^{m_j} R_b\right]_{j=\ell}^{k} &= [\operatorname{ad}_{R_a}^{m_\ell} R_b, \cdots [\operatorname{ad}_{R_a}^{m_{k-1}} R_b, \operatorname{ad}_{R_a}^{m_k} R_b] \cdots], \qquad 1 \leq \ell \leq k-1. \end{aligned} \tag{8.1}$$

LEMMA 8.2. *Suppose that a series S of the form (2.7) represents a system of the form (1.7), $q \geq 2$, $\operatorname{rank}\{v_i\}_{i=0}^{q-2} = r$, and $\ell_1 < \cdots < \ell_r$ are indices of the first r linearly independent vectors from the sequence $\{v_i\}_{i=0}^{q-2}$. Let also*

$$v_{m_1 \dots m_k} \overset{m,S}{\cong} 0$$

for any $k \geq 2$, $m_1, \dots, m_k \geq 0$ such that $m_1 + \cdots + m_k + k = m \leq q-1$. Then

$$\prod_{j=1}^{i-1} \operatorname{ad}_{R_a}^{m_j} R_b \circ \left[\operatorname{ad}_{R_a}^{m_j} R_b\right]_{j=i}^{p-1} \circ \prod_{j=p}^{k} \operatorname{ad}_{R_a}^{m_j} R_b E(x)\Big|_{\substack{t=0 \\ x=0}} \overset{q,S}{\cong} 0$$

for any $k \geq 3$, $m_1, \dots, m_k \geq 0$ such that $m_1 + \cdots + m_k + k = q$ and any $p = 3, \dots, k$, $i = 1, \dots, p-2$. As a consequence,

(i) *for any $k \geq 3$, $m_1, \dots, m_k \geq 0$ such that $m_1 + \cdots + m_k + k = q$, and any permutation $Z(1, \dots, k-1)$ of numbers $\{1, \dots, k-1\}$ we have*

$$v_{m_1 \dots m_k} - v_{Z(m_1, \dots, m_{k-1}) m_k} \overset{q,S}{\cong} 0;$$

(ii) *for any* $k \geq 2$, $m_1, \dots, m_k \geq 0$ *such that* $m_1 + \dots + m_k + k = q$ *and* $\{m_1, \dots, m_{k-1}\} \not\subset \{\ell_1, \dots, \ell_r\}$, *we have* $v_{m_1 \dots m_k} \overset{q,S}{\cong} 0$.

LEMMA 8.3 (on equivalence of series having the same values on bang-bang controls). *Let a series* S *of the form* (2.7) *represent a system of the form* (1.7). *Let* $q \geq 2$ *be such a number that*

$$v_{m_1 \dots m_k} \overset{m,S}{\cong} 0$$

for any $k \geq 2$, $m_1, \dots, m_k \geq 0$ *such that* $m = m_1 + \dots + m_k + k \leq q - 1$. *Put* $r = \operatorname{rank}\{v_j\}_{j=0}^{q-2}$, *and suppose*

$$(v_{q-1} - w)\xi_{q-1}(1, u_r) + \sum_{\substack{m_1 + \dots + m_k + k = q \\ k \geq 2, m_j \geq 0}} v_{m_1 \dots m_k} \xi_{m_1 \dots m_k}(1, u_r) \overset{q,S}{\cong} 0 \tag{8.2}$$

for any $u_r(t) \in U^r$, *where* $w \in \mathbb{R}^n$. *Then*

$$v_{q-1} \overset{q,S}{\cong} w, \qquad v_{m_1 \dots m_k} \overset{q,S}{\cong} 0 \tag{8.3}$$

for any $k \geq 2$, $m_1, \dots, m_k \geq 0$ *such that* $m_1 + \dots + m_k + k = q$.

Proof. Suppose $r = 0$; then $v_j = 0$ as $j \leq q - 2$ and $v_{m_1 \dots m_k} = 0$ as $k \geq 2$, $m_1 + \dots + m_k + k = q$ due to assertion (ii) of Lemma 8.2. Then (8.2) implies $v_{q-1} = w$, which proves (8.3).

Now let $r \geq 1$. We proceed by induction *on the last index* m_k. Suppose $p = 0$ or $1 \leq p \leq q - 1$ and the second relation of (8.3) holds for any set of indices $m_1, \dots, m_k$ such that $k \geq 2$, $0 \leq m_k \leq p - 1$, $m_1, \dots, m_{k-1} \geq 0$, $m_1 + \dots + m_k + k = q$. Our aim is to show that (8.3) is valid for any $m_1, \dots, m_k$ such that $m_k = p$, $m_1, \dots, m_{k-1} \geq 0$, $m_1 + \dots + m_{k-1} + p + k = q$. We have from (8.2) and the induction supposition that for any $u_r(t) \in U^r$,

$$(v_{q-1} - w)\xi_{q-1}(1, u_r) + \sum_{\substack{m_1 + \dots + m_k + k = q \\ k \geq 2,\ m_1, \dots, m_{k-1} \geq 0, m_k \geq p}} v_{m_1 \dots m_k} \xi_{m_1 \dots m_k}(1, u_r) \overset{q,S}{\cong} 0.$$

Let $u_r(t)$ have at least one point of discontinuity $\alpha_1 \in (0, 1)$. Then, using Lemma 8.1 and the equality $\xi_{q-1}(1, u_r) = \pm\frac{1}{q}(2\sum_{j=1}^{r}(-1)^{j-1}\alpha_j^q + (-1)^r)$, we obtain $v_{q-1} \overset{q,S}{\cong} w$ if $p = q - 1$ (the first of relations (8.3)) and

$$\sum_{\substack{m_1 + \dots + m_{k-1} + p + k = q \\ k \geq 2, m_j \geq 0}} v_{m_1 \dots m_{k-1} p}\ \xi_{m_1 \dots m_{k-1}}(1, u_{r-1}) \overset{q,S}{\cong} 0 \quad \text{for any} \quad u_{r-1}(t) \in U^{r-1} \tag{8.4}$$

if $p \leq q - 2$.

Let $\ell_1 < \dots < \ell_r$ be indices of the first r linearly independent vectors from the sequence $\{v_j\}_{j=0}^{q-2}$. Due to assertion (ii) of Lemma 8.2, (8.4) reads

$$\sum_{\substack{m_1 + \dots + m_{k-1} + p + k = q \\ k \geq 2, m_j \geq 0 \\ m_1, \dots, m_{k-1} \in \{\ell_1, \dots, \ell_r\}}} v_{m_1 \dots m_{k-1} p}\ \xi_{m_1 \dots m_{k-1}}(1, u_{r-1}) \overset{q,S}{\cong} 0. \tag{8.5}$$

Let Z denote the set of all distinct permutations of numbers $m_1, \dots, m_{k-1}$, and N denotes the number of such permutations. Since

$$\sum_{(q_1,\dots,q_{k-1})\in Z} v_{q_1\dots q_{k-1}p}\,\xi_{q_1\dots q_{k-1}} = v_{m_1\dots m_{k-1}p} \sum_{(q_1,\dots,q_{k-1})\in Z} \xi_{q_1\dots q_{k-1}} + \sum_{(q_1,\dots,q_{k-1})\neq(m_1,\dots,m_{k-1})} (v_{q_1\dots q_{k-1}p} - v_{m_1\dots m_{k-1}p})\xi_{q_1\dots q_{k-1}},$$

and

$$\sum_{(q_1,\dots,q_{k-1})\in Z} \xi_{q_1\dots q_{k-1}} = \frac{N}{(k-1)!}\prod_{i=1}^{k-1}\xi_{m_i},$$

we get from (8.5) and assertion (i) of Lemma 8.2 that

$$\sum_{\substack{m_1+\dots+m_{k-1}+p+k=q\\ k\geq 2, m_j\geq 0\\ m_1\leq\dots\leq m_{k-1}\in\{\ell_1,\dots,\ell_r\}}} v_{m_1\dots m_{k-1}p}\frac{N}{(k-1)!}\prod_{i=1}^{k-1}\xi_{m_i}(\theta,\bar u_{r-1}) \overset{q,S}{\cong} 0 \tag{8.6}$$

for any $\theta\geq 0$ and control $\bar u_{r-1}(t) = u_{r-1}(t/\theta)$, $t\in[0,\theta]$, where $u_{r-1}(t)$ runs through the set U^{r-1}. Since $\{\xi_{\ell_j}(\theta,\bar u_{r-1})\}_{j=1}^r$ are independent as polynomials of r variables (points of discontinuity of $\bar u_{r-1}(t)$ and θ), we have from (8.6) that $v_{m_1\dots m_{k-1}p} \overset{q,S}{\cong} 0$ if $m_1\leq\dots\leq m_{k-1}$ and $m_1,\dots,m_{k-1}\in\{\ell_1,\dots,\ell_r\}$. Finally, Lemma 8.2 yields that (8.3) holds for any $m_1,\dots,m_k$ such that $m_k = p$, $m_1,\dots,m_{k-1}\geq 0$, $m_1+\dots+m_{k-1}+p+k=q$. □

LEMMA 8.4. *Let $S^{(p)}\in\mathcal{L}_e^p$, $p\geq 1$. Denote $S^{(p+1)} = F_{p+1}(S^{(p)})$, where F_{p+1} is defined by* (5.2). *Then $S^{(p+1)}\in\mathcal{L}_e^{p+1}$ under the two following conditions satisfied for any $m_1,\dots,m_{p+1}\geq 0$.*

(i) *For an arbitrary permutation $Z(1,\dots,p+1)$ of numbers $\{1,\dots,p+1\}$*

$$v^{(p)}_{m_1\dots m_{p+1}} \overset{m,S^{(p)}}{\cong} v^{(p)}_{Z(m_1,\dots,m_{p+1})},\quad m = m_1+\dots+m_{p+1}+p+1.$$

(ii) *If $v^{(p)}_{m_1} \overset{m,S^{(p)}}{\cong} \sum_{\ell=1}^{n_1}\alpha_\ell v^{(p)}_{q_\ell}$, then $v^{(p)}_{m_1\dots m_{p+1}} \overset{m,S^{(p)}}{\cong} \sum_{\ell=1}^{n_1}\alpha_\ell v^{(p)}_{q_\ell m_2\dots m_{p+1}}$, where $m = m_1+\dots+m_{p+1}+p+1$, $\alpha_\ell\in\mathbb{R}$.*

LEMMA 8.5. *Let $S^{(p)}\in\mathcal{L}_e^p$ represent a control system $\dot x = a^{(p)}(t,x)+b^{(p)}(t,x)u$, which satisfies condition* (E). *Then (under the notations analogous to* (8.1)*) the following equalities hold:*

$$\prod_{j=1}^{\ell}\mathrm{ad}^{m_j}_{R_{a^{(p)}}}R_{b^{(p)}}\circ\Big[\mathrm{ad}^{m_j}_{R_{a^{(p)}}}R_{b^{(p)}}\Big]^q_{j=l+1}\circ\prod_{j=q+1}^{p+1}\mathrm{ad}^{m_j}_{R_{a^{(p)}}}R_{b^{(p)}}E(x)\Big|_{\substack{t=0\\ x=0}} \overset{m,S^{(p)}}{\cong} 0$$

for any $q=2,\dots,p+1$, $\ell=0,\dots,q-2$; $m_1,\dots,m_{p+1}\geq 0$, where $m = m_1+\dots+m_{p+1}+p$ as $q<p+1$ and $m_1+\dots+m_{p+1}+p+1$ as $q=p+1$. As a consequence, conditions (i) *and* (ii) *of Lemma* 8.4 *are satisfied.*

Acknowledgments. The authors thank the two anonymous referees and the associate editor for their comments and recommendations that significantly improved the exposition of the paper.

REFERENCES

[1] A. A. Agrachev and R. W. Gamkrelidze, *Exponential representation of flows and chronological calculus*, Mat. Sb., 107 (149) (1978), pp. 467–532, 639 (in Russian).

[2] R. W. Brockett, *Volterra series and geometric control theory*, Automatica J. IFAC, 12 (1976), pp. 167–176.

[3] A. F. Filippov, *On certain questions in the theory of optimal control*, Vestnik Moscov. Univ. Ser. I, Mat. Mekh., 2 (1959), pp. 25–32, SIAM J. Control, 1 (1962), pp. 76–84.

[4] M. Fliess, *Fonctionnelles causales non linéaires et indéterminées non commutatives*, Bull. Soc. Math. France, 109 (1981), pp. 3–40.

[5] M. Fliess, *Une approche algébrique du développement fonctionnel des solutions d'équations différentielles non linéaires forcées*, Astérisque, 75–76 (1980), pp. 95–103.

[6] M. Fliess, *Vers une notion de dérivation fonctionnelle causale*, Ann. Inst. H. Poincaré Anal. Non Linéaire, 3 (1986), pp. 67–76.

[7] M. Fliess and S. T. Glad, *An algebraic approach to linear and nonlinear control*, in Essays on Control: Perspectives in the Theory and its Applications, Birkhäuser Boston, Cambridge, MA, 1993, pp. 223–265.

[8] E. G. Gilbert, *Functional expansions for the response of nonlinear differential systems*, IEEE Trans. Automat. Control, AC-22 (1977), pp. 909–921.

[9] A. Isidori, *Nonlinear control systems. An introduction*, Springer-Verlag, New York, 1989.

[10] V. I. Korobov, *Controllability, stability of certain nonlinear systems*, Differentsial'nye Uravneniya, 9 (1973), pp. 614–619 (in Russian).

[11] V. I. Korobov and G. M. Sklyar, *The Markov moment problem on the smallest possible interval*, Dokl. Akad. Nauk SSSR, 308 (1989), pp. 525–528; Soviet Math. Dokl., 40 (1990), pp. 334–337 (in English).

[12] V. I. Korobov and G. M. Sklyar, *Time optimality and the power moment problem*, Mat. Sb., 134 (1987), pp. 186–206 (in Russian); Sb. Math., 62 (1989), pp. 185–205 (in English).

[13] V. I. Korobov and G. M. Sklyar, *Time optimality and the trigonometric moment problem*, Izv. Akad. Nauk SSSR Ser. Mat., 53 (1989), pp. 868–885 (in Russian); Math. USSR Izv., 35 (1990), pp. 203–220 (in English).

[14] V. I. Korobov and G. M. Sklyar, *Markov power min-moment problem with periodic gaps*, J. Math. Sci., 80 (1996), pp. 1559–1581.

[15] M. G. Krein, *L-problem in an abstract linear normalized space*, in On Some Questions of the Moment Theory, N. I. Akhiezer and M. G. Krein, eds., GONTI, Kharkov, Ukraine, 1938, pp. 171–199.

[16] C. Lesiak and A. J. Krener, *The existence and uniqueness of Volterra series for nonlinear systems*, IEEE Trans. Automat. Control, AC-23 (1978), pp. 1090–1095.

[17] A. A. Markov, *Selected Works*, Gostekhizdat, Moscow, 1948.

[18] G. M. Sklyar and S. Yu. Ignatovich, *A classification of linear time-optimal control problems in a neighborhood of the origin*, J. Math. Anal. Appl., 203 (1996), pp. 791–811.

SIAM J. CONTROL OPTIM.
Vol. 38, No. 6, pp. 1729–1756

OBSERVABILITY INEQUALITIES FOR SHALLOW SHELLS*

PENG-FEI YAO†

Abstract. We consider some observability inequalities from boundary for a general shallow shell with a middle surface of any shape. At first, an estimate is established by the geometric multiplier method in the case that no boundary conditions are imposed under some checkable geometric conditions. Then our results yield continuous observability estimates for two kinds of boundary conditions which have a physical meaning with an explicit observability time and hence, by duality, exact controllability results.

Key words. shallow shell, Bochner's technique, observability inequality

AMS subject classifications. 35A, 35l, 35q, 49A, 49B, 49E

PII. S0363012999338692

1. Introduction: Statement of main results.

The purpose of this paper is to establish some observability estimates for the shallow shell from which some boundary exact controllability results can be derived. This problem has been well understood in the case of wave equations and plates and, in particular, in the constant coefficient case; see Komornik [19], Lagnese and Lions [21]. It is, in general, hard to handle the variable coefficient case in which some special tools are often needed in addition to the classical multiplier method, for instance, the microlocal analysis method of Bardos, Lebeau, and Rauch [1], the pseudodifferential method of Tataru [28], [29], [30], and the geometric method of Yao [36]. In the case of thin shells with a middle surface of any shape, very little is apparently known in the context of control/stabilization theory, partly because thin shell problems are always of variable coefficient (at least about space variables). Generally, direct adaptation of the techniques, traditionally developed in assuming that the middle surface is defined by one coordinate, would not be fully adequate when dealing with some observability estimates since the presence of the Christoffel symbols Γ_{jk}^{k} can often make computing too complicated.

First, we shall briefly introduce the classical shallow model in a version produced by Yao [37], in which the middle surface is viewed as a Riemann manifold with the induced metric in $\mathbb{R}^3$. One of the advantages in doing this is to build a bridge to modern geometry. For instance, the Bochner technique, which cannot be applied once fixed in one coordinate, is the key ingredient used throughout this paper to overcome the complexity of computation when we deal with all the estimates. This technique, which describes a method initiated by Bochner some fifty years ago for proving some identities of geometric interest, is not so easily described, but it offers the greatest computational simplification on some variable coefficient problems. For details, we refer to Wu [35].

Next, we set up an assumption (H2) on the middle surface under which an estimate for the shallow model is established in the case that no boundary conditions are imposed (Theorem 1.1). In fact, this assumption also works for some observability inequalities of Naghdi's models; see Yao [40]. In subsection 2.1 we shall show that

*Received by the editors May 19, 1999; accepted for publication (in revised form) November 15, 1999; published electronically June 15, 2000. This work was supported by the National Natural Science Foundation of China and the National Key Project of China.
http://www.siam.org/journals/sicon/38-6/33869.html
†Institute of Systems Science, Academy of Mathematics and Systems Science, Chinese Academy of Science, Beijing 100080, P. R. China (pfyao@iss03.iss.ac.cn).

the main assumption (H2) always holds locally for the middle surface of any shape (Proposition 2.2) and how to find a vector field to meet it when the middle surface of the shallow shell is of constant curvature or revolution (Propositions 2.3 and 2.4). In particular, several examples of the middle surface that verify the main assumption (H2) are presented in subsection 1.4.

Finally, the estimates in Theorem 1.1 will produce continuous observability inequalities in both Dirichlet and Neumann cases (Theorems 1.2 and 1.3), respectively, by a compactness/uniqueness argument to absorb the lower order terms as in Bardos, Lebeau, and Rauch [1] or in Zuazua [41]. Fortunately, all the uniqueness results we need here can, in both cases, be derived from certain old uniqueness issues (Proposition 2.13), i.e., Hörmander [16], or Shirota [27].

In addition, regularities of solutions to all the shallow equations needed here should be an intrinsic issue. Since we are mainly concerned with estimates of inequalities, it is assumed that all the regularities of solutions we need hold in this paper.

We mention that some works have been done on control problems of some special shallow shells, for example, Chen, Coleman, and Liu [8] for circular cylindrical shells, Lasiecka, Triggiani, and Valente [23], Triggiani [32] for spherical shells, and Geymont, Loreti, and Valente [12].

1.1. Some notation. We introduce some notation in preparation for shallow shell equations and the observability inequalities. It is mentioned that all definitions and notation in this subsection are standard and classical in the literature.

Denote the usual inner product in $\mathbb{R}^3$ by $\langle\cdot,\cdot\rangle$, i.e., the dot product. Let M be a surface in $\mathbb{R}^3$. For simplicity, M is assumed to be smooth. Surface M produces a natural Riemannian manifold of dimension 2 with the induced metric in $\mathbb{R}^3$. We denote this induced metric on surface M by g or by $\langle\cdot,\cdot\rangle$, as convenient. For each $x \in M$, M_x is the tangential space of M at x. It is assumed that surface M is orientable with the unit normal field N on M. Denote the set of all vector fields on M by $\mathcal{X}(M)$. Denote the set of all k-order tensor fields and the set of all k-forms on M by $T^k(M)$ and $\Lambda^k(M)$, respectively, where k is a nonnegative integer. Then

$$\Lambda^k(M) \subset T^k(M). \tag{1.1}$$

In particular, $\Lambda^0(M) = T^0(M) = C^\infty(M)$ is the set of all C^∞ functions on M and

$$T^1(M) = T(M) = \Lambda(M) = \mathcal{X}(M), \tag{1.2}$$

where $\Lambda(M) = \mathcal{X}(M)$ is in the following isomorphism: for $X \in \mathcal{X}(M)$ given,

$$U(Y) = \langle Y, X\rangle \quad \forall\, Y \in \mathcal{X}(M) \tag{1.3}$$

determines a unique $U \in \Lambda(M)$.

It is well known that, for each $x \in M$, k-order tensor space T^k_x on M_x is an inner product space defined as follows. Let e_1, e_2 be an orthonormal basis of M_x. For any α, $\beta \in T^k_x$, $x \in M$, the inner product is given by

$$\langle \alpha, \beta\rangle_{T^k_x} = \sum_{i_1,\dots,i_k=1}^{2} \alpha(e_{i_1},\dots,e_{i_k})\beta(e_{i_1},\dots,e_{i_k}) \quad \text{at } x. \tag{1.4}$$

In particular, for $k = 1$ definition (1.4) becomes

$$g(\alpha,\beta) = \langle \alpha, \beta\rangle_{T_x} = \langle \alpha, \beta\rangle \quad \forall\, \alpha,\, \beta \in M_x, \tag{1.5}$$

that is, the induced inner product of M_x in $\mathbb{R}^3$.

Let Ω be a bounded region of surface M with a regular boundary Γ or without boundary (when Γ is empty). From (1.4), $T^k(\Omega)$ are then inner product spaces in the following sense:

$$(T_1, T_2)_{T^k(\Omega)} = \int_\Omega \langle T_1, T_2 \rangle_{T_x^k}\, dx, \quad T_1,\, T_2 \in T^k(\Omega), \tag{1.6}$$

where dx is the volume element of surface M in its Riemannian metric g.

The completions of $T^k(\Omega)$ in inner products (1.6) are denoted by $L^2(\Omega, T^k)$. In particular, $L^2(\Omega, \Lambda) = L^2(\Omega, T)$. $L^2(\Omega)$ is the completion of $C^\infty(\Omega)$ in the following inner product:

$$(f, h)_{L^2(\Omega)} = \int_\Omega f(x)h(x)\, dx, \quad f,\, h \in C^\infty(\Omega). \tag{1.7}$$

Let D be the Levi–Civita connection on M in the induced metric g of surface M. For $U \in \mathcal{X}(M)$, DU is the covariant differential of U which is a 2-order covariant tensor field in the following sense:

$$DU(X, Y) = D_Y U(X) = \langle D_Y U,\, X \rangle \quad \forall\, X,\, Y \in M_x,\, x \in M. \tag{1.8}$$

We also define $D^*U \in T^2(M)$ by

$$D^*U(X, Y) = DU(Y, X) \quad \forall\, X,\, Y \in M_x,\, x \in M, \tag{1.9}$$

that is, $D^*U \in T^2(M)$ is the transpose of DU. For any $T \in T^2(M)$, the trace of T at $x \in M$ is defined by

$$\mathrm{tr}T = \sum_{i=1}^{2} T(e_i, e_i) \tag{1.10}$$

where e_1, e_2 is an orthonormal basis of M_x. It is obvious that $\mathrm{tr}T \in C^\infty(M)$ if $T \in T^2(M)$.

For $T \in T^k(M)$ and $X \in \mathcal{X}(M)$, we define $l_X T \in T^{k-1}(M)$ by

$$l_X T(X_1, \ldots, X_{k-1}) = T(X, X_1, \ldots, X_{k-1}) \quad \forall\, X_1, \ldots, X_{k-1} \in \mathcal{X}(M). \tag{1.11}$$

For X, $Y \in \mathcal{X}(M)$, the curvature operator $R_{XY}\colon \mathcal{X}(M) \to \mathcal{X}(M)$ is defined by

$$R_{XY} = -D_X D_Y + D_Y D_X + D_{[X,Y]}, \tag{1.12}$$

where $[X, Y]$ is the Lie product of vector fields X and Y. We have the following identity (see Wu [33, section 2, Lem. 4]):

$$D^2T(\ldots, X, Y) = D^2T(\ldots, Y, X) + (R_{XY}T)(\ldots), \tag{1.13}$$

for $T \in T^k(M)$, $\ldots$, X, $Y \in \mathcal{X}(M)$.

The curvature tensor R of the Levi–Civita connection is given by

$$R(X, Y, Z, W) = \langle R_{XY}Z,\, W \rangle, \quad X,\, Y,\, Z,\, W \in M_x,\, x \in M, \tag{1.14}$$

which is a 4-order tensor field on M.

The Sobolev space $H^k(\Omega)$ is the completion of $C^\infty(\Omega)$ with respect to the norm

$$\|f\|^2_{H^k(\Omega)} = \sum_{i=1}^{k} \|D^k f\|^2_{L^2(\Omega,T^k)} + \|f\|^2_{L^2(\Omega)}, \quad f \in C^\infty(\Omega), \tag{1.15}$$

where $D^k f$ is the kth covariant differential of f in the induced metric g of M, which is a k-order tensor field on Ω, and $\|\cdot\|_{L^2(\Omega,T^k)}$ and $\|\cdot\|_{L^2(\Omega)}$ are the induced norms in inner products (1.6)–(1.7), respectively. For details on Sobolev spaces on Riemannian manifolds, we refer to Hebey [15] or Taylor [31].

Another important Sobolev space for us is $H^k(\Omega,\Lambda)$, defined by

$$H^k(\Omega,\Lambda) = \{U \,|\, U \in L^2(\Omega,\Lambda),\ D^i U \in L^2(\Omega,T^{i+1}),\ 1 \le i \le k\} \tag{1.16}$$

with inner product

$$(U,V)_{H^k(\Omega,\Lambda)} = \sum_{i=0}^{k} (D^i U, D^i V)_{L^2(\Omega,T^{i+1})} \quad \forall\ U,\ V \in H^k(\Omega,\Lambda) \tag{1.17}$$

(for example, see Wu [34]). In particular, $H^0(\Omega,\Lambda) = L^2(\Omega,\Lambda)$.

For $\hat{\Gamma} \subset \Gamma$, set

$$H^1_{\hat{\Gamma}}(\Omega,\Lambda) = \{W \,|\, W \in H^1(\Omega,\Lambda),\ W|_{\hat{\Gamma}} = 0\}; \tag{1.18}$$

$$H^2_{\hat{\Gamma}}(\Omega) = \left\{ w \,|\, w \in H^2(\Omega),\ w|_{\hat{\Gamma}} = \frac{\partial w}{\partial n}|_{\hat{\Gamma}} = 0 \right\}. \tag{1.19}$$

In particular, $H^1_0(\Omega,\Lambda) = H^1_\Gamma(\Omega,\Lambda)$ and $H^2_0(\Omega) = H^2_\Gamma(\Omega)$.

1.2. Model. Let us assume that the middle surface of the shell occupies a bounded region Ω of surface M in $\mathbb{R}^3$. The shell, a body in $\mathbb{R}^3$, is defined by

$$\mathcal{S} = \{\, p \,|\, p = x + zN(x),\ x \in \Omega,\ -h/2 < z < h/2 \,\}, \tag{1.20}$$

where h is the thickness of the shell, i.e., "small"; see Ciarlet and Paumier [9].

Denote by $\zeta(x)$ the displacement vector of point x of the middle surface. We decompose displacement vector ζ into sums:

$$\zeta(x) = W(x) + w(x)N(x), \quad x \in \Omega,\ W(x) \in M_x, \tag{1.21}$$

i.e., W and w are components of ζ on the tangent plane and on the normal of the undeformed middle surface Ω, respectively. The linearized strain tensor and the change of curvature tensor of the middle surface Ω are given by

$$\Upsilon(\zeta) = \frac{1}{2}(DW + D^*W) + w\Pi \tag{1.22}$$

and

$$\rho(\zeta) = -D^2 w \tag{1.23}$$

in a coordinate-free form, respectively, where Π is the second fundamental form of surface M and $D^2 w$ the Hessian of w, which are justified for a shallow shell. For (1.22) and (1.23), we refer to Niordson [25, p. 355] or to Koiter [17, p. 27].

The shell strain energy associated to a displacement field ζ of the middle surface Ω can be written as

$$\mathcal{B}_1(\zeta,\zeta) = \frac{Eh}{1-\mu^2}\int_\Omega B(\zeta,\zeta)\,dx, \tag{1.24}$$

where

$$B(\zeta,\zeta) = a(\Upsilon(\zeta),\Upsilon(\zeta)) + \gamma a(\rho(\zeta),\rho(\zeta)), \quad \gamma = h^2/12, \tag{1.25}$$

$$a(\Upsilon(\zeta),\Upsilon(\zeta)) = (1-\mu)\langle\Upsilon(\zeta),\Upsilon(\zeta)\rangle_{T_x^2} + \mu(\mathrm{tr}\Upsilon(\zeta))^2, \tag{1.26}$$

for $x\in\Omega$, where E, μ denote, respectively, Young's modulus and Poison's coefficient of the material. For (1.24), we refer to Bernadou and Boisserie [3, p. 15].

Thus, with expression (1.24), we are able to associate the following symmetric bilinear form, directly defined on the middle surface Ω:

$$\mathcal{B}(\zeta,\eta) = \int_\Omega B(\zeta,\eta)\,dx, \tag{1.27}$$

where ζ is given in (1.21) and

$$\eta = U + uN, \quad U(x)\in M_x,\ x\in\Omega. \tag{1.28}$$

Denote by H and by k the mean curvature and the Gauss curvature of surface M, respectively. From Yao [37], we have the following Green formula for a shallow shell.

Formula I. Let bilinear form $\mathcal{B}(\cdot,\cdot)$ be given in (1.27). For $\zeta=(W,w)$, $\eta=(U,u)\in H^1(\Omega,\Lambda)\times H^2(\Omega)$, we have

$$\mathcal{B}(\zeta,\eta) = (\mathcal{A}\zeta,\eta)_{L^2(\Omega,\Lambda)\times L^2(\Omega)} + \int_\Gamma \partial(\mathcal{A}\zeta,\eta)\,d\Gamma, \tag{1.29}$$

where

$$\begin{aligned}\partial(\mathcal{A}\zeta,\eta) &= B_1(W,w)\langle U,n\rangle + B_2(W,w)\langle U,\tau\rangle\\ &+\gamma\left[(\Delta w + (1-\mu)B_3 w)\frac{\partial u}{\partial n} - \left(\frac{\partial\Delta w}{\partial n} + (1-\mu)B_4 w\right)u\right];\end{aligned} \tag{1.30}$$

n, τ are the normal and the tangential along curve Γ, respectively;

$$\mathcal{A}\zeta = \begin{pmatrix} -\boldsymbol{\Delta}_\mu W - (1-\mu)kW - \mathcal{F}(w)\\ \gamma[\Delta^2 w - (1-\mu)\delta(kdw)] + (H^2 - 2(1-\mu)k)w + \mathcal{G}(W)\end{pmatrix}, \tag{1.31}$$

$\boldsymbol{\Delta}_\mu$ is of the Hodge–Laplacian type, applied to 1-forms (or equivalently, vector fields), defined by

$$\boldsymbol{\Delta}_\mu = -\left(\frac{1-\mu}{2}\delta d + d\delta\right), \tag{1.32}$$

where d is the exterior differential, δ the formal adjoint of d, Δ the Laplacian on manifold M,

$$\begin{cases}\mathcal{F}(w) = (1-\mu)l_{dw}\Pi + \mu H dw + w dH,\\ \mathcal{G}(W) = (1-\mu)\langle DW,\Pi\rangle_{T_x^2} - \mu H\delta W,\end{cases} \tag{1.33}$$

and

$$\begin{cases} B_1(W,w) = (1-\mu)\Upsilon(\zeta)(n,n) + \mu(wH - \delta W), \\ B_2(W,w) = (1-\mu)\Upsilon(\zeta)(n,\tau), \\ B_3 w = -D^2 w(\tau,\tau), \\ B_4 w = \dfrac{\partial}{\partial\tau}(D^2 w(\tau,n)) + k(x)\dfrac{\partial w}{\partial n}. \end{cases} \tag{1.34}$$

By the "principle of virtual work" and Formula I, we obtain the following displacement equations for a shallow shell (see Yao [37]) after changing t to t/λ with $\lambda^2 E/(1-\mu^2) = 1$.

Formula II. We assume that there are no external loads on the shell and that the shell is clamped along a portion Γ_0 of Γ and is free on Γ_1, where $\Gamma_0 \cup \Gamma_1 = \Gamma$ and $\Gamma_0 \cap \Gamma_1 = \emptyset$. Then the displacement vector $\zeta = (W,w)$ satisfies the following boundary value problem:

$$\left.\begin{aligned} &W_{tt} - [\boldsymbol{\Delta}_\mu W + (1-\mu)kW + F(w)] = 0 \\ &w_{tt} - \gamma\Delta w_{tt} + \gamma\left(\Delta^2 w - (1-\mu)\delta(kdw)\right) \\ &\qquad\qquad +(H^2 - 2(1-\mu)k)w + \mathcal{G}(W)] = 0 \\ &\zeta(0) = \zeta_0, \quad \zeta_t(0) = \zeta_1 \end{aligned}\right\} \quad \text{in } Q_\infty, \tag{1.35}$$

$$\left.\begin{aligned} &W = 0 \\ &w = \frac{\partial w}{\partial n} = 0 \end{aligned}\right\} \quad \text{on } \Sigma_{0\infty}, \tag{1.36}$$

$$\left.\begin{aligned} &B_1(W,w) = B_2(W,w) = 0 \\ &\Delta w + (1-\mu)B_3 w = 0 \\ &\frac{\partial \Delta w}{\partial n} + (1-\mu)B_4 w - \frac{\partial w_{tt}}{\partial n} = 0 \end{aligned}\right\} \quad \text{on } \Sigma_{1\infty}, \tag{1.37}$$

where

$$Q_\infty = \Omega \times (0,\infty), \quad \Sigma_{0\infty} = \Gamma_0 \times (0,\infty), \quad \Sigma_{1\infty} = \Gamma_1 \times (0,\infty). \tag{1.38}$$

Remark 1.1. In the literature, for some special cases the displacement equations are expressed in terms of three displacement components of the shell and their derivatives such as spherical shells (Lasiecka, Trigginia, and Valente [23]) and circular cylindrical ones (Chen, Coleman, and Liu [8]). For a shell with a general middle surface of any shape, this method may not be possible (see some comments by Koiter [18, p. 33]) and we have to draw support from Formula II.

Remark 1.2. If the shell is flat, a plate, (1.35) is uncoupled. The equation on component w is the same as in Lagnese [20, pp. 15–16], a Kirchhoff plate (see Yao [37]).

1.3. Observability inequalities. In obtaining observability inequalities, the ellipticity of the shell strain energy is necessary, which is assumed throughout; that is, there is constant $\lambda_0 \geq 1$ such that (H.1)

$$\lambda_0 \mathcal{B}(\zeta,\zeta) \geq \|DW\|^2_{L^2(\Omega,T^2)} + \gamma\|D^2 w\|^2_{L^2(\Omega,T^2)} \tag{1.39}$$

for $\zeta = (W, w) \in H^1(\Omega, \Lambda) \times H^2(\Omega)$. The above inequality is established if Π and $D\Pi$ are small enough (see Bernadou and Oden [5]) and proved if there is some information on the curvature of the middle surface (Yao [37]).

Main Assumption (H.2). Suppose that there is a vector field $V \in \mathcal{X}(M)$ such that

$$DV(X, X) = b(x)|X|^2, \quad X \in M_x,\ x \in \overline{\Omega}, \tag{1.40}$$

where b is a function on Ω. Set

$$a(x) = \frac{1}{2}\langle DV, \mathcal{E}\rangle_{T_x^2}, \quad x \in \overline{\Omega}, \tag{1.41}$$

where $\mathcal{E}$ is the volume element of M. Moreover, suppose that b and a meet inequality

$$2 \min_{x\in\overline{\Omega}} b(x) > \lambda_0(1+\mu) \max_{x\in\overline{\Omega}} |a(x)|. \tag{1.42}$$

We say that middle surface Ω satisfies assumption (H2) if there is a vector field V such that conditions (1.40) and (1.42) hold.

Set

$$\sigma_0 = \max_{x\in\Omega} |V|; \quad \sigma_1 = \min_{x\in\Omega} b(x) - \frac{\lambda_0(1+\mu)}{2} \max_{x\in\Omega} |a(x)|; \tag{1.43}$$

$$Q = \Omega \times (0, T); \quad \Sigma_0 = \Gamma_0 \times (0, T); \quad \Sigma_1 = \Gamma_1 \times (0, T). \tag{1.44}$$

Remark 1.3. Geometric condition (1.40) is used in Yao [38] for some observability inequalities of the Euler–Bernoulli equation with variable coefficients. If the shell is flat, a plate, then $M = \mathbb{R}^2$. For any $x^0 \in \mathbb{R}^2$, set $V = x - x^0$. It is easily checked that

$$DV = g, \quad x \in \mathbb{R}^2,$$

where g is the dot product in R^2, with $b = 1$ and $a = 0$. For any M, we will show that there always exists a vector field to meet condition (1.40) on Ω (Proposition 2.2 of subsection 2.1). In addition, relations (2.8) in Proposition 2.2 mean that Assumption (H2) always holds locally. Indeed, for any $x_0 \in M$ fixed, by relations (2.8) there are a vector field V, defined by (2.7), and $\epsilon > 0$ such that

$$2 \min_{x\in B(\epsilon)} b(x) > \lambda_0(1+\mu) \max_{x\in B(\epsilon)} |a(x)|,$$

where $B(\epsilon)$ is the geodesic ball with radius ϵ and centered at x_0. The above inequality means that Assumption (H2) is true if middle surface $\Omega \subset B(\epsilon)$.

Let surface M be of constant curvature or revolution. Propositions 2.3 and 2.4 in subsection 2.1 will show that there exists a vector field V such that relation (1.40) holds on the whole surface M with $a(x) = 0\,\forall\, x \in M$. So if middle surface $\Omega \subset M$ such that $b(x) \neq 0\,\forall\, x \in \overline{\Omega}$, then Assumption (H2) holds for Ω.

The total energy of the shell is to be defined by

$$E(t) = \frac{1}{2}[\|W_t\|^2_{L^2(\Omega,\Lambda)} + \|w_t\|^2_{L^2(\Omega)} + \gamma\|Dw_t\|^2_{L^2(\Omega,\Lambda)} + \mathcal{B}(\eta, \eta)]. \tag{1.45}$$

For $\eta = (W, w)$, we set

$$\eta_1 = (W, 0); \quad \eta_2 = (0, w); \tag{1.46}$$

$$L(t) = \|W(t)\|^2_{L^2(\Omega,\Lambda)} + \|w(t)\|^2_{L^2(\Omega)} + \gamma\|w_t(t)\|^2_{L^2(\Omega)} + \|Dw(t)\|^2_{L^2(\Omega,\Lambda)}. \tag{1.47}$$

THEOREM 1.1. *Let Assumptions* (H.1) *and* (H.2) *hold. Let* $\eta = (W, w)$ *solve the problem*

$$\eta_{tt} - \gamma(0, \Delta w_{tt}) + \mathcal{A}\eta = 0 \tag{1.48}$$

such that all the terms on the left-hand side of inequality (1.49) *below are well defined.*

Given $T > 0$*, then for any* $\epsilon > 0$*, there is* $C_\epsilon > 0$*, independent of* η*, such that*

$$SB|_\Sigma + C_\epsilon \left[L(0) + L(T) + \int_0^T L(t)dt \right] + (\sigma_0\lambda_0 + \epsilon)[E(0) + E(T)] \geq \sigma_1 \int_0^T E(t)dt, \tag{1.49}$$

where

$$\begin{aligned} SB|_\Sigma &= \frac{1}{2}\int_\Sigma [|\eta_t|^2 + \gamma|Dw_t|^2 - B(\eta,\eta)]\langle V, n\rangle \, d\Sigma \\ &+ \int_\Sigma \left[\partial\left(\mathcal{A}\eta,\, m(\eta) - \frac{1}{2}b\eta_2 + \frac{1}{2}h\eta_1\right) + \gamma\left(V(w) - \frac{1}{2}bw\right)\frac{\partial w_{tt}}{\partial n} \right] d\Sigma; \end{aligned} \tag{1.50}$$

$$m(\eta) = (D_V W, V(w)); \quad h = 2b - \sigma_1. \tag{1.51}$$

Dirichlet control. First, we consider the Dirichlet mixed problem in unknown $\zeta = (\Phi, \phi)$:

$$\begin{cases} \zeta_{tt} - \gamma(0, \Delta\phi_{tt}) + \mathcal{A}\zeta = 0 \quad \text{in} \quad \Omega \times (0,T), \\ \zeta(0) = \zeta^0, \quad \zeta_t(0) = \zeta^1 \quad \text{on} \quad \Omega, \\ \Phi|_{\Gamma_1} = 0, \quad \Phi|_{\Gamma_0} = U, \quad 0 < t < T, \\ \phi|_{\Gamma_1} = \dfrac{\partial\phi}{\partial n}|_{\Gamma_1} = 0, \quad 0 < t < T, \\ \phi|_{\Gamma_0} = u, \quad \dfrac{\partial\phi}{\partial n}|_{\Gamma_0} = v, \quad 0 < t < T, \end{cases} \tag{1.52}$$

with control functions U, u, and v. Its dual version in $\eta = (W, w)$ follows:

$$\begin{cases} \eta_{tt} - \gamma(0, \Delta w_{tt}) + \mathcal{A}\eta = 0 \quad \text{in} \quad Q, \\ \eta(0) = \eta^0, \quad \eta_t(0) = \eta^1 \quad \text{on} \quad \Omega, \\ W = 0 \quad \text{on} \quad \Sigma, \\ w = \dfrac{\partial w}{\partial n} = 0 \quad \text{on} \quad \Sigma. \end{cases} \tag{1.53}$$

Remark 1.4. In the flat case, for the normal component, one control function $\frac{\partial\phi}{\partial n}|_{\Gamma_0} = v$ is enough; see Lagnese and Lions [21]. We here add another control function $\phi|_{\Gamma_0} = u$ for problem (1.52) in order to avoid the following uniqueness assumption: the problem

$$\begin{cases} \lambda^2\eta - \lambda^2\gamma(0, \Delta w) + \mathcal{A}\eta = 0 \quad \text{on} \quad \Omega, \\ W = 0 \quad \text{on} \quad \Gamma, \\ w = \dfrac{\partial w}{\partial n} = 0 \quad \text{on} \quad \Gamma, \\ D_n W = \Delta w = 0 \quad \text{on} \quad \Gamma_0 \end{cases} \tag{1.54}$$

admits the unique zero solution. The above uniqueness result does not fall into a class of systems to which the Holmgren theorem may be applied even if the coefficients are

analytic since it is not the Cauchy problem for component w(it does for component W). For the flat case, it has been proved in Lagnese and Lions [21].

Continuous observability inequality in the Dirichlet case. From Proposition 2.12 of subsection 2.2, it is easily checked that if $\eta = (W, w)$ solves problem (1.53), then

$$SB|_{\Sigma} = \frac{1}{2}\int_{\Sigma} B(\eta,\eta)\langle V, n\rangle \, d\Sigma. \tag{1.55}$$

The exact controllability of problem (1.52) then leads to the following observability inequality: to seek constant $T_0 > 0$ such that, for any $T > T_0$, there is $c > 0$ satisfying

$$\int_{\Sigma_0} \left[B(\eta,\eta) + \gamma \left(\frac{\partial \Delta w}{\partial n} \right)^2 \right] d\Sigma \geq cE(0), \tag{1.56}$$

where $\eta = (W, w)$ is a solution to problem (1.53) with the initial data $(\eta^0, \eta^1) \in \big(L^2(\Omega,\Lambda) \times H_0^1(\Omega)\big) \times \big(H^{-1}(\Omega,\Lambda) \times L^2(\Omega)\big)$, and

$$\int_{\Sigma_0} B(\eta,\eta)\, d\Sigma = \gamma \int_{\Sigma_0} (\Delta w)^2 \, d\Sigma + \int_{\Sigma_0} \left[(DW(n,n))^2 + \frac{1-\mu}{2} \left(DW(\tau,n)\right)^2 \right] d\Sigma. \tag{1.57}$$

We have the following theorem.

THEOREM 1.2 (Dirichlet case). *Let Assumptions* (H.1) *and* (H.2) *hold. Then for any $T > T_0$, there exists $c > 0$ such that observability inequality* (1.56) *holds, where*

$$T_0 = 2\lambda_0 \sigma_0 / \sigma_1; \tag{1.58}$$

$$\Gamma_0 = \{x \,|\, x \in \Gamma,\, V(x) \cdot n(x) > 0\}. \tag{1.59}$$

Neumann control. Here we let $\overline{\Gamma_1} \neq \emptyset$, $\overline{\Gamma_0} \cap \overline{\Gamma_1} = \emptyset$ and consider problem $\zeta = (\Phi, \phi)$:

$$\begin{cases} \zeta_{tt} - \gamma(0, \Delta\phi_{tt}) + \mathcal{A}\zeta = 0 \quad \text{in} \quad Q, \\ \zeta(0) = \zeta^0, \quad \zeta_t(0) = \zeta^1 \quad \text{on} \quad \Omega. \end{cases} \tag{1.60}$$

We can act on $\Sigma_1 = \Gamma_1 \times (0,T)$ by

$$\begin{cases} \Phi = 0 \quad \text{on} \quad \Sigma_1, \\ \phi = \dfrac{\partial \phi}{\partial n} = 0 \quad \text{on} \quad \Sigma_1, \end{cases} \tag{1.61}$$

and we can act on Σ_0 by

$$\begin{cases} B_1(\Phi,\phi) = u_1 \quad B_2(\Phi,\phi) = u_2 \quad \text{on} \quad \Sigma_0, \\ \Delta\phi + (1-\mu)B_3\phi = v_1 \quad \text{on} \quad \Sigma_0, \\ \dfrac{\partial \Delta\phi}{\partial n} + (1-\mu)B_4\phi - \dfrac{\partial \phi_{tt}}{\partial n} = v_2 \quad \text{on} \quad \Sigma_0. \end{cases} \tag{1.62}$$

The dual problem for the above is the following in $\eta = (W, w)$:

$$\begin{cases} \eta_{tt} - \gamma(0, \Delta w_{tt}) + \mathcal{A}\eta = 0 \quad \text{in} \quad Q, \\ \eta(0) = \eta^0, \quad \eta_t(0) = \eta^1 \quad \text{on} \quad \Omega, \end{cases} \tag{1.63}$$

subject to the boundary condition

$$\begin{cases} W = 0 \quad \text{on} \quad \Sigma_1, \\ w = \dfrac{\partial w}{\partial n} = 0 \quad \text{on} \quad \Sigma_1, \end{cases} \tag{1.64}$$

$$\begin{cases} B_1(W,w) = B_2(W,w) = 0 \quad \text{on} \quad \Sigma_0, \\ \Delta w + (1-\mu)B_3 w = 0 \quad \text{on} \quad \Sigma_0, \\ \dfrac{\partial \Delta w}{\partial n} + (1-\mu)B_4 w - \dfrac{\partial w_{tt}}{\partial n} = 0 \quad \text{on} \quad \Sigma_0. \end{cases} \tag{1.65}$$

Continuous observability inequality in the Neumann case. Let $\eta = (W, w)$ solve problem (1.63)–(1.65). It is easy to check from Proposition 2.12 of subsection 2.2 and boundary conditions (1.64) and (1.65) that

$$SB|_\Sigma = \frac{1}{2}\int_{\Sigma_0} [|\eta_t|^2 + \gamma|Dw_t|^2 - B(\eta,\eta)]\langle V, n\rangle \, d\Sigma + \frac{1}{2}\int_{\Sigma_1} B(\eta,\eta)\langle V, n\rangle \, d\Sigma. \tag{1.66}$$

It follows from (1.66) that to obtain the observability inequality is to seek $T_0 > 0$ such that for any $T > T_0$, there is $c > 0$ satisfying

$$\int_{\Sigma_0} [|\eta_t|^2 + \gamma|Dw_t|^2] \, d\Sigma \geq cE(0) \tag{1.67}$$

for all initial data $(\eta^0, \eta^1) \in \left(H^1_{\Gamma_1}(\Omega,\Lambda) \times H^2_{\Gamma_1}(\Omega)\right) \times \left(L^2(\Omega,\Lambda) \times L^2(\Omega)\right)$ for which the left-hand side of (1.67) is finite.

We have the following theorem.

THEOREM 1.3 (Neumann case). *Let Assumptions* (H.1) *and* (H.2) *hold. Then for any $T > T_0$, there is $c > 0$ such that inequality* (1.67) *holds, where T_0 and Γ_0 are defined by* (1.58) *and* (1.59)*, respectively.*

Remark 1.5. Exact controllability results (in suitable function spaces) for $T > T_0$ follow from (1.56) and (1.67) and duality.

Remark 1.6. Let the shell be flat, that is, $M = \mathbb{R}^2$. Then system (1.35) becomes two systems, where one is a wave equation on component W and the other is a plate on component w. We may take $\lambda_0 = 1$. If we set $V = x - x_0$, x_0 a fixed point in $\mathbb{R}^2$, then inequalities (1.56) and (1.67) on component w are exactly the same as in Lagnese and Lions [21]. In this case, $\sigma_1 = 1$. It follows that $T_0 = 2diameter(\Omega)$, which is the best for wave component W; see Komornik [19]. In this sense, T_0 in (1.58) is the best.

1.4. Some examples. We give here some examples that verify Assumption (H2).

Example 1.1. Let middle surface Ω be of constant curvature. Suppose that the curvature of manifold (M, g) is constant k. Given $x_0 \in M$, let ρ be the distance function from $x \in M$ to x_0 on (M, g), i.e., $\rho(x) = dis(x_0, x)$. Set $V = h(\rho)D\rho$, where $h(\rho)$ is defined by

$$h(\rho) = \begin{cases} \sin(\sqrt{k}\rho(x)), & k > 0, \\ \rho(x), & k = 0, \\ \sinh(\sqrt{-k}\rho(x)), & k < 0. \end{cases} \tag{1.68}$$

By Proposition 2.3 of subsection 2.1, we have

$$DV = b(x)g \quad \text{and} \quad a(x) = 0, \tag{1.69}$$

where

$$b(x) = \begin{cases} \sqrt{k}\cos\left(\sqrt{k}\rho(x)\right), & k > 0, \\ 1, \quad k = 0, \\ \sqrt{-k}\cosh\left(\sqrt{-k}\rho(x)\right), & k < 0. \end{cases} \tag{1.70}$$

It follows that Assumption (H2) holds with vector field V if and only if $\min_{x\in\Omega} b(x) > 0$. By expression (1.70), we have the following conclusions:

(a) If $k > 0$, Assumption (H2) holds when $\overline{\Omega}$ is contained by a geodesic ball with radius $\pi/(2\sqrt{k})$;

(b) If $k \le 0$, Assumption (H2) holds for any $\Omega \subset M$. □

Example 1.2. Let middle surface Ω be a portion of surface M of revolution, where

$$M = \{(x, y, z) \,|\, (x, y) \in \mathbb{R}^2,\ z = \log(1 + x^2 + y^2)\}. \tag{1.71}$$

Set $f(r) = \log(1 + r^2)$. It is easy to check that the curvature is

$$k(p) = -\frac{2(1 + r^2)}{[1 + (1 + r^2)^2]^2} < 0, \tag{1.72}$$

where $p = (x, y, f(r))$ and $r = \sqrt{x^2 + y^2}$ so that inequality (2.25) in Proposition 2.4 holds. By applying Proposition 2.4 of subsection 2.1 to this case, there are a vector field V on M and a function $b(p) > 0$ on M such that

$$DV = b(p)g \quad \text{and} \quad a(p) = 0 \quad \forall p \in M, \tag{1.73}$$

where g is the induced metric of M in $\mathbb{R}^3$ and $a(p) = \frac{1}{2}\langle DV, \mathcal{E}\rangle_{T_p^2}$ is defined by (1.41), that is, Assumption (H2) holds for any $\Omega \subset M$. □

Finally, we give an example with Assumption (H2) holding but $a(x) = \frac{1}{2}\langle DV, \mathcal{E}\rangle_{T_x^2} \neq 0$.

Example 1.3. Consider a helicoid, defined by

$$M = \{\alpha(t, s) \,|\, (t, s) \in \mathbb{R}^2,\ t > 0\}, \tag{1.74}$$

where

$$\alpha(t, s) = (t\cos s, t\sin s, c_0 s), \quad c_0 > 0. \tag{1.75}$$

The Gauss curvature is $-c_0^2/(t^2 + c_0^2)^2$.

We set

$$E_1 = \alpha_t = (\cos s, \sin s, 0); \tag{1.76}$$

$$E_2 = \frac{1}{\sqrt{t^2 + c_0^2}}\alpha_s = \frac{1}{\sqrt{t^2 + c_0^2}}(-t\sin s, t\cos s, c_0). \tag{1.77}$$

Then E_1, E_2 makes up a frame field on the whole surface M. We may obtain

$$D_{E_1}E_1 = 0; \quad D_{E_2}E_1 = \frac{t}{t^2 + c_0^2}E_2; \tag{1.78}$$

$$D_{E_1}E_2 = 0; \quad D_{E_2}E_2 = \frac{-t}{t^2 + c_0^2}E_1, \tag{1.79}$$

where D is the Levi–Civita connection of surface M.

For any $c > 0$, we set

$$V_c = f_c E_1 + hE_2, \tag{1.80}$$

where

$$f_c = \sqrt{t^2 + c_0^2}\left(\int_0^t \frac{dt}{\sqrt{t^2 + c_0^2}} + c\right); \quad h = \sqrt{t^2 + c_0^2}s. \tag{1.81}$$

We then have

$$DV_c = b_c g + a\mathcal{E} \quad \text{for} \quad c > 0, \tag{1.82}$$

where g is the induced metric of M in $\mathbb{R}^3$, $\mathcal{E}$ is the volume element of M, and

$$b_c = 1 + \frac{t}{t^2 + c_0^2} f_c; \quad a = -\frac{st}{\sqrt{t^2 + c_0^2}}. \tag{1.83}$$

It is clear that, for any $\Omega \subset M$ bounded with $\overline{\Omega} \subset M$ and for any constant $c_1 > 0$, there is $c > 0$ large enough such that

$$\min_{x\in\Omega} b_c \geq c_1 \max_{x\in\Omega} |a|; \tag{1.84}$$

that is, for any $\Omega \subset M$, we can find a vector field V_c ($c > 0$ large enough), defined by (1.80), to meet geometric conditions (1.40) and (1.42). □

Remark 1.7. It is easy to check by the curvature information and Yao [37] that Examples 1.1–1.3 satisfy Assumption (H1), too.

2. Geometric conditions; proofs of main results.

2.1. On geometric conditions. We give some information on how to find a vector field to meet conditions (1.40) and (1.42).

PROPOSITION 2.1. *Let vector field V be given such that* (1.40) *holds. If $T \in T^2(M)$ is a symmetric,* 2*-order tensor field, then*

$$\langle T, T(\cdot, D.V)\rangle_{T_x^2} = b|T|^2_{T_x^2} \quad \text{at } x \in M; \tag{2.1}$$

$$\mathrm{tr}T\mathrm{tr}T(\cdot, D.V) = b(\mathrm{tr}T)^2 \quad \text{at } x \in M, \tag{2.2}$$

where "·" denotes the position of the variable.

Proof. Let $T \in T^2(M)$ be symmetric. Given $x \in M$, there is an orthonormal basis e_1, e_2 of M_x such that

$$T(e_1, e_2) = 0, \quad \text{at } x, \tag{2.3}$$

since $T(\cdot, \cdot)$ is a symmetric, bilinear form on $M_x \times M_x$. Conditions (1.40) and (2.3) yield

$$\begin{aligned} T(e_i, D_{e_i}V) &= T(e_i,\ DV(e_i, e_i)e_i + DV(e_j, e_i)e_j) \\ &= bT(e_i, e_i) + DV(e_j, e_i)T(e_i, e_j) = bT(e_i, e_i) \end{aligned} \tag{2.4}$$

for $j \neq i,\ i = 1, 2$. By (2.3) and (2.4), we obtain

$$\langle T, T(\cdot, D.V)\rangle_{T_x^2} = \sum_{ij=1}^{2} T(e_i, e_j)T(e_i, D_{e_j}V) = \sum_{i=1}^{2} T(e_i, e_i)T(e_i, D_{e_i}V)$$
$$= b\sum_{i=1}^{2}(T(e_i, e_i))^2 = b\sum_{ij=1}^{2}(T(e_i, e_j))^2 = b|T|^2_{T_x^2} \quad \text{at } x$$

and

$$\mathrm{tr}T(\cdot, D.V) = \sum_{i=1}^{2} T(e_i, D_{e_i}V) = b\sum_{i=1}^{2} T(e_i, e_i) = b\mathrm{tr}T \quad \text{at } x. \qquad \square$$

Consider a set Θ which consists of all the C^∞ functions q such that there is a region $\aleph \subset M$ satisfying $\overline{\Omega} \subset \aleph$ and

$$\Delta q = k(x) \quad \forall\ x \in \aleph, \tag{2.5}$$

where k is the Gaussian curvature function of M. It is easily checked that Θ is nonempty by an elliptic boundary value problem for the Laplacian since $\Omega \subset M$ is a bounded region with a nonempty boundary Γ. For a given $q \in \Theta$, consider a metric on $\aleph$ given by

$$\hat{g} = e^{2q}g, \tag{2.6}$$

where g is the induced metric of M in $\mathbb{R}^3$. Denote by $(\aleph, \hat{g})$ the Riemannian manifold with metric (2.6) and by $\hat{\rho}(x)$ the distance function on $(\aleph, \hat{g})$ from $x^0 \in \aleph$ to $x \in \aleph$, respectively. Let $\hat{D}$ be the covariant differential of $(\aleph, \hat{g})$ in metric $\hat{g}$.

PROPOSITION 2.2. *Given $q \in \Theta$ and $x^0 \in \aleph$, set*

$$V(x) = \hat{\rho}(x)\hat{D}\hat{\rho}(x), \quad x \in \aleph. \tag{2.7}$$

Then vector field V on $\aleph$ (so that on Ω) meets condition (1.40) *where $b = 1 - V(q)$ and $a = \langle Dq \otimes V, \mathcal{E}\rangle_{T_x^2}$ such that*

$$\lim_{x\to x_0} b(x) = 1 \quad \textit{and} \quad \lim_{x\to x_0} a(x) = 0, \tag{2.8}$$

that is, conditions (1.40) *and* (1.42) *hold locally.*

Proof. We do a computation in a coordinate. Let (x_1, x_2) be a coordinate on $\aleph$. Set

$$g_{ij} = \left\langle \frac{\partial}{\partial x_i}, \frac{\partial}{\partial x_j} \right\rangle \quad \text{and} \quad \hat{g}_{ij} = \hat{g}\left(\frac{\partial}{\partial x_i}, \frac{\partial}{\partial x_j}\right) \tag{2.9}$$

for $1 \leq i, j \leq 2$. It follows from (2.6) and (2.9) that $\hat{g}_{ij} = e^{2q}g_{ij}$.

Denote the coefficients of connections D and $\hat{D}$ by Γ_{ij}^k and by $\hat{\Gamma}_{ij}^k$, respectively. From Schoen and Yau [26, Chapter 5], we have formulae

$$\Gamma_{ij}^k = \hat{\Gamma}_{ij}^k - \delta_{ik}\frac{\partial q}{\partial x_j} - \delta_{jk}\frac{\partial q}{\partial x_i} + g_{ij}\sum_{l=1}^{2} g^{kl}\frac{\partial q}{\partial x_l} \tag{2.10}$$

and

$$\Delta q - k + \hat{k}e^{2q} = 0 \quad \text{in} \quad \aleph, \tag{2.11}$$

where $\hat{k}$ is the Gaussian curvature function of $(\aleph, \hat{g})$ in metric $\hat{g}$ and $(g^{ij}) = (g_{ij})^{-1}$. Relations (2.5) and (2.11) yield

$$\hat{k} = 0 \quad \text{in} \quad \aleph, \tag{2.12}$$

i.e., $(\aleph, \hat{g})$ is of zero curvature. It is well known that

$$\hat{D}V = \frac{1}{2}\hat{D}^2\hat{\rho}^2 = \hat{g} \quad \text{in} \quad \aleph \tag{2.13}$$

because of (2.12).

Let $V = h_1\dfrac{\partial}{\partial x_1} + h_2\dfrac{\partial}{\partial x_2}$. For any vector field $X = X_1\dfrac{\partial}{\partial x_1} + X_2\dfrac{\partial}{\partial x_2}$, it follows from (2.10) that

$$\begin{aligned} D_X V &= \sum_{j=1}^{2} X(h_j)\frac{\partial}{\partial x_j} + \sum_{ij=1}^{2} X_i h_j D_{\frac{\partial}{\partial x_i}}\frac{\partial}{\partial x_j} = \sum_{j=1}^{2} X(h_j)\frac{\partial}{\partial x_j} + \sum_{k=1}^{2}\sum_{ij=1}^{2} X_i h_j \Gamma_{ij}^{k}\frac{\partial}{\partial x_k} \\ &= \hat{D}_X V - V(q)X - X(q)V + \langle X, V\rangle Dq. \end{aligned} \tag{2.14}$$

By (2.14), (2.6), and (2.13), we obtain

$$\begin{aligned} DV(X,X) &= \langle \hat{D}_X V, X\rangle - V(q)|X|^2 = e^{-2q}\hat{D}V(X,X) - V(q)|X|^2 \\ &= e^{-2q}\hat{g}(X,X) - V(q)|X|^2 = (1 - V(q))|X|^2 \end{aligned} \tag{2.15}$$

$\forall\ X \in M_x,\ x \in \aleph$, and

$$\begin{aligned} a(x) &= \frac{1}{2}\langle DV, \mathcal{E}\rangle = \frac{1}{2}[DV(e_1,e_2) - DV(e_2,e_1)] \\ &= \langle e_1, V\rangle e_2(q) - \langle e_2, V\rangle e_1(q) = \langle Dq \otimes V, \mathcal{E}\rangle_{T_x^2}, \end{aligned} \tag{2.16}$$

where e_1, e_2 is an orthonormal basis of M_x with the positive orientation. Since $e^q|\hat{D}\hat{\rho}| = 1$ we then obtain $|V| = \hat{\rho}(x)e^{-q}$ so that (2.8) follows from (2.16) and $b = 1 - V(q)$. □

Given $x^0 \in M$. Let r: $(0,\infty) \to \mathbb{R}^n$ be a geodesic with $r(0) = x^0$ parameterized by arc length in Riemannian metric g. Denote the distance on (M, g) by dis. For sufficiently small $t > 0$ we know that $dis(r(t), x^0) = t$, since the exponential map $\exp_{x^0}$: $M_{x^0} \to M$ is injective on a sufficiently small ball in (M, g). We recall that $r(t_0)$ is called the cut point of r with respect to x^0, if $t_0 > 0$ is such that $d_g(r(t), x^0) = t$, $0 \le t < t_0$, and $d_g(r(t), x^0) < t\ \forall\ t > t_0$; see Cheeger and Ebin [6]. The union of all cut points is called the cut locus of x^0 and denoted by $cut(x^0)$. For any $X \in M_{x^0}$, $|X| = 1$, there is at most one cut point on the geodesic $\exp_{x^0} tX$ $(t \ge 0)$. Thus $cut(x^0)$ is the image of the exponential map on some closed subset of S^{n-1} and the n-dimensional measure of $cut(x^0)$ is zero. Set $\mu(X) = d_g(x^0, r(t_0))$, if $r(t_0)$ is the cut point of x^0 along $r(t) = \exp_{x^0} tX$; $\mu(X) = \infty$, when there is no cut point of x^0 along r, where $X \in S^{n-1} \subset M_{x^0}$. We define

$$E(x_0) = \left\{ tX \,|\, 0 \le t < \mu(X),\ X \in S^{n-1} \subset M_{x^0} \right\}.$$

Then $\exp_{x^0}\colon E(x_0) \to \exp_{x^0} E(x_0)$ is a diffeomorphism. It is obvious that $\exp_{x^0} E(x_0)$ is a star domain and

$$M = \exp_{x^0}(E) \cup cut(x^0).$$

PROPOSITION 2.3. *Let M be of constant curvature k. Given $x_0 \in M$, let ρ be the distance function from $x \in M$ to x_0 on (M, g), i.e., $\rho(x) = dis(x_0, x)$. Set $V = h(\rho)D\rho$, where $h(\rho)$ is defined by*

$$h(\rho) = \begin{cases} \sin(\sqrt{k}\rho(x)), & k > 0, \\ \rho(x), & k = 0, \\ \sinh(\sqrt{-k}\rho(x)), & k < 0, \end{cases} \tag{2.17}$$

for $x \in \exp_{x_0} E(x_0)$. Then

$$DV = b(x)g \quad \text{and} \quad a(x) = 0, \quad x \in \exp_{x_0} E(x_0), \tag{2.18}$$

where

$$b(x) = \begin{cases} \sqrt{k}\cos(\sqrt{k}\rho(x)), & k > 0, \\ 1, & k = 0, \\ \sqrt{-k}\cosh(\sqrt{-k}\rho(x)), & k < 0. \end{cases} \tag{2.19}$$

Proof. It is well known that

$$D^2\rho(\tau, \tau) = \begin{cases} \sqrt{k}\cot(\sqrt{k}\rho(x)), & k > 0, \\ \dfrac{1}{\rho(x)}, & k = 0, \\ \sqrt{-k}\coth(\sqrt{-k}\rho(x)), & k < 0, \end{cases} \qquad x \in \exp_{x_0} E(x_0), \tag{2.20}$$

where $\tau \in M_x$ such that $D\rho$, τ is an orthonormal basis of M_x and $D^2\rho$ is the Hessian of distance function ρ at x. Since $D^2\rho(D\rho, X) = 0$ for any $X \in M_x$, it follows from (2.20) that

$$DV(X, Y) = h'(\rho)\langle D\rho, X\rangle\langle D\rho, Y\rangle + h(\rho)D^2\rho(X, Y) = b(x)\langle X, Y\rangle, \tag{2.21}$$

for any $X, Y \in M_x$. Since DV is symmetric, $a(x) = 0$. ∎

PROPOSITION 2.4. *Let f be a function on the zy-plane, $z = f(y)$ for $y > 0$. Consider a surface, given by*

$$M = \left\{(x, y, z) \,|\, z = f(r), r = \sqrt{x^2 + y^2}, (x, y) \in \mathbb{R}^2\right\}. \tag{2.22}$$

The curvature is

$$k(p) = \frac{f'(r)f''(r)}{r(1 + f'^2)^2}, \quad p = (x, y, z) \in M. \tag{2.23}$$

Then there exist a vector field V and a function b such that

$$DV = b(p)g \quad \text{and} \quad a(p) = 0, \quad p \in M, \tag{2.24}$$

where g is the induced metric of M in $\mathbb{R}^3$. Furthermore, if

$$\int_{k(t)>0} tk(t)dt \le 1, \tag{2.25}$$

then

$$b(p) > 0 \quad \forall p \in M, \tag{2.26}$$

where

$$k(t) = \frac{f'(r(t))f''(r(t))}{r(t)(1+f'^2(r(t)))^2}, \quad t > 0, \tag{2.27}$$

and $r(t)$ is defined by equation

$$t = \int_0^{r(t)} \sqrt{1+f'^2(s)}ds, \tag{2.28}$$

for $t \geq 0$.

Proof. Denote $p_0 = (0, 0, f(0))$. Let $\rho(p) = dis(p_0, p)$ be the distance function from $p \in M$ to p_0 on (M, g). Set $V = \varphi(\rho)D\rho$, where φ is the solution to the problem

$$\begin{cases} \varphi''(t) + k(t)\varphi(t) = 0, \quad t > 0, \\ \varphi(0) = 0, \quad \varphi'(0) = 1. \end{cases} \tag{2.29}$$

We will show that $DV = \varphi'(\rho(p))g$ for any $p \in M$.

Let $p = (x, y, f(r)) \in M$ with $r = \sqrt{x^2+y^2}$. Since M is of revolution, it is easily checked that curve σ: $[0, \rho(p)] \to M$, given by

$$\sigma(t) = \left(r(t)\frac{x}{r},\, r(t)\frac{y}{r},\, f(r(t))\right),$$

is the unique minimizing geodesic, parameterized by arc length, which joins p_0 to p. Since $k(\sigma(t)) = k(t)$ for $0 \leq t \leq \rho(p)$, it is well known that

$$D^2\rho(\tau, \tau) = \frac{\varphi'(\rho)}{\varphi(\rho)}, \tag{2.30}$$

where $\tau \in M_p$ such that $D\rho$, τ is an orthonormal basis of M_p. It follows from (2.30) that

$$\begin{aligned} DV(X,Y) &= \langle D_Y V, X\rangle = \varphi'(\rho)\langle D\rho, X\rangle\langle D\rho, Y\rangle + \varphi(\rho)D^2\rho(X,Y) \\ &= \varphi'(\rho)\langle D\rho, X\rangle\langle D\rho, Y\rangle + \varphi'(\rho)\langle X, \tau\rangle\langle Y, \tau\rangle = \varphi'(\rho)\langle X, Y\rangle, \end{aligned} \tag{2.31}$$

for any $X, Y \in M_p$.

Let inequality (2.25) hold. We next prove $\varphi'(t) > 0$ for $t \geq 0$.

Set $\widetilde{k}(t) = \max(k, 0)$. Let ϕ be the solution to the problem

$$\begin{cases} \phi''(t) + \widetilde{k}(t)\phi = 0, \quad t > 0, \\ \phi(0) = 0, \quad \phi'(0) = 1. \end{cases} \tag{2.32}$$

Inequality (2.25) means that

$$\int_0^\infty t\widetilde{k}(t)\, dt \leq 1. \tag{2.33}$$

By Proposition 4.2 of Green and Wu [13], it follows from (2.32) and (2.33) that

$$\phi > 0 \quad \text{and} \quad \phi' > 0 \quad \text{on} \quad (0, \infty). \tag{2.34}$$

In addition, Sturm's theorem gives

$$\varphi \geq \phi > 0 \quad \text{on} \quad (0,\infty) \tag{2.35}$$

since $\widetilde{k}(t) \geq k(t)$ for $t \geq 0$. To complete our proof, by (2.34) it will suffice to prove $\varphi' \geq \phi'$ on $(0,\infty)$. Set $u(t) = \varphi'\phi - \varphi\phi'$. By (2.29), (2.32), and (2.35), we obtain

$$u'(t) = \varphi''\phi - \varphi\phi'' = (\widetilde{k} - k)\varphi\phi \geq 0 \tag{2.36}$$

on $(0,\infty)$. It follows from (2.36) and $u(0) = 0$ that $u(t) \geq 0$ on $(0,\infty)$. We then have, from (2.34) and (2.35),

$$\left(\frac{\varphi'}{\phi'}\right)' = \frac{1}{\phi'^2}[(\widetilde{k} - k)\varphi\phi' + \widetilde{k}u(t)] \geq 0 \tag{2.37}$$

on $(0,\infty)$. Inequality (2.37) gives $\varphi' \geq \phi'$ on $(0,\infty)$ since $\varphi'(0) = \phi'(0) = 1$.

2.2. Some multiplier identities.. To estimate inequality (1.49), we need some multiplier identities which have been built for the classical wave equations; see Lions [24], Lasiecka and Triggiani [22], and Chen [7]; for the Kirchhoff plates, see Lagnese and Lions [21]; and for variably coefficient wave equations and Euler–Bernoulli equations see Yao [36], [38]. We now consider their generalizations for the shallow shell.

In the following, let vector field V be such that condition (1.40) holds. Denote by $lot(\eta)$ the lower order term with respect to the energy level, that is, for any $\epsilon > 0$ there is $c_\epsilon > 0$ such that

$$|lot(\eta)| \leq \epsilon \int_0^T E(t)dt + c_\epsilon \left[L(0) + L(T) + \int_0^T L(t)dt \right], \tag{2.38}$$

where $E(t)$ and $L(t)$ are defined in (1.45) and (1.47), respectively.

PROPOSITION 2.5. *Let $\eta = (W, w)$ solve problem* (1.48) *such that all the terms in* (2.39) *and* (2.40) *below are well defined. Let f be a function on Ω. Then*

$$\int_Q f[|W_t|^2 - B(\eta,\eta_1)]\, dQ = -\int_\Sigma \partial(\mathcal{A}\eta, f\eta_1)\, d\Sigma + lot(\eta); \tag{2.39}$$

$$\int_Q f[w_t^2 + \gamma|Dw_t|^2 - B(\eta,\eta_2)]\, dQ = -\int_\Sigma \left[\partial(\mathcal{A}\eta, f\eta_2) + \gamma f w \frac{\partial w_{tt}}{\partial n}\right] d\Sigma + lot(\eta), \tag{2.40}$$

where η_1, η_2 are given by (1.46).

Proof. We multiply (1.48) by $f\eta_2$. Then

$$\begin{aligned}(w_t, fw)_{L^2(\Omega)}|_0^T - \int_Q f w_t^2 dQ - \gamma \int_0^T (\Delta w_{tt}, fw)_{L^2(\Omega)} dt \\ + \int_0^T (\mathcal{A}\eta, f\eta_2)_{L^2(\Omega,\Lambda)\times L^2(\Omega)} dt = 0.\end{aligned} \tag{2.41}$$

The third term equals

$$\begin{aligned}
&-\gamma(\Delta w_t, fw)_{L^2(\Omega)}|_0^T + \gamma\int_0^T (\Delta w_t, fw_t)_{L^2(\Omega)}dt\\
&\qquad = \gamma(Dw_t, D(fw))_{L^2(\Omega,\Lambda)}|_0^T - \gamma\int_Q \langle Dw_t, D(fw_t)\rangle dQ\\
&\qquad\quad -\gamma\int_\Gamma fw\frac{\partial w_t}{\partial n}d\Gamma|_0^T + \gamma\int_\Sigma fw_t\frac{\partial w_t}{\partial n}d\Sigma\\
&\qquad = -\gamma\int_Q f|Dw_t|^2dQ - \gamma\int_\Sigma fw\frac{\partial w_{tt}}{\partial n}d\Sigma + lot(\eta). \qquad (2.42)
\end{aligned}$$

The fourth term equals (using formula (1.29))

$$\int_Q fB(\eta,\eta_2)dQ - \int_\Sigma \partial(\mathcal{A}\eta, f\eta_2)d\Sigma + lot(\eta), \qquad (2.43)$$

since we have

$$D^2(fw) = fD^2w + Df\otimes Dw + Dw\otimes Df + wD^2f, \qquad (2.44)$$

where " $\otimes$ " is the tensor product, and

$$\int_Q B(\eta, f\eta_2)dQ = \gamma\int_Q a(\rho(\eta),\rho(f\eta_2))\,dQ = \int_Q fB(\eta,\eta_2)\,dQ + lot(\eta). \qquad (2.45)$$

Inserting (2.42) and (2.43) into (2.41), we obtain identity (2.40).

We multiply (1.48) by $f\eta_1$ and obtain identity (2.39) through a similar computation. □

LEMMA 2.6. *For any W, $V\in\mathcal{X}(M)$, we have*

$$D(D_VW) = D_V(DW) + R(V,\cdot,W,\cdot) + DW(\cdot, D_\cdot V), \qquad (2.46)$$

where $R(\cdot,\cdot,\cdot,\cdot)$ is the tensor of curvature, given in (1.14)*, and "$\cdot$" denotes the position of variable.*

Proof. Given $x\in M$, let E_1, E_2 be a frame field normal at x. We then have, from (1.14) and (1.13),

$$\begin{aligned}
&D(D_VW)(E_i,E_j) = E_j(D_VW(E_i)) = E_j(DW(E_i,V))\\
&= D^2W(E_i,V,E_j) + DW(E_i, D_{E_j}V)\\
&= D^2W(E_i,E_j,V) + R_{VE_j}W(E_i) + DW(E_i, D_{E_j}V)\\
&= D_V(DW)(E_i,E_j) + R(V,E_j,W,E_i) + DW(E_i, D_{E_j}V), \quad \text{at } x, \qquad (2.47)
\end{aligned}$$

since $(D_{E_j}E_i)(x) = 0$ and $(D_VE_i)(x) = \sum_{k=1}^2\langle V,E_k\rangle(D_{E_k}E_i)(x) = 0$ for $1\le i, j\le 2$.

Identity (2.46) follows from (2.47). □

LEMMA 2.7. *Let $V\in\mathcal{X}(\Omega)$ be given. Set*

$$G(V,\eta)(X,Y) = \frac{1}{2}[DW(X,D_YV) + DW(Y,D_XV)], \quad X, Y\in M_x,\ x\in M. \qquad (2.48)$$

Then

$$\Upsilon(m(\eta)) = D_V\Upsilon(\eta) + G(V,\eta) + l(\eta); \qquad (2.49)$$

$$\rho(m(\eta)) = D_V\rho(\eta) + \rho(\eta)(\cdot, D.V) + \rho(\eta)(D.V, \cdot) + l(\eta), \tag{2.50}$$

where $l(\eta)$ *denote all the terms such that* $\int_0^T |l(\eta)|dt = lot(\eta)$.

Proof. First, by (1.22) and (2.46), we have

$$\begin{aligned} D_V\Upsilon(\eta) &= \frac{1}{2}[D_V(DW) + D_V(D^*W)] + V(w)\Pi + wD_V\Pi \\ &= \frac{1}{2}[D(D_VW) + (D(D_VW))^* - R(V,\cdot,W,\cdot) - R(W,\cdot,V,\cdot)] \\ &\quad -G(V,\eta) + V(w)\Pi + wD_V\Pi = \Upsilon(m(\eta)) - G(V,\eta) - l(\eta), \end{aligned} \tag{2.51}$$

that is, identity (2.49), where

$$l(\eta) = \frac{1}{2}[R(V,\cdot,W,\cdot) + R(W,\cdot,V,\cdot)] - wD_V\Pi. \tag{2.52}$$

Given $x \in \Omega$, let E_1, E_2 be a frame field normal at x. By (1.13), we then have

$$\begin{aligned} D^2(V(w))(E_i,E_j) &= E_jE_i(Dw(V)) = E_j(D^2w(E_i,V) + Dw(D_{E_i}V)) \\ &= D^3w(E_i,W,E_j) + D^2w(E_i,D_{E_j}V) + E_j\langle Dw, D_{E_i}V\rangle \\ &= D_V(D^2w)(E_i,E_j) + R(Dw,E_i,W,E_j) \\ &\quad +D^2w(E_i,D_{E_j}V) + E_j\langle Dw, D_{E_i}V\rangle \quad \text{at } x, \end{aligned} \tag{2.53}$$

since $(D_{E_i}E_j)(x) = 0$ for $1 \le i, j \le 2$. The last term in (2.53) can be computed as follows:

$$\begin{aligned} E_j(DV(Dw,E_i)) &= D^2V(Dw,E_i,E_j) + DV(D_{E_j}Dw,E_i) \\ &= l_{Dw}D^2V(E_i,E_j) + D^2w(D_{E_i}V,E_j), \quad \text{at } x. \end{aligned} \tag{2.54}$$

Inserting (2.54) into (2.53) yields

$$D^2(V(w)) = D_V(D^2w) + D^2w(\cdot, D.V) + D^2w(D.V,\cdot) - l(\eta), \tag{2.55}$$

where

$$l(\eta) = -R(Dw,\cdot,V,\cdot) - l_{Dw}D^2V, \tag{2.56}$$

that is, identity (2.50) since $\rho(\eta) = -D^2w$. □

LEMMA 2.8. *Let* $V \in \mathcal{X}(\Omega)$ *be such that* (1.40) *holds. Then*

$$\begin{aligned} \mathcal{B}(\eta, m(\eta)) &= \frac{1}{2}\int_\Gamma B(\eta,\eta)\langle V,n\rangle\, d\Gamma + \int_\Omega a(\Upsilon(\eta), G(V,\eta))\, dx \\ &\quad + \int_\Omega b[B(\eta,\eta_2) - B(\eta,\eta_1)]\, dx + l(\eta). \end{aligned} \tag{2.57}$$

Proof. First, we compute $\int_\Omega a(\Upsilon(\eta), \Upsilon(m(\eta))dx$.

By (2.49), we obtain

$$\langle \Upsilon(\eta), \Upsilon(m(\eta))\rangle_{T_x^2} = \frac{1}{2}V(|\Upsilon(\eta)|^2_{T_x^2}) + \langle \Upsilon(\eta), G(V,\eta)\rangle_{T_x^2} + \langle \Upsilon(\eta), l(\eta)\rangle_{T_x^2} \tag{2.58}$$

for $x \in \Omega$. It follows from (2.58) and by divergence formula that

$$\begin{aligned}\int_\Omega \langle \Upsilon(\eta), \Upsilon(m(\eta))\rangle_{T_x^2}\, dx =& \frac{1}{2}\int_\Gamma |\Upsilon(\eta)|^2_{T_x^2}\langle V, n\rangle\, d\Gamma - \int_\Omega b|\Upsilon(\eta)|^2_{T_x^2}\, dx \\ &+ \int_\Omega \langle \Upsilon(\eta), G(V,\eta)\rangle_{T_x^2}\, dx + l(\eta),\end{aligned} \tag{2.59}$$

since $\mathrm{div} V = 2b$.

Given $x \in \Omega$. Let E_1, E_2 be a frame field normal at x. We then have

$$V(\mathrm{tr}\Upsilon(\eta)) = \sum_{i=1}^{2} D_V \Upsilon(\eta)(E_i, E_i) = \mathrm{tr} D_V \Upsilon(\eta) \quad \text{at } x, \tag{2.60}$$

since $(D_V E_i)(x) = 0$ for $i = 1, 2$. From (2.49) and (2.60), an argument similar to (2.59) gives

$$\begin{aligned}\int_\Omega \mathrm{tr}\Upsilon(\eta)\mathrm{tr}\Upsilon(m(\eta))\, dx =& \frac{1}{2}\int_\Gamma (\mathrm{tr}\Upsilon(\eta))^2\langle V, n\rangle\, d\Gamma - \int_\Omega b(\mathrm{tr}\Upsilon(\eta))^2\, dx \\ &+ \int_\Omega \mathrm{tr}\Upsilon(\eta)\mathrm{tr}G(V,\eta)\, dx + l(\eta).\end{aligned} \tag{2.61}$$

Using (2.59), (2.61), and (1.26), we obtain the following identity:

$$\begin{aligned}\int_\Omega a(\Upsilon(\eta), \Upsilon(m(\eta)))\, dx =& \frac{1}{2}\int_\Gamma a(\Upsilon(\eta), \Upsilon(\eta))\langle V, n\rangle\, d\Gamma - \int_\Omega bB(\eta,\eta_1)\, dx \\ &+ \int_\Omega a(\Upsilon(\eta), G(V,\eta))\, dx + l(\eta),\end{aligned} \tag{2.62}$$

since $B(\eta, \eta_1) = a(\Upsilon(\eta), \Upsilon(\eta)) + l(\eta)$.

We now compute $\int_\Omega \gamma a(\rho(\eta), \rho(m(\eta)))dx$.

Since $\rho(\eta)$ is symmetric, we have by properties (2.1), (2.2), and identity (2.50)

$$\langle \rho(\eta), \rho(m(\eta))\rangle_{T_x^2} = \frac{1}{2}V(|\rho(\eta)|^2_{T_x^2}) + 2b|\rho(\eta)|^2 + l(\eta); \tag{2.63}$$

$$\mathrm{tr}\rho(\eta)\mathrm{tr}\rho(m(\eta)) = \frac{1}{2}V((\mathrm{tr}\rho(\eta))^2) + 2b(\mathrm{tr}\rho(\eta))^2 + l(\eta). \tag{2.64}$$

Combining (2.63) with (2.64) we obtain the following identity by divergence formula:

$$\begin{aligned}\int_\Omega \gamma a(\rho(\eta), \rho(m(\eta)))\, dx =& \frac{1}{2}\int_\Gamma \gamma a(\rho(\eta), \rho(\eta))\langle V, n\rangle\, d\Gamma \\ &+ \int_\Omega bB(\eta,\eta_2)\, dx + l(\eta).\end{aligned} \tag{2.65}$$

Adding up (2.62) and (2.65), we obtain identity (2.57). □

PROPOSITION 2.9. *Let $\eta = (W, w)$ solve problem (1.48). Then*

$$\begin{aligned}&\frac{1}{2}\int_\Sigma [|\eta_t|^2 + \gamma|Dw_t|^2 - B(\eta,\eta)]\langle V, n\rangle\, d\Sigma \\ &+ \int_\Sigma \left[\partial(\mathcal{A}\eta, m(\eta)) + \gamma V(w)\frac{\partial w_{tt}}{\partial n}\right] d\Sigma \\ &= Z|_0^T + \int_Q b[|\eta_t|^2 + B(\eta,\eta_2) - B(\eta,\eta_1)]\, dQ \\ &+ \int_Q a(\Upsilon(\eta), G(V,\eta))\, dQ + lot(\eta),\end{aligned} \tag{2.66}$$

where

$$Z = (\eta_t, m(\eta))_{L^2(\Omega,\Lambda)\times L^2(\Omega)} + \gamma(Dw_t, D(V(w)))_{L^2(\Omega)}. \tag{2.67}$$

Proof. We multiply (1.48) by $m(\eta)$. Then

$$\begin{aligned}&(\eta_t, m(\eta))_{L^2(\Omega,\Lambda)\times L^2(\Omega)}|_0^T - \int_0^T (\eta_t, m(\eta_t))_{L^2(\Omega,\Lambda)\times L^2(\Omega)}\, dt\\ &-\gamma\int_0^T (\Delta w_{tt}, V(w))_{L^2(\Omega)}dt + \int_0^T (\mathcal{A}\eta, m(\eta))_{L^2(\Omega,\Lambda)\times L^2(\Omega)}\, dt = 0.\end{aligned} \tag{2.68}$$

Let us compute each term in (2.68) separately.

By the divergence formula

$$\text{second term} = -\frac{1}{2}\int_Q V(|\eta_t|^2)dQ = -\frac{1}{2}\int_\Sigma |\eta_t|^2\langle V, n\rangle d\Sigma + \int_Q b|\eta_t|^2 dQ. \tag{2.69}$$

By Yao [36, Lemma 2.1] and (1.40), we have

$$\begin{aligned}\langle Dw_t, D(V(w_t))\rangle &= DV(Dw_t, Dw_t) + \frac{1}{2}\mathrm{div}(|Dw_t|^2V) - \frac{1}{2}|Dw_t|^2\mathrm{div}V\\ &= \frac{1}{2}\mathrm{div}(|Dw_t|^2V)\end{aligned} \tag{2.70}$$

since $\mathrm{div}V = 2b$. It follows from (2.70) and Green's formula that

$$-(\Delta w_t, V(w_t))_{L^2(\Omega)} = \frac{1}{2}\int_\Gamma |Dw_t|^2\langle V, n\rangle\, d\Gamma - \int_\Gamma V(w_t)\frac{\partial w_t}{\partial n}\, d\Gamma. \tag{2.71}$$

By (2.71), we obtain

$$\begin{aligned}\text{third term} &= -\gamma(\Delta w_t, V(w))_{L^2(\Omega)}|_0^T + \gamma\int_0^T (\Delta w_t, V(w_t))_{L^2(\Omega)}dt\\ &= \gamma(Dw_t, D(V(w)))_{L^2(\Omega,\Lambda)}|_0^T - \frac{\gamma}{2}\int_\Sigma |Dw_t|^2\langle V, n\rangle d\Sigma\\ &\quad -\gamma\int_\Sigma V(w)\frac{\partial w_{tt}}{\partial n}\, d\Sigma.\end{aligned} \tag{2.72}$$

By formulae (1.29) and (2.57), we have

$$\begin{aligned}\text{third term} &= \int_0^T \mathcal{B}(\eta, m(\eta))\, dt - \int_\Sigma \partial(\mathcal{A}\eta, m(\eta))\, d\Sigma\\ &= \frac{1}{2}\int_\Sigma B(\eta,\eta)\langle V, n\rangle\, d\Sigma - \int_\Sigma \partial(\mathcal{A}\eta, m(\eta))\, d\Sigma + lot(\eta)\\ &\quad + \int_Q b[B(\eta,\eta_2) - B(\eta,\eta_1)]\, dQ + \int_Q a(\Upsilon(\eta), G(V,\eta))\, dQ.\end{aligned} \tag{2.73}$$

Adding up all the terms, we obtain identity (2.66). ∎

LEMMA 2.10. *Let Z be given in* (2.67). *Then, for any $\epsilon > 0$, there is $c_\epsilon > 0$, independent of η, such that*

$$|Z(t)| \le (\lambda_0\sigma_0 + \epsilon)E(t) + c_\epsilon L(t), \tag{2.74}$$

where λ_0 and σ_0 are defined in (1.39) *and* (1.43)*, respectively.*

Proof. Given $\epsilon > 0$, it is easily checked that there is $c_\epsilon > 0$ such that

$$|(w_t, V(w))_{L^2(\Omega)} + \gamma(Dw, D_{Dw_t}V)_{L^2(\Omega,\Lambda)}| \le \epsilon E(t) + c_\epsilon L(t). \tag{2.75}$$

Therefore by formula $D(V(w)) = Dw(D.V) + l_V D^2 w$ and (2.75), we obtain the following estimate:

$$\begin{aligned} |Z| &= |(W_t, D_V W)_{L^2(\Omega,\Lambda)} + \gamma(Dw_t, l_V D^2 w)_{L^2(\Omega,\Lambda)} \\ &\quad + (w_t, V(w))_{L^2(\Omega)} + \gamma(Dw, D_{Dw_t}V)_{L^2(\Omega,\Lambda)}| \\ &\le \sigma_0[\|W_t\|_{L^2(\Omega,\Lambda)}\|DW\|_{L^2(\Omega,T^2)} + \gamma\|Dw_t\|_{L^2(\Omega,\Lambda)}\|D^2 w\|_{L^2(\Omega,T^2)}] \\ &\quad + \epsilon E(t) + c_\epsilon L(t) \\ &\le (\sigma_0\lambda_0 + \epsilon)E(t) + c_\epsilon L(t). \qquad \square \end{aligned} \tag{2.76}$$

LEMMA 2.11. *Let $V \in \mathcal{X}(\Omega)$ be such that conditions* (1.40) *and* (1.42) *hold. Then*

$$\int_\Omega a(\Upsilon(\eta_1), G(V, \eta_1))\, dx \ge \sigma_1 \int_\Omega B(\eta_1, \eta_1)\, dx. \tag{2.77}$$

Proof. It is easily checked that condition (1.40) implies

$$DV = bg + a\mathcal{E}, \quad x \in \Omega, \tag{2.78}$$

where g is the induced metric of M in $\mathbb{R}^3$ and $\mathcal{E}$ the volume element of M.

Given $x \in \Omega$, since $\Upsilon(\eta_1)$ is symmetric, we may take e_1, e_2, as an orthonormal basis of M_x, with the positive orientation such that

$$DW(e_1, e_2) + DW(e_2, e_1) = 0 \quad \text{at } x. \tag{2.79}$$

Formula (2.78) implies

$$D_{e_1}V = be_1 - ae_2, \quad D_{e_2}V = ae_1 + be_2 \quad \text{at } x. \tag{2.80}$$

Denote $W_{ij} = DW(e_i, e_j)$ for $1 \le i, j \le 2$. Then formulae (2.80) and (2.48) yield

$$G(V, \eta_1)(e_1, e_1) = DW(e_1, D_{e_1}V) = bW_{11} - aW_{12}; \tag{2.81}$$

$$G(V, \eta_1)(e_2, e_2) = DW(e_2, D_{e_2}V) = aW_{21} + bW_{22} \quad \text{at } x. \tag{2.82}$$

It follows from (2.81), (2.82), and (2.79) that

$$\langle \Upsilon(\eta_1), G(V, \eta_1)\rangle_{T_x^2} = b(W_{11}^2 + W_{22}^2) + aW_{21}(W_{11} + W_{22}). \tag{2.83}$$

Similarly, we may obtain

$$\mathrm{tr}\Upsilon(\eta_1)\mathrm{tr}G(V, \eta_1) = b(W_{11} + W_{22})^2 + 2aW_{21}(W_{11} + W_{22}) \quad \text{at } x. \tag{2.84}$$

Thus relations (2.83) and (2.84) yield

$$\begin{aligned} a(\Upsilon(\eta_1), G(V, \eta_1)) &= ba(\Upsilon(\eta_1), \Upsilon(\eta_1)) + (1 + \mu)aW_{21}(W_{11} + W_{22}) \\ &\ge ba(\Upsilon(\eta_1), \Upsilon(\eta_1)) - \frac{1 + \mu}{2}|a||DW|_{T_x^2}^2. \end{aligned} \tag{2.85}$$

By inequality (1.39) and (2.85), we obtain inequality (2.77). □

PROPOSITION 2.12. *Let* $\eta = (W, w) \in H^1(\Omega, \Lambda) \times H^2(\Omega)$ *and* $\hat{\Gamma} \subset \Gamma$ *be relatively open such that*

$$W|_{\hat{\Gamma}} = 0, \quad w|_{\hat{\Gamma}} = \frac{\partial w}{\partial n}|_{\hat{\Gamma}} = 0. \tag{2.86}$$

Then

$$\text{(i)} \quad B(\eta, \eta) = \gamma(\Delta w)^2 + (DW(n,n))^2 + \frac{1-\mu}{2}(DW(\tau, n))^2 \quad \text{on} \quad \hat{\Gamma}; \tag{2.87}$$

$$\text{(ii)} \quad \partial(\mathcal{A}\eta, m(\eta)) = B(\eta, \eta)\langle V, n\rangle \quad \text{on} \quad \hat{\Gamma}. \tag{2.88}$$

Proof. Given $x \in \hat{\Gamma}$, conditions $W|_{\hat{\Gamma}} = w|_{\hat{\Gamma}} = 0$ imply

$$D_\tau W = 0; \tag{2.89}$$

$$\Upsilon(\eta)(\tau, \tau) = DW(\tau, \tau) = \langle D_\tau W, \tau\rangle = 0; \tag{2.90}$$

$$\Upsilon(\eta)(n, \tau) = \frac{1}{2}[DW(\tau, n) + \langle D_\tau W, n\rangle] = \frac{1}{2}DW(\tau, n); \tag{2.91}$$

$$\Upsilon(\eta)(n, n) = DW(n, n). \tag{2.92}$$

It follows from (2.89)–(2.92) that

$$|\Upsilon(\eta)|^2_{T^2_x} = (DW(n,n))^2 + \frac{1}{2}(DW(\tau, n))^2; \quad (\text{tr}\Upsilon(\eta))^2 = (DW(n,n))^2, \tag{2.93}$$

so that we obtain

$$a(\Upsilon(\eta), \Upsilon(\eta)) = (DW(n,n))^2 + \frac{1-\mu}{2}(DW(\tau, n))^2 \quad \text{at } x. \tag{2.94}$$

A similar computation yields

$$a(\rho(\eta), \rho(\eta)) = (\Delta w)^2 \quad \text{at } x. \tag{2.95}$$

Inserting (2.94) and (2.95) into (1.25), we obtain (2.87).

By Yao [38, Lemma 2.5] and (2.86), we have

$$\frac{\partial(V(w))}{\partial n} = \langle V, n\rangle \Delta w \quad \text{and} \quad V(w) = \langle Dw, V\rangle = 0 \quad \text{on} \quad \hat{\Gamma}. \tag{2.96}$$

Since $D_V W = \langle V, n\rangle D_n W + \langle V, \tau\rangle D_\tau W = \langle V, n\rangle D_n W$, we obtain

$$\langle D_V W, n\rangle = \langle V, n\rangle DW(n,n) \quad \text{and} \quad \langle D_V W, \tau\rangle = \langle V, n\rangle DW(\tau, n) \quad \text{on} \quad \hat{\Gamma}. \tag{2.97}$$

In addition, condition (2.86) implies

$$B_1(W, w) = (1-\mu)DW(n,n) + \mu DW(n,n) + \mu\langle D_\tau W, \tau\rangle = DW(n,n); \tag{2.98}$$

$$B_2(W, w) = \frac{1-\mu}{2}[DW(\tau, n) + \langle D_\tau W, n\rangle] = \frac{1-\mu}{2}DW(\tau, n); \tag{2.99}$$

$$B_3 w = -D^2 w(\tau, \tau) = -\langle D_\tau(Dw), \tau\rangle = 0; \tag{2.100}$$

$$B_4 w = \frac{\partial}{\partial \tau}\left(D^2 w(\tau, n)\right) = \frac{\partial}{\partial \tau}\left(D^2 w(n, \tau)\right) = \frac{\partial}{\partial \tau}\left(\langle D_\tau(Dw), n\rangle\right) = 0. \tag{2.101}$$

Inserting (2.96)–(2.101) into (1.30), we obtain (2.88) by (i). □

PROPOSITION 2.13. *Let λ be a complex number and $\hat{\Gamma} \subset \Gamma$ be relatively open. Let $\eta = (W, w)$ solve the problem*

$$\lambda^2 \eta - \lambda^2 \gamma(0, \Delta w) + \mathcal{A}\eta = 0 \quad on \quad \Omega \tag{2.102}$$

subject to boundary conditions

$$\begin{cases} W = DW = 0 \quad on \quad \hat{\Gamma}, \\ w = \dfrac{\partial w}{\partial n} = \Delta w = \dfrac{\partial \Delta w}{\partial n} = 0 \quad on \quad \hat{\Gamma}. \end{cases} \tag{2.103}$$

Then

$$W = w = 0 \quad on \quad \Omega. \tag{2.104}$$

Proof. It will suffice to show that this uniqueness result is the Cauchy problem consisting of a system of three equations of the fourth order with the same principal part Δ^2, where Δ is the Laplacian on manifold M. Therefore, this proposition is covered by Shirota [27] .

Denote by $LF(\eta)$ some terms for which there is constant C such that

$$|LF(\eta)|^2 \le C \sum_{i=0}^{3} (|D^i W|^2 + |D^i w|^2) \quad \text{on} \quad \Omega. \tag{2.105}$$

It is easily checked from (2.102) and (1.31) that we have the following equation on component W:

$$\frac{1-\mu}{2} \delta d W + d\delta W = (k - \mu k - \lambda^2) W + \mathcal{F}(w). \tag{2.106}$$

Applying $d\delta$ to both sides of equation (2.106), we obtain

$$d\delta d\delta W = d\delta[(k - \mu k - \lambda^2) W + \mathcal{F}(w)], \tag{2.107}$$

since $\delta^2 = 0$. Similarly, it follows that

$$\delta d \delta d W = \frac{2}{1-\mu} \delta d[(k - \mu k - \lambda^2) W + \mathcal{F}(w)]. \tag{2.108}$$

Relations (2.107) and (2.108) yield

$$\boldsymbol{\Delta}^2 W = LF(\eta), \tag{2.109}$$

where $\boldsymbol{\Delta}$ is the Hodge–Laplacian, defined by $\boldsymbol{\Delta} = \delta d + d\delta$.

Let (x_1, x_2) be a coordinate and $W = (w_1, w_2)$ in this coordinate. It is easy to check that

$$\boldsymbol{\Delta}^2 W = (\Delta^2 w_1, \Delta^2 w_2) + LF(w_1, w_2). \tag{2.110}$$

By (2.109) and (2.110), system (2.102) is equivalent to the following

$$\begin{cases} \Delta^2 w_1 = LF(w_1, w_2, w), \\ \Delta^2 w_2 = LF(w_1, w_2, w), \\ \Delta^2 w = LF(w_1, w_2, w). \end{cases} \tag{2.111}$$

To complete the proof, we have to show that there are the Cauchy data on $\hat{\Gamma}$ enough for problem (2.111). For this purpose, by boundary conditions (2.103), it will suffice to show

$$D^2 W|_{\hat{\Gamma}} = D^3 W|_{\hat{\Gamma}} = 0. \tag{2.112}$$

Given $x \in \hat{\Gamma}$, let E_1, E_2 be a frame field normal at x such that $E_1(x) = n$ and $E_2(x) = \tau$. By (2.103), we obtain at x

$$D^2 W(E_i, E_j, \tau) = \tau(DW(E_i, E_j)) - DW(D_\tau E_i, E_j) - DW(E_i, D_\tau E_j) = 0, \tag{2.113}$$

for $1 \le i, j \le 2$. It follows from (1.13), (2.103), and (2.113) that at x

$$D^2 W(E_i, \tau, E_j) = D^2 W(E_i, E_j, \tau) + R(\tau, E_j, W, E_i) = 0, \tag{2.114}$$

for $1 \le i, j \le 2$.

By Wu [35, pp. 305–306] and (2.113–2.114), we have at x

$$\begin{aligned} \delta dW &= -\sum_{i=1}^{2} D_{E_i} D_{E_i} W + \sum_{ij=1}^{2} \langle D_{E_j} D_{E_i} W, E_j \rangle E_i \\ &= -\sum_{i=1}^{2} [D^2 W(n, E_i, E_i) n + D^2 W(\tau, E_i, E_i)\tau] + \sum_{i=1}^{2} \left[\sum_{j=1}^{2} D^2 W(E_j, E_i, E_j) \right] E_i \\ &= -D^2 W(\tau, n, n)\tau; \end{aligned} \tag{2.115}$$

$$d\delta W = -\sum_{i=1}^{2} \left[\sum_{j=1}^{2} D^2 W(E_j, E_j, E_i) \right] E_i = -D^2 W(n, n, n) n. \tag{2.116}$$

After inserting (2.115) and (2.116) into (2.106), we obtain by (2.103)

$$-D^2 W(n, n, n) n - \frac{(1-\mu) D^2 W(\tau, n, n)}{2} \tau = (k - \mu k - \lambda^2) W + \mathcal{F}(w) = 0 \tag{2.117}$$

at x, that is, $D^2 W(n, n, n) = D^2 W(\tau, n, n) = 0$, which implies

$$D^2 W = 0 \quad \text{at } x. \tag{2.118}$$

Similarly, we may obtain $D^3 W|_{\hat{\Gamma}} = 0$. □

2.3. Proofs of main results.

Proof of Theorem 1.1. Since $\Upsilon(\eta) = \Upsilon(\eta_1) + l(\eta)$ and $G(V,\eta) = G(V,\eta_1)$, inequality (2.77) yields

$$\int_\Omega a(\Upsilon(\eta), G(V,\eta))\,dx \geq \sigma_1 \int_\Omega B(\eta,\eta_1)\,dx + l(\eta). \tag{2.119}$$

In addition, it is easily checked that $B(\eta_1,\eta_2) = l(\eta)$ so that we have

$$B(\eta,\eta_2) = B(\eta_2,\eta_2) + l(\eta). \tag{2.120}$$

We now are ready to establish inequality (1.49).

By (1.50), (2.66), and (2.119), we obtain the following:

$$\begin{aligned} SB|_\Sigma = {} & Z|_0^T + \int_Q b[|\eta_t|^2 + B(\eta,\eta_2) - B(\eta,\eta_1)]\,dQ \\ & + \int_Q a(\Upsilon(\eta), G(V,\eta))\,dQ + lot(\eta) \\ & + \frac{1}{2}\int_\Sigma \left[\partial(\mathcal{A}\eta, -b\eta_2 + h\eta_1) - \gamma b w \frac{\partial w_{tt}}{\partial n}\right] d\Sigma \\ \geq {} & Z|_0^T + \int_Q b(|W_t|^2 + w_t^2)\,dQ \\ & + \int_Q bB(\eta,\eta_2)\,dQ + \int_Q (\sigma_1 - b)B(\eta,\eta_1)\,dQ + lot(\eta) \\ & + \frac{1}{2}\int_\Sigma \left[\partial(\mathcal{A}\eta, -b\eta_2 + h\eta_1) - \gamma b w \frac{\partial w_{tt}}{\partial n}\right] d\Sigma. \end{aligned} \tag{2.121}$$

Now rewrite the sum of the second, the third, and the fourth terms in the right-hand side of inequality (2.121), as follows in terms of identities (2.39) and (2.40):

$$\begin{aligned} & \int_Q b(|W_t|^2 + w_t^2)\,dQ + \int_Q bB(\eta,\eta_2)\,dQ + \int_Q (\sigma_1 - b)B(\eta,\eta_1)\,dQ \\ & = \sigma_1 \int_0^T E(t)dt + \frac{1}{2}\int_Q (b - \sigma_1)[w_t^2 + \gamma|Dw_t|^2 + B(\eta_2,\eta_2)]\,dQ + \int_Q b w_t^2\,dQ \end{aligned} \tag{2.122}$$

$$\begin{aligned} & + \frac{1}{2}\int_Q (2b - \sigma_1)[|W_t|^2 - B(\eta,\eta_1)]\,dQ + \frac{1}{2}\int_Q b[B(\eta,\eta_2) - w_t^2 - \gamma|Dw_t|^2]\,dQ + lot(\eta) \\ & \geq \sigma_1 \int_0^T E(t)dt + \frac{1}{2}\int_\Sigma \left[\partial(\mathcal{A}\eta, b\eta_2 - h\eta_1) + \gamma b w \frac{\partial w_{tt}}{\partial n}\right] d\Sigma + lot(\eta), \end{aligned} \tag{2.123}$$

where $h = 2b - \sigma_1$. Inserting (2.123) into (2.121), we obtain the desired inequality (1.49) by means of inequality (2.74).

Proofs of Theorems 1.2 *and* 1.3. By a compactness/uniqueness argument as in Bardos, Lebeau, and Rauch [1] or in Zuazua [41], the lower order terms in estimate (1.49) can be absorbed and we obtain both inequalities (1.56) and (1.67), since in both cases the uniqueness we need is covered by Proposition 2.13. It is mentioned that in both cases we also have $E(t) = E(0)$. □

Acknowledgments. Some results of this paper were obtained when the author was visiting the Institute of Applied Mathematics and Mechanics, University of Virginia. The author would like to thank Professors I. Lasiecka and R. Triggiani for bringing this topic to him and for their hospitality.

REFERENCES

[1] C. Bardos, G. Lebeau, and J. Rauch, *Sharp sufficient conditions for the observation, control, and stabilization of wave from the boundary*, SIAM J. Control Optim., 30 (1992), pp. 1024–1065.

[2] M. Bernadou, *Méthodes d'Éléments Finis pour les Problèmes de Coques Minces*, Masson, Paris, 1994.

[3] M. Bernadou and J. M. Boisserie, *The finite element method in thin shell theory: Application to arch dam simulations*, Progress in Scientific Computing I, Birkhäuser, Boston, 1982.

[4] M. Bernadou, P. G. Ciarlet, and B. Miara, *Existence theorems for two-dimensional linear shell theories*, J. Elasticity, 34, (1994), pp. 111–138.

[5] M. Bernadou and J. T. Oden, *An existence theorem for a class of nonlinear shallow shell problems*, J. Math. Pures Appl., 60 (1981), pp. 285–308.

[6] J. Cheeger and D. Ebin, *Comparison Theorem in Riemannian Geometry*, North-Holland, Amsterdam, 1975.

[7] G. Chen, *Energy decay estimates and exact boundary value controllability for the wave equation in a bounded domain*, J. Math. Pures Appl., 58 (1979), pp. 249–274.

[8] G. Chen, M. P. Coleman, and K. Liu, *Boundary stabilization of Donnell's shallow circular cylindrical shell*, J. Sound Vibration, 209 (1998), pp. 265–298.

[9] P. G. Ciarlet and J. C. Paumier, *Justification of the Marguerre-von Karman equations*, Comput. Mech., 1 (1986), pp. 177–202.

[10] P. G. Ciarlet and B. Miara, *Justification of the two-dimensional equations of a linearly elastic shallow shell*, Comm. Pure Appl. Math., 45 (1992), pp. 327–360.

[11] L. H. Donnell, *Beams, Plates, and Shells*, McGraw-Hill, New York, 1976.

[12] G. Geymont, P. Loreti, and V. Valente, *Exact controllability of a shallow shell model*, in Internat. Ser. Numer. Math. 107, Birkhäuser, Basel, 1992, pp. 85–97.

[13] R. E. Greene and H. Wu, *Function Theory on Manifolds which Possess a Pole*, Lecture Notes in Math. 699, Springer-Verlag, New York, 1979.

[14] A. E. Green and W. Zerna, *Theoretical Elasticity*, 2nd ed., Clarendon, Oxford, UK, 1968.

[15] E. Hebey, *Sobolev Spaces on Riemannian Manifolds*, Lecture Notes in Math. 1635, Springer-Verlag, New York, 1996.

[16] L. Hörmander, *On the uniqueness of the Cauchy problem* II, Math. Scand., 7 (1959), pp. 177–190.

[17] W. T. Koiter, *A consistent first approximation in the general theory of thin elastic shells*, in Proceedings of the IUTAM Symposium on the Theory of Thin Shells, Delft (August 1959), North-Holland, Amsterdam, 1960, pp. 12–33.

[18] W. T. Koiter, *On the nonlinear theory of thin elastic shells*, Proc. Konink. Nederl. Akad. Wetensch., B69 (1966), pp. 1–54.

[19] V. Komornik, *Exact Controllability and Stabilization*, Research in Applied Mathematics, CSC, Masson, John Wiley, Paris, New York, 1994.

[20] J. Lagnese, *Boundary Stabilization of Thin Plates*, SIAM Stud. Appl. Math., SIAM, Philadelphia, PA, 1989.

[21] J. Lagnese and J.L. Lions, *Modelling, Analysis and Control of Thin Plates*, Recherches en Mathématiques Appliquées, Masson, Paris, 1988.

[22] I. Lasiecka and R. Triggiani, *Exact controllability of the wave equation with Neumann boundary control*, Appl. Math. Optim., 19 (1989), pp. 243–290.

[23] I. Lasiecka, R.Triggiani, and V. Valente, *Uniform stabilization of spherical shells by boundary dissipation*, Adv. Differential Equations, 1 (1996), pp. 635–674.

[24] J. L. Lions, *Exact controllability, stabilization and perturbations for distributed systems*, SIAM Rev. 30 (1988), pp. 1–68.

[25] F. I. Niordson, *Shell Theory*, North–Holland Series in Applied Mathematics and Mechanics 29, North-Holland, Amsterdam, 1985.

[26] R. Schoen and S.-T. Yau, *Lectures on Differential Geometry*, Conference Proceedings and Lecture Notes in Geometry and Topology I, International Press, Cambridge, MA, 1994.

[27] T. Shirota, *A remark on the unique continuation theorem for certain fourth order elliptic equations*, Proc. Japan Acad. Ser. A. Math. Sci., 36 (1960), pp. 571–573.
[28] D. Tataru, *A-priori estimates of Carleman's type in domains with boundary*, J. Math. Pures Appl., 73 (1994), pp. 355–387.
[29] D. Tataru, *Boundary controllability for conservative PDEs*, Appl. Math. Optim., 31 (1995), pp. 257–295.
[30] D. Tataru, *Carleman estimates and unique continuation for solutions to boundary value problems*, J. Math. Pures Appl., 75 (1996), pp. 367–408.
[31] M. E. Taylor, *Partial Differential Equations* I, Spring-Verlag, New York, 1996.
[32] R. Triggiani, *Regularity theory, exact controllability and optimal quadratic cost problem for spherical shells with physical boundary controls*, Control Cybernet., 25 (1996), pp. 553–568.
[33] H. Wu, C. L. Shen, and Y. L. Yu, *An Introduction to Riemannian Geometry*, University of Beijing, Beijing, 1989 (in Chinese).
[34] H. Wu, *Selected Lectures in Riemannian Geometry*, University of Beijing, Beijing, 1981 (in Chinese).
[35] H. Wu, *The Bochner technique in differential geometry*, Math. Rep., 3, (1988), pp. 289–538.
[36] P. F. Yao, *On the observability inequalities for exact controllability of wave equations with variable coefficients*, SIAM J. Control and Optim., 37 (1999), pp. 1568–1599.
[37] P. F. Yao, *On the Shallow Equations*, manuscript.
[38] P. F. Yao, *The Observability Inequalities for the Euler–Bernoulli Equations with Variable Coefficients*, manuscript.
[39] P. F. Yao, *Koiter's Model and Observability Inequalities*, manuscript.
[40] P. F. Yao, *Naghdi's Model and Observability Inequalities*, manuscript.
[41] E. Zuazua, *Contrôlabilité exacte de quelques modèles de plaques en temps arbitrairement petit*, Appendix I in Contrôlabilité exacte, stabilisation et perturbations de systèmes distribués. Tome 1. Contrôlabilité exacte, Masson, Paris, Recherches en Mathématiques Appliquées 8, pp. 465–491.

SIAM J. CONTROL OPTIM.
Vol. 38, No. 6, pp. 1757–1793

STABILITY RADIUS AND INTERNAL VERSUS EXTERNAL STABILITY IN BANACH SPACES: AN EVOLUTION SEMIGROUP APPROACH*

STEPHEN CLARK†, YURI LATUSHKIN‡, STEPHEN MONTGOMERY-SMITH‡, AND TIMOTHY RANDOLPH†

Abstract. In this paper the theory of evolution semigroups is developed and used to provide a framework to study the stability of general linear control systems. These include autonomous and nonautonomous systems modeled with unbounded state-space operators acting on Banach spaces. This approach allows one to apply the classical theory of strongly continuous semigroups to time-varying systems. In particular, the complex stability radius may be expressed explicitly in terms of the generator of an (evolution) semigroup. Examples are given to show that classical formulas for the stability radius of an autonomous Hilbert-space system fail in more general settings. Upper and lower bounds on the stability radius are proven for Banach-space systems. In addition, it is shown that the theory of evolution semigroups allows for a straightforward operator-theoretic analysis of internal stability as determined by classical frequency-domain and input-output operators, even for nonautonomous Banach-space systems. In particular, for the nonautonomous setting, internal stability is shown to be equivalent to input-output stability for stabilizable and detectable systems. For the autonomous setting, an explicit formula for the norm of input-output operator is given.

Key words. evolution semigroups, stability radius, exponential stability, external stability, spectral mapping theorem, transfer function

AMS subject classifications. 47D06, 34G10, 93C25, 93D09, 93D25

PII. S036301299834212X

1. Introduction.

Presented here is a study of stability of infinite-dimensional linear control systems which is based on the relatively recent development of the theory of evolution semigroups. These semigroups have been used in the study of exponential dichotomy of time-varying differential equations and more general hyperbolic dynamical systems; see [7, 23, 24, 27, 29, 34, 43] and the bibliographies therein. The intent of this paper is to show how the theory of evolution semigroups can be used to provide a clarifying perspective, and prove new results, on the uniform exponential stability for general linear control systems, $\dot{x}(t) = A(t)x(t) + B(t)u(t)$, $y(t) = C(t)x(t)$, $t \geq 0$. The operators $A(t)$ are generally unbounded operators on a Banach space X, while the operators $B(t)$ and $C(t)$ may act on Banach spaces U and Y, respectively. In addressing the general settings, difficulties arise both from the time-varying aspect and from a loss of Hilbert-space properties. This presentation, however, provides some relatively simple operator-theoretic arguments for properties that extend classical theorems of autonomous systems in finite dimensions. The topics

*Received by the editors July 20, 1998; accepted for publication (in revised form) November 1, 1999; published electronically June 15, 2000. A portion of this article appeared in preliminary form in *Proceedings of the* 36*th IEEE Conference on Decision and Control*, San Diego, CA, 1997.
http://www.siam.org/journals/sicon/38-6/34212.html

†Department of Mathematics and Statistics, University of Missouri–Rolla, Rolla, MO 65409 (sclark@umr.edu, randolph@umr.edu). The research of the fourth author was supported by the Missouri Research Board; a portion of the work by this author was carried out while visiting the Department of Mathematics, University of Missouri–Columbia.

‡Mathematics Department, University of Missouri–Columbia, Columbia, MO 65201 (yuri@math.missouri.edu, stephen@math.missouri.edu). The research of both authors was supported by the National Science Foundation. The research of the second author was supported by the Missouri Research Board.

covered here include characterizing internal stability of the nominal system in terms of appropriate input-state-output operators and, subsequently, using these properties to obtain new explicit formulas for bounds on the stability radius. Nonautonomous systems are generally considered, but some results apply only to autonomous ones, such as the upper bound for the stability radius (section 4.3), the formula for the norm of the input-output operator in Banach spaces (section 4.4), and a characterization of stability that is related to this formula (section 5.2).

Although practical considerations usually dictate that U and Y are Hilbert spaces (indeed, finite dimensional), the Banach-space setting addressed here may be motivated by the problem of determining optimal sensor (or actuator) location. For this, it may be natural to consider $U = X$ and $B = I_X$ (or $Y = X$ and $C = I_X$) [5]; if the natural state space X is a Banach space, then, as will be shown in this paper, Hilbert-space characterizations of internal stability or its robustness do not apply. We also show that even in the case of Hilbert spaces U and Y, known formulas for the stability radius involving the spaces $L^2(\mathbb{R}_+, U)$ and $L^2(\mathbb{R}_+, Y)$ do not apply if the L^2 norm is replaced by, say, the L^1 norm—see the examples in subsection 4.5. In addition to the general setting of nonautonomous systems on Banach spaces, autonomous and Hilbert-space systems are considered.

For the autonomous case, the primary observation we make about general Banach-space settings versus the classical L^2 and Hilbert-space setting can be explained using the notion of L^p–Fourier multipliers. For this, let $H(s) = C(A - is)^{-1}B$, $s \in \mathbb{R}$, denote the transfer function, and let $\mathbf{F}$ denote the Fourier transform. The transfer function H is said to be an *L^p–Fourier multiplier* if the operator $u \mapsto \mathbf{F}^{-1}H(\cdot)\mathbf{F}u$ can be extended from the Schwartz class of rapidly decaying U-valued functions to a bounded operator from $L^p(\mathbb{R}; U)$ to $L^p(\mathbb{R}; Y)$; see, e.g., [1] for the definitions. As shown in Theorem 4.11, the norm of this operator is equal to the norm of the input-output operator. If U and Y are Hilbert spaces and $p = 2$, then H is an L^2–Fourier multiplier if and only if $\|H(\cdot)\|$ is bounded on $\mathbb{R}$; see formula (4.20). For Banach spaces and/or $p \neq 2$, this latter condition is necessary but *not sufficient* for H to be an L^p–Fourier multiplier. As a result, our formula (4.18) for the norm of the input-output operator is more involved.

To motivate the methods, recall Lyapunov's stability theorem which says that if A is a bounded linear operator on X and if the spectrum of A is contained in the open left half of the complex plane, then the solution of the autonomous differential equation $\dot{x}(t) = Ax(t)$ on X is uniformly exponentially stable; equivalently, the spectrum $\sigma(e^{tA})$ is contained in the open unit disk $\{z \in \mathbb{C} : |z| < 1\}$ for $t > 0$. This is a consequence of the fact that when A is a bounded operator, then the spectral mapping theorem holds: $\sigma(e^{tA}) \setminus \{0\} = e^{t\sigma(A)}$, $t \neq 0$. Difficulties with Lyapunov's theorem arise when the operators A are allowed to be unbounded. In particular, it is well known that there exist strongly continuous semigroups $\{e^{tA}\}_{t\geq 0}$ that are *not* uniformly exponentially stable even though $\mathrm{Re}\lambda \leq \omega < 0$ for all $\lambda \in \sigma(A)$; see, e.g., [2, 29, 41]. For nonautonomous equations the situation is worse. Indeed, even for finite-dimensional X it is possible for the spectra of $A(t)$ to be the same for all $t > 0$ *and* contained in the open left half-plane, yet the corresponding solutions to $\dot{x}(t) = A(t)x(t)$ are not uniformly exponentially stable (see [15, Ex. 7.1] for a classical example). In the development that follows we plan to show how these difficulties can be overcome by the construction of an "evolution semigroup." This is a family of operators defined on a superspace of functions from $\mathbb{R}$ into X, such as $L^p(\mathbb{R}, X)$, $1 \leq p < \infty$, or $C_0(\mathbb{R}, X)$.

Section 2 sets up the notation and provides background information. Section 3 presents the basic properties of the evolution semigroups. Included here is the property that the spectral mapping theorem always holds for these semigroups when they are defined on X-valued functions on the *half-line*, such as $L^p(\mathbb{R}_+, X)$. A consequence of this is a characterization of exponential stability for nonautonomous systems in terms of the invertibility of the generator Γ of the evolution semigroup. This operator and its role in determining exponential stability is the basis for many of the subsequent developments. In particular, the semigroup $\{e^{tA}\}_{t\geq 0}$ *is* uniformly exponentially stable provided $\operatorname{Re}\lambda < 0$ for all $\lambda \in \sigma(\Gamma)$.

Section 4 addresses the topic of the (complex) stability radius; that is, the size of the smallest disturbance, $\Delta(\cdot)$, under which the perturbation, $\dot{x}(t) = (A(t)+\Delta(t))x(t)$, of an exponentially stable system, $\dot{x}(t) = A(t)x(t)$, loses exponential stability. Results address structured and unstructured perturbations of autonomous and nonautonomous systems in both Banach- and Hilbert-space settings. Examples are given which highlight some important differences between these settings. Also included in this section is a discussion about the transfer function for infinite-dimensional *time-varying* systems. This concept arises naturally in the context of evolution semigroups.

In section 5 the explicit relationship between internal and external stability is studied for general linear systems. This material expands on the ideas begun in [37]. A classical result for autonomous systems in Hilbert space is the fact that exponential stability of the nominal system (internal stability) is, under the hypotheses of stabilizability and detectability, equivalent to the boundedness of the transfer function in the right half-plane (external stability). Such a result does not apply to nonautonomous systems, and a counterexample shows that this property fails to hold for Banach-space systems. Properties from section 4 provide a natural Banach-space extension of this result: the role of transfer function is replaced by the input-output operator. Moreover, for autonomous systems we provide an explicit formula relating the norm of this input-output operator to that of the transfer function. Finally, we prove two theorems—one for nonautonomous and one for autonomous systems—which characterize internal stability in terms of the various input-state-output operators.

This introduction concludes with a brief synopsis of the main results. The characterization of uniform exponential stability in terms of an evolution semigroup and its generator is given in Theorem 3.2, Theorem 3.5, and Corollary 3.6. Although these results are essentially known, the proofs are approached in a new way. In particular, Theorem 3.5 identifies the operator $\mathbb{G} = -\Gamma^{-1}$ used to determine stability throughout the paper. Theorem 4.2 records the main observation that the input-output operator, $\mathbb{L} = \mathcal{C}\mathbb{G}\mathcal{B}$, for a general nonautonomous system is related to the inverse of the generator of the evolution semigroup. A very short proof of the known fact that the stability radius for such a system is bounded from below by $\|\mathbb{L}\|^{-1}$ is also provided here. The upper bound for the stability radius, being given in terms of the transfer function, applies only to autonomous systems and is proven in subsection 4.3. The upper bound, as identified here for Banach spaces, seems to be new although our proof is based on the idea of the Hilbert-space result of [18, Thm. 3.5]. In subsection 4.3 we also introduce the pointwise stability radius and dichotomy radius. Estimates for the former are provided by Theorems 4.3 and 4.4 while the latter is addressed in Lemma 4.5. Examples 4.13 and 4.15 show that, for autonomous Banach-space systems, both inequalities for the upper and lower bounds on the stability radius (see Theorem 4.1) can be strict. In view of the possibility of the strict inequality $\|\mathbb{L}\| > \sup_{s\in\mathbb{R}} \|C(A-is)^{-1}B\|$, Theorem 4.11 provides a new Banach-space formula for $\|\mathbb{L}\|$ in terms of A, B, and C. In

section 5 this expression for $\|L\|$ is used to relate state-space versus frequency-domain stability—concepts which are *not* equivalent for Banach-space systems. A special case of this expression gives a new formula for the growth bound of a semigroup on a Banach space; see Theorem 5.4 and the subsequent paragraph. Finally, Theorem 5.3 extends a classical characterization of stability for stabilizable and detectable control systems as it applies to nonautonomous Banach-space settings.

2. Notation and preliminaries. Throughout the paper, $\mathcal{L}(X,Y)$ will denote the set of bounded linear operators between complex Banach spaces X and Y. If A is a linear operator on X, $\sigma(A)$ will denote the spectrum of A, $\rho(A)$ will denote the resolvent set of A relative to $\mathcal{L}(X) = \mathcal{L}(X,X)$, and $\|A\|_\bullet = \|A\|_{\bullet,X} := \inf\{\|Ax\| : x \in \mathrm{Dom}(A), \|x\| = 1\}$. In particular, if A is invertible in $\mathcal{L}(X)$, $\|A\|_\bullet = 1/\|A^{-1}\|_{\mathcal{L}(X)}$. Also, let $\mathbb{C}_+ = \{\lambda \in \mathbb{C} : \mathrm{Re}\ \lambda > 0\}$.

If A generates a strongly continuous (or C_0) semigroup $\{e^{tA}\}_{t\geq 0}$ on a Banach space X, the following notation will be used: $s(A) = \sup\{\mathrm{Re}\lambda : \lambda \in \sigma(A)\}$ denotes the spectral bound; $s_0(A) = \inf\{\omega \in \mathbb{R} : \ \{\lambda : \mathrm{Re}\lambda > \omega\} \subset \rho(A) \quad \text{and} \quad \sup_{\mathrm{Re}\lambda>\omega} \|(A - \lambda)^{-1}\| < \infty\}$ is the abscissa of uniform boundedness of the resolvent; and $\omega_0(e^{tA}) = \inf\{\omega \in \mathbb{R} : \ \|e^{tA}\| \leq Me^{t\omega} \quad \text{for some} \quad M \geq 0 \quad \text{and all} \quad t \geq 0\}$ denotes the growth bound of the semigroup. In general, $s(A) \leq s_0(A) \leq \omega_0(e^{tA})$ (see, e.g., [29]) with *strict* inequalities possible; see [2, 29, 41] for examples. However, when X is a Hilbert space, the following spectral mapping theorem of Gearhart holds (see, e.g., [2, p. 95] or [29, 33]).

THEOREM 2.1. *If A generates a strongly continuous semigroup $\{e^{tA}\}_{t\geq 0}$ on a Hilbert space, then $s_0(A) = \omega_0(e^{tA})$. Moreover, $1 \in \rho(e^{2\pi A})$ if and only if $i\mathbb{Z} \subset \rho(A)$ and $\sup_{k\in\mathbb{Z}} \|(A - ik)^{-1}\| < \infty$.*

In particular, this result shows that on a Hilbert space X the semigroup $\{e^{tA}\}_{t\geq 0}$ is uniformly exponentially stable if and only if $\sup_{\lambda\in\mathbb{C}_+} \|(A - \lambda)^{-1}\| < \infty$ [19].

Now consider operators $A(t)$, $t \geq 0$, with domain $\mathrm{Dom}(A(t))$ in a Banach space X. If the abstract Cauchy problem

$$\dot{x}(t) = A(t)x(t), \quad x(\tau) \in \mathrm{Dom}(A(\tau)), \qquad t \geq \tau \geq 0, \tag{2.1}$$

is well-posed in the sense that there exists an evolution (solving) family of operators $\mathcal{U} = \{U(t,\tau)\}_{t\geq\tau}$ on X which gives a differentiable solution, then $x(\cdot) : t \mapsto U(t,\tau)x(\tau)$, $t \geq \tau$, in $\mathbb{R}$ is differentiable, $x(t)$ is in $\mathrm{Dom}(A(t))$ for $t \geq 0$, and (2.1) holds. The precise meaning of the term evolution family used here is as follows.

DEFINITION 2.2. *A family of bounded operators $\{U(t,\tau)\}_{t\geq\tau}$ on X is called an* evolution family *if*

(i) *$U(t,\tau) = U(t,s)U(s,\tau)$ and $U(t,t) = I$ for all $t \geq s \geq \tau$;*

(ii) *for each $x \in X$ the function $(t,\tau) \mapsto U(t,\tau)x$ is continuous for $t \geq \tau$.*

An evolution family $\{U(t,\tau)\}_{t\geq\tau}$ is called exponentially bounded *if, in addition,*

(iii) *there exist constants $M \geq 1$, $\omega \in \mathbb{R}$ such that*

$$\|U(t,\tau)\| \leq Me^{\omega(t-\tau)}, \quad t \geq \tau.$$

Remarks 2.3.

(a) An evolution family $\{U(t,\tau)\}_{t\geq\tau}$ is called *uniformly exponentially stable* if in part (iii), ω can be taken to be strictly less than zero.

(b) Evolution families appear as solutions for abstract Cauchy problems (2.1). Since the definition requires that $(t,\tau) \mapsto U(t,\tau)$ is merely strongly continuous, the operators $A(t)$ in (2.1) can be unbounded.

(c) In the autonomous case where $A(t) \equiv A$ is the infinitesimal generator of a strongly continuous semigroup $\{e^{tA}\}_{t\geq 0}$ on X, then $U(t,\tau) = e^{(t-\tau)A}$ for $t \geq \tau$ is a strongly continuous exponentially bounded evolution family.

(d) The existence of a *differentiable* solution to (2.1) plays little role in this paper, so the starting point will usually not be (2.1), but rather the existence of an exponentially bounded evolution family.

In the next section we will define the evolution semigroup relevant to our interests for the nonautonomous Cauchy problem (2.1) on the half-line $\mathbb{R}_+ = [0,\infty)$. For now, we begin by considering the autonomous equation $\dot{x}(t) = Ax(t)$, $t \in \mathbb{R}$, where A is the generator of a strongly continuous semigroup $\{e^{tA}\}_{t\geq 0}$ on X. If $\mathcal{F}_{\mathbb{R}}$ is a space of X-valued functions, $f : \mathbb{R} \to X$, define

$$(E^t_{\mathbb{R}} f)(\tau) = e^{tA} f(\tau - t) \quad \text{for } f \in \mathcal{F}_{\mathbb{R}}. \tag{2.2}$$

If $\mathcal{F}_{\mathbb{R}} = L^p(\mathbb{R}, X)$, $1 \leq p < \infty$, or $\mathcal{F}_{\mathbb{R}} = C_0(\mathbb{R}, X)$, the space of continuous functions vanishing at infinities (or another Banach function space as in [34]) this defines a strongly continuous semigroup of operators $\{E^t_{\mathbb{R}}\}_{t\geq 0}$ whose generator will be denoted by $\Gamma_{\mathbb{R}}$. In the case $\mathcal{F}_{\mathbb{R}} = L^p(\mathbb{R}, X)$, $\Gamma_{\mathbb{R}}$ is the closure (in $L^p(\mathbb{R}, X)$) of the operator $-d/dt + \mathcal{A}$, where $(\mathcal{A}f)(t) = Af(t)$ and

$$\begin{aligned}\operatorname{Dom}(-d/dt + \mathcal{A}) &= \operatorname{Dom}(-d/dt) \cap \operatorname{Dom}(\mathcal{A}) \\ &= \{v \in L^p(\mathbb{R}, X) : v \in AC(\mathbb{R}, X),\ v' \in L^p(\mathbb{R}, X), \\ &\qquad v(s) \in \operatorname{Dom}(\mathcal{A}) \text{ for almost every } s, \text{ and } -v' + Av \in L^p(\mathbb{R}, X)\}.\end{aligned}$$

The important properties of this "evolution semigroup" are summarized in the following remarks; see [23] and also further developments in [29, 34, 43]. The unit circle in $\mathbb{C}$ is denoted here by $\mathbb{T} = \{z \in \mathbb{C} : |z| = 1\}$.

Remarks 2.4. The spectrum $\sigma(E^t_{\mathbb{R}})$ for $t > 0$ is invariant with respect to rotations centered at the origin, and $\sigma(\Gamma_{\mathbb{R}})$ is invariant with respect to translations along $i\mathbb{R}$. Moreover, the following are equivalent.

(i) $\sigma(e^{tA}) \cap \mathbb{T} = \emptyset$ on X.

(ii) $\sigma(E^t_{\mathbb{R}}) \cap \mathbb{T} = \emptyset$ on $\mathcal{F}_{\mathbb{R}}$.

(iii) $0 \in \rho(\Gamma_{\mathbb{R}})$ on $\mathcal{F}_{\mathbb{R}}$. As a consequence,

$$\sigma(E^t_{\mathbb{R}}) \setminus \{0\} = e^{t\sigma(\Gamma_{\mathbb{R}})}, \quad t > 0. \tag{2.3}$$

Note that $\{E^t_{\mathbb{R}}\}_{t\geq 0}$ has the spectral mapping property (2.3) on $\mathcal{F}_{\mathbb{R}}$ even if the underlying semigroup $\{e^{tA}\}_{t\geq 0}$ does not have the spectral mapping property on X. In the latter case, it may be that the exponential stability of the solutions to $\dot{x} = Ax$ on $\mathbb{R}$ are not determined by the spectrum of A. However, such stability *is determined* by the spectrum of $\Gamma_{\mathbb{R}}$. This is made explicit by the following corollary of Remarks 2.4. The spectral bound $s(\Gamma_{\mathbb{R}})$ and the growth bound $\omega_0(E^t_{\mathbb{R}})$ for the evolution semigroup coincide and are equal to the growth bound of $\{e^{tA}\}_{t\geq 0}$:

$$s(\Gamma_{\mathbb{R}}) = \omega_0(E^t_{\mathbb{R}}) = \omega_0(e^{tA}).$$

One of the difficulties related to nonautonomous problems is that their associated evolution families are two-parameter families of operators. From this point of view, it would be of interest to define a *one-parameter* semigroup that is associated to the solutions of the nonautonomous Cauchy problem (2.1). For such a semigroup to be useful, its properties should be closely connected to the asymptotic behavior of

the original nonautonomous problem. Ideally, this semigroup would have a generator that plays the same significant role in determining the stability of the solutions as the operator A played in Lyapunov's classical stability theorem for finite-dimensional autonomous systems $\dot{x} = Ax$. This can, in fact, be done, and the operator of interest is the generator of the following *evolution semigroup* that is induced by the two-parameter evolution family: if $\mathcal{U} = \{U(t,\tau)\}_{t\geq\tau}$ is an evolution family, define operators $E^t_{\mathbb{R}}$, $t \geq 0$, on $\mathcal{F}_{\mathbb{R}} = L^p(\mathbb{R}, X)$ or $\mathcal{F}_{\mathbb{R}} = C_0(\mathbb{R}, X)$ by

$$(E^t_{\mathbb{R}} f)(\tau) = U(\tau, \tau - t) f(\tau - t), \quad \tau \in \mathbb{R}, \quad t \geq 0. \tag{2.4}$$

When $\mathcal{U}$ is exponentially bounded, this defines a strongly continuous evolution semigroup on $\mathcal{F}_{\mathbb{R}}$ whose generator will be denoted by $\Gamma_{\mathbb{R}}$. As shown in [23] and [34], the spectral mapping theorem (2.3) holds for this semigroup. Moreover, the existence of an exponential dichotomy for solutions to $\dot{x}(t) = A(t)x(t)$, $t \in \mathbb{R}$, is characterized by the condition that $\Gamma_{\mathbb{R}}$ is invertible on $\mathcal{F}_{\mathbb{R}}$. Note that in the autonomous case where $U(t,\tau) = e^{(t-\tau)A}$, this is the evolution semigroup defined in (2.2). In the nonautonomous case, the construction of an evolution semigroup is a way to "autonomize" a time-varying Cauchy problem by replacing the time-dependent differential equation $\dot{x} = A(t)x$ on X by an autonomous differential equation $\dot{f} = \Gamma f$ on a superspace of X-valued functions.

3. Evolution semigroups and Cauchy problems. In order to tackle the problem of characterizing the exponential stability of solutions to the nonautonomous Cauchy problem (2.1) on the half-line $\mathbb{R}_+$, the following variant of the above evolution semigroup is needed. As before, let $\{U(t,\tau)\}_{t\geq\tau}$ be an exponentially bounded evolution family, and define operators E^t, $t \geq 0$, on functions $f : \mathbb{R}_+ \to X$ by

$$(E^t f)(\tau) = \begin{cases} U(\tau, \tau - t) f(\tau - t), & 0 \leq t \leq \tau, \\ 0, & 0 \leq \tau < t. \end{cases} \tag{3.1}$$

This defines a strongly continuous semigroup of operators on the space of functions $\mathcal{F} = L^p(\mathbb{R}_+, X)$, and the generator of this *evolution semigroup* will be denoted by Γ. This also defines a strongly continuous semigroup on $C_{00}(\mathbb{R}_+, X) = \{f \in C_0(\mathbb{R}_+, X) : f(0) = 0\}$. For more information on evolution semigroups on the half-line, see also [27, 28, 29, 43].

3.1. Stability. The primary goal of this subsection is to identify the useful properties of the semigroup of operators defined in (3.1) which will be used in the subsequent sections. In particular, the following spectral mapping theorem will allow this semigroup to be used in characterizing the exponential stability of solutions to (2.1) on $\mathbb{R}_+$. See also [34, 43] for different proofs. The spectral symmetry portion of this theorem is due to Rau [38].

THEOREM 3.1. *Let $\mathcal{F}$ denote $C_{00}(\mathbb{R}_+, X)$ or $L^p(\mathbb{R}_+, X)$. The spectrum $\sigma(\Gamma)$ is a half-plane, the spectrum $\sigma(E^t)$ is a disk centered at the origin, and*

$$e^{t\sigma(\Gamma)} = \sigma(E^t) \setminus \{0\}, \quad t > 0. \tag{3.2}$$

Proof. The arguments for the two cases $\mathcal{F} = C_{00}(\mathbb{R}_+, X)$ and $\mathcal{F} = L^p(\mathbb{R}_+, X)$ are similar, so only the first one is considered here.

We first note that $\sigma(\Gamma)$ is invariant under translations along $i\mathbb{R}$, and $\sigma(E^t)$ is invariant under rotations about zero. This spectral symmetry is a consequence of the

fact that for $\xi \in \mathbb{R}$,

$$E^t e^{i\xi\cdot} f = e^{i\xi\cdot} e^{-i\xi t} E^t f, \quad \text{and} \quad \Gamma e^{i\xi\cdot} = e^{i\xi\cdot}(\Gamma - i\xi). \tag{3.3}$$

The inclusion $e^{t\sigma(\Gamma)} \subseteq \sigma(E^t) \setminus \{0\}$ follows from the standard spectral inclusion for strongly continuous semigroups [2]. In view of the spectral symmetry, it suffices to show that $\sigma(E^t) \cap \mathbb{T} = \emptyset$ whenever $0 \in \rho(\Gamma)$. To this end, we replace the Banach space X in Remarks 2.4 by $C_{00}(\mathbb{R}_+, X)$ and consider two semigroups $\{\tilde{E}^t\}_{t\geq 0}$ and $\{\mathcal{E}^t\}_{t\geq 0}$ with generators $\tilde{\Gamma}$ and $\mathcal{G}$, respectively, acting on the space $C_0(\mathbb{R}, C_{00}(\mathbb{R}_+, X))$. These semigroups are defined by

$$(\tilde{E}^t h)(\tau,\theta) = \begin{cases} U(\theta, \theta - t)h(\tau - t, \theta - t) & \text{for} \quad \theta \geq t, \\ 0 & \text{for} \quad 0 \leq \theta < t, \end{cases}$$

$$(\mathcal{E}^t h)(\tau,\theta) = \begin{cases} U(\theta, \theta - t)h(\tau, \theta - t) & \text{for} \quad \theta \geq t, \\ 0 & \text{for} \quad 0 \leq \theta \leq t, \end{cases}$$

where $\tau \in \mathbb{R}$ and $h(\tau, \cdot) \in C_{00}(\mathbb{R}_+, X)$. Note that if $H \in C_0(\mathbb{R}, C_{00}(\mathbb{R}_+, X))$, then $h(\tau, \cdot) := H(\tau) \in C_{00}(\mathbb{R}_+, X)$ and we recognize $\{\tilde{E}^t\}_{t\geq 0}$ as the evolution semigroup induced by $\{E^t\}_{t\geq 0}$, as in (2.2):

$$(\tilde{E}^t H)(\tau) = E^t H(\tau - t).$$

Also, the semigroup $\{\mathcal{E}^t\}_{t\geq 0}$ is the family of multiplication operators given by

$$(\mathcal{E}^t H)(\tau) = E^t H(\tau).$$

The generator $\mathcal{G}$ of this semigroup is the operator of multiplication by Γ: $(\mathcal{G}H)(\tau) = \Gamma(H(\tau))$, where $H(\tau) \in \mathrm{Dom}(\Gamma)$ for $\tau \in \mathbb{R}$. In particular, if $0 \in \rho(\Gamma)$ on $\mathcal{F}$, then $(\mathcal{G}^{-1}H)(\tau) = \Gamma^{-1}(H(\tau))$, and so $0 \in \rho(\mathcal{G})$.

Let J denote the isometry on $C_0(\mathbb{R}, C_{00}(\mathbb{R}_+, X))$ given by $(Jh)(\tau,\theta) = h(\tau+\theta, \theta)$ for $\tau \in \mathbb{R}$, $\theta \in \mathbb{R}_+$. Then J satisfies the identity

$$(\mathcal{E}^t J h)(\tau,\theta) = (J\tilde{E}^t h)(\tau,\theta), \quad \tau \in \mathbb{R}, \quad \theta \in \mathbb{R}_+.$$

It follows that $\mathcal{G}JH = J\tilde{\Gamma}H$ for $H \in \mathrm{Dom}(\tilde{\Gamma})$, and $J^{-1}\mathcal{G}H = \tilde{\Gamma}J^{-1}H$ for $H \in \mathrm{Dom}(\mathcal{G})$. Consequently, $\sigma(\mathcal{G}) = \sigma(\tilde{\Gamma})$ on $C_0(\mathbb{R}, C_{00}(\mathbb{R}_+, X))$. In particular, $0 \in \rho(\tilde{\Gamma})$. Therefore, $\sigma(E^t) \cap \mathbb{T} = \emptyset$ follows from Remarks 2.4 applied to the semigroup $\{E^t\}_{t\geq 0}$ on $\mathcal{F}$ in place of $\{e^{tA}\}_{t\geq 0}$ on X.

The facts that $\sigma(\Gamma)$ is a half-plane and $\sigma(E^t)$ is a disk follow from the spectral mapping property (3.2) and [38, Prop. 2]. □

An important consequence of this theorem is the property that the growth bound $\omega_0(E^t)$ equals the spectral bound $s(\Gamma)$. This leads to the following simple result on stability.

THEOREM 3.2. *Let $\mathcal{F}$ denote $C_{00}(\mathbb{R}_+, X)$ or $L^p(\mathbb{R}_+, X)$. An exponentially bounded evolution family $\{U(t,\tau)\}_{t\geq\tau}$ is exponentially stable if and only if the growth bound $\omega_0(E^t)$ of the induced evolution semigroup on $\mathcal{F}$ is negative.*

Proof. Let $\mathcal{F} = C_{00}(\mathbb{R}_+, X)$. If $\{U(t,\tau)\}_{t\geq\tau}$ is exponentially stable, then there exist $M > 1$, $\beta > 0$ such that $\|U(t,\tau)\|_{\mathcal{L}(X)} \leq Me^{-\beta(t-\tau)}$, $t \geq \tau$. For $\tau \geq 0$ and $f \in C_{00}(\mathbb{R}_+, X)$,

$$\begin{aligned} \|E^\tau f\|_{C_{00}(\mathbb{R}_+,X)} = \sup_{t>0} \|E^\tau f(t)\|_X &= \sup_{t>\tau} \|U(t, t-\tau) f(t-\tau)\|_X \\ &\leq \sup_{t>\tau} \|U(t, t-\tau)\|_{\mathcal{L}(X)} \|f(t-\tau)\|_X \\ &\leq Me^{-\beta\tau} \|f\|_{C_{00}(\mathbb{R}_+,X)}. \end{aligned}$$

Conversely, assume there exist $M > 1$, $\alpha > 0$ such that $\|E^t\| \le Me^{-\alpha t}$, $t \ge 0$. Let $x \in X$, $\|x\| = 1$. For fixed $t > \tau > 0$, choose $f \in C_{00}(\mathbb{R}_+, X)$ such that $\|f\|_{C_{00}(\mathbb{R}_+,X)} = 1$ and $f(\tau) = x$. Then,

$$\begin{aligned}\|U(t,\tau)x\|_X = \|U(t,\tau)f(\tau)\|_X &= \|E^{(t-\tau)}f(t)\|_X \\ &\le \sup_{\theta>0} \|E^{(t-\tau)}f(\theta)\|_X \\ &= \|E^{(t-\tau)}f\|_{C_{00}(\mathbb{R}_+,X)} \\ &\le Me^{-\alpha(t-\tau)}.\end{aligned}$$

A similar argument works for $\mathcal{F} = L^p(\mathbb{R}_+, X)$. □

The remainder of this subsection focuses on the operator used for determining exponential stability. In fact, stability is characterized by the boundedness of this operator which, as seen below, is equivalent to the invertibility of Γ, the generator of the evolution semigroup. We begin with the autonomous case.

Datko and van Neerven have characterized the exponential stability of solutions for autonomous equations $\dot{x} = Ax$, $t \ge 0$, in terms of a convolution operator $\mathbb{G}$ induced by $\{e^{tA}\}_{t\ge0}$. In this autonomous setting,

$$(E^t f)(\tau) = \begin{cases} e^{tA}f(\tau - t), & 0 \le t \le \tau, \\ 0, & 0 \le \tau < t, \end{cases} \tag{3.4}$$

and the convolution operator takes the following form: for $f \in L^1_{loc}(\mathbb{R}_+, X)$,

$$(\mathbb{G}f)(t) := \int_0^t e^{\tau A}f(t-\tau)\,d\tau = \int_0^\infty (E^\tau f)(t)\,d\tau, \quad t \ge 0. \tag{3.5}$$

For the reader's convenience, we cite Theorem 1.3 of [28] (see also [13]) in the following remarks.

Remarks 3.3. If $\{e^{tA}\}_{t\ge0}$ is a strongly continuous semigroup on X, and $1 \le p < \infty$, then the following are equivalent.

(i) $\omega_0(e^{tA}) < 0$.

(ii) $\mathbb{G}f \in L^p(\mathbb{R}_+, X)$ for all $f \in L^p(\mathbb{R}_+, X)$.

(iii) $\mathbb{G}f \in C_0(\mathbb{R}_+, X)$ for all $f \in C_0(\mathbb{R}_+, X)$.

Remarks 3.4.

(a) Note that condition (ii) is equivalent to the boundedness of $\mathbb{G}$ on $L^p(\mathbb{R}_+, X)$. To see this, it suffices to show that the map $f \mapsto \mathbb{G}f$ is a closed operator on $L^p(\mathbb{R}_+, X)$ and then to apply the closed graph theorem. For this, let $f_n \to f$ and $\mathbb{G}f_n \to g$ in $L^p(\mathbb{R}_+, X)$. Then $(\mathbb{G}f_n)(t) \to (\mathbb{G}f)(t)$ for each $t \in \mathbb{R}$. Also, every norm-convergent sequence in $L^p(\mathbb{R}_+, X)$ contains a subsequence that converges pointwise almost everywhere. Thus $(\mathbb{G}f_{n_k})(t) \to g(t)$ for almost all t. This implies that $\mathbb{G}f = g$, as claimed.

(b) Also, condition (iii) is equivalent to the boundedness of $\mathbb{G}$ on $C_0(\mathbb{R}_+, X)$. This follows from the uniform boundedness principle applied to the operators $\mathbb{G}_t : f \mapsto \int_0^t e^{\tau A}f(t-\tau)\,d\tau$.

We now extend this result so that it may be used to describe exponential stability for a nonautonomous equation. For this define an operator $\mathbb{G}$ in an analogous way: let $\{U(t,\tau)\}_{t\ge\tau}$ be an evolution family and $\{E^t\}_{t\ge0}$ the evolution semigroup in (3.1).

Then define $\mathbb{G}$ for $f \in L^1_{loc}(\mathbb{R}_+, X)$ as

$$(\mathbb{G}f)(t) := \int_0^\infty (E^\tau f)(t)\,d\tau = \int_0^t U(t, t-\tau)f(t-\tau)\,d\tau = \int_0^t U(t,\tau)f(\tau)\,d\tau, \qquad t \geq 0. \tag{3.6}$$

For $\mathbb{G}$ acting on $\mathcal{F} = C_{00}(\mathbb{R}_+, X)$ or $L^p(\mathbb{R}_+, X)$, standard semigroup properties show that $\mathbb{G}$ equals $-\Gamma^{-1}$ provided the semigroup $\{E^t\}_{t\geq 0}$ or the evolution family is uniformly exponentially stable. Parts (i) $\Leftrightarrow$ (ii) of Remarks 3.3 and the nonautonomous version below are the classical results by Datko [13]. Our proof uses the evolution semigroup and creates a formally autonomous problem so that Remarks 3.3 can be applied.

THEOREM 3.5. *The following are equivalent for the evolution family of operators* $\{U(t,\tau)\}_{t\geq\tau}$ *on* X.

(i) $\{U(t,\tau)\}_{t\geq\tau}$ *is exponentially stable.*

(ii) $\mathbb{G}$ *is a bounded operator on* $L^p(\mathbb{R}_+, X)$.

(iii) $\mathbb{G}$ *is a bounded operator on* $C_0(\mathbb{R}_+, X)$.

Before proceeding with the proof, note that statement (ii) is equivalent to the statement: $\mathbb{G}f \in L^p(\mathbb{R}_+, X)$ for each $f \in L^p(\mathbb{R}_+, X)$. This is seen as in Remark 3.4. See also [6] for similar facts.

Proof. By Theorem 3.2, (i) implies that $\{E^t\}_{t\geq 0}$ is exponentially stable, and formula (3.6) implies (ii) and (iii). The implication (ii)$\Rightarrow$(i) will be proved here; the argument for (iii)$\Rightarrow$(i) is similar. The main idea is again to use the "change-of-variables" technique, as in the proof of Theorem 3.1.

Consider the operator $\tilde{\mathbb{G}}$ on $L^p(\mathbb{R}, L^p(\mathbb{R}_+, X)) = L^p(\mathbb{R}\times\mathbb{R}_+, X)$ defined as multiplication by $\mathbb{G}$. More precisely, for $h \in L^p(\mathbb{R}\times\mathbb{R}_+, X)$ with $\mathbf{h}(\theta) := h(\theta, \cdot) \in L^p(\mathbb{R}_+, X)$, define

$$(\tilde{\mathbb{G}}h)(\theta, t) = \mathbb{G}(\mathbf{h}(\theta))(t) = \int_0^t U(t, t-\tau)h(\theta, t-\tau)\,d\tau, \quad t \in \mathbb{R}_+, \quad \theta \in \mathbb{R}.$$

In view of statement (ii), this operator is bounded. For the isometry J defined on the space $L^p(\mathbb{R}, L^p(\mathbb{R}_+, X))$ by $(Jh)(\theta, t) = h(\theta + t, t)$, we have

$$(J^{-1}\tilde{\mathbb{G}}Jh)(\theta, t) = \int_0^t U(t, t-\tau)h(\theta-\tau, t-\tau)\,d\tau. \tag{3.7}$$

Next, let $\{E^t\}_{t\geq 0}$ be the evolution semigroup (3.1) induced by $\{U(t,\tau)\}_{t\geq\tau}$, and define $\mathbb{G}_*$ to be the operator of convolution with this semigroup as in (3.5); that is,

$$(\mathbb{G}_*\mathbf{h})(\theta) = \int_0^\infty E^\tau \mathbf{h}(\theta-\tau)\,d\tau, \quad \mathbf{h} \in L^p(\mathbb{R}, L^p(\mathbb{R}_+, X)). \tag{3.8}$$

If $h(\theta, \cdot) = \mathbf{h}(\theta) \in L^p(\mathbb{R}_+, X)$, then by definition (3.1), evaluating (3.8) at t gives

$$[(\mathbb{G}_*\mathbf{h})(\theta)](t) = (\mathbb{G}_*h)(\theta, t) = \int_0^t U(t, t-\tau)h(\theta-\tau, t-\tau)\,d\tau, \quad t \in \mathbb{R}_+,\ \theta \in \mathbb{R}. \tag{3.9}$$

From (3.7) it follows that $\mathbb{G}_* = J^{-1}\tilde{\mathbb{G}}J$ is a bounded operator on $L^p(\mathbb{R}, L^p(\mathbb{R}_+, X))$.

Now, each function $\mathbf{h}_+ \in L^p(\mathbb{R}_+, L^p(\mathbb{R}_+, X))$ is an $L^p(\mathbb{R}_+, X)$-valued function on the half-line $\mathbb{R}_+$. We extend each such $\mathbf{h}_+$ to a function $\mathbf{h} \in L^p(\mathbb{R}, L^p(\mathbb{R}_+, X))$ by setting $\mathbf{h}(\theta) = \mathbf{h}_+(\theta)$ for $\theta \geq 0$ and $\mathbf{h}(\theta) = 0$ for $\theta < 0$. Note that $\mathbb{G}_*\mathbf{h} \in L^p(\mathbb{R}, L^p(\mathbb{R}_+, X))$ because $\mathbb{G}_*$ is bounded on $L^p(\mathbb{R}, L^p(\mathbb{R}_+, X))$. Consider the function $\mathbf{f}_+ : \mathbb{R}_+ \to L^p(\mathbb{R}_+, X)$ defined by

$$\mathbf{f}_+(t) = \int_0^t E^\tau \mathbf{h}_+(t-\tau)\, d\tau = \int_0^\infty E^\tau \mathbf{h}(t-\tau)\, d\tau, \quad t \in \mathbb{R}_+.$$

To complete the proof of the theorem, it suffices to prove the following claim:

$$\mathbf{f}_+ \in L^p(\mathbb{R}_+, L^p(\mathbb{R}_+, X)).$$

Indeed, the operator $\mathbf{h}_+ \mapsto \mathbf{f}_+$ is the convolution operator as in (3.5) defined by the semigroup operators E^t instead of e^{tA}. An application of Remarks 3.3 to E^t on $L^p(\mathbb{R}_+, X)$ (in place of e^{tA} on X) shows that the semigroup $\{E^t\}_{t\geq 0}$ is exponentially stable on $L^p(\mathbb{R}_+, X)$ provided that

$$\mathbf{f}_+ \in L^p(\mathbb{R}_+, L^p(\mathbb{R}_+, X)) \quad \text{for each } \mathbf{h}_+ \in L^p(\mathbb{R}_+, L^p(\mathbb{R}_+, X)).$$

But if $\{E^t\}_{t\geq 0}$ is exponentially stable, the evolution family $\{U(t,\tau)\}_{t\geq\tau}$ is exponentially stable by Theorem 3.2.

To prove the claim, apply formula (3.9) for $h(\theta, t) = h_+(\theta, t)$, $\theta \geq 0$, and $h(\theta, t) = 0$, $\theta < 0$, $t \in \mathbb{R}_+$, where $\mathbf{h}_+(\theta) = h_+(\theta, \cdot)$. This gives

$$(\mathbb{G}_* h)(\theta, t) = \begin{cases} \displaystyle\int_0^{\min\{\theta,t\}} U(t, t-\tau) h_+(\theta-\tau, t-\tau)\, d\tau & \text{for} \quad \theta \geq 0,\ t \in \mathbb{R}_+, \\ (\mathbb{G}_* h)(\theta, t) = 0 & \text{for} \quad \theta < 0,\ t \in \mathbb{R}_+. \end{cases}$$

Thus the function

$$\theta \mapsto (\mathbb{G}_* h)(\theta, \cdot) = (\mathbb{G}_* \mathbf{h})(\theta) \in L^p(\mathbb{R}_+, X)$$

is in the space $L^p(\mathbb{R}_+, L^p(\mathbb{R}_+, X))$. On the other hand, denoting $f_+(\theta, \cdot) := \mathbf{f}_+(\theta) \in L^p(\mathbb{R}_+, X)$, we have that

$$f_+(\theta, t) = \int_0^{\min\{\theta,t\}} U(t, t-\tau) h_+(\theta-\tau, t-\tau)\, d\tau, \quad \theta, t \in \mathbb{R}_+.$$

Thus $\theta \mapsto f_+(\theta, \cdot) = (\mathbb{G}_* h)(\theta, \cdot)$ is a function in $L^p(\mathbb{R}_+, L^p(\mathbb{R}_+, X))$, and the claim is proved. □

This theorem makes explicit, in the case of the half-line $\mathbb{R}_+$, the relationship between the stability of an evolution family $\{U(t,\tau)\}_{t\geq\tau}$ and the generator Γ of the corresponding evolution semigroup (3.1). Indeed, as shown above, stability is equivalent to the boundedness of $\mathbb{G}$, in which case $\mathbb{G} = -\Gamma^{-1}$. Combining Theorems 3.1, 3.2, and 3.5 yields the following corollary.

COROLLARY 3.6. *Let $\{U(t,\tau)\}_{t\geq\tau}$ be an exponentially bounded evolution family and let Γ denote the generator of the induced evolution semigroup on $L^p(\mathbb{R}_+, X)$, $1 \leq p < \infty$, or $C_{00}(\mathbb{R}_+, X)$. The following are equivalent.*

(i) *$\{U(t,\tau)\}_{t\geq\tau}$ is exponentially stable.*

(ii) *Γ is invertible with $\Gamma^{-1} = -\mathbb{G}$.*

(iii) *$s(\Gamma) < 0$.*

For more information on stability and dichotomy of evolution families on the semiaxis see [27].

3.2. Perturbations and robust stability. This subsection briefly considers perturbations of (2.1) of the form

$$\dot{x}(t) = (A(t) + D(t))x(t), \quad t \geq 0. \tag{3.10}$$

It will not, however, be assumed that (3.10) has a differentiable solution. For example, let $\{e^{tA_0}\}_{t\geq 0}$ be a strongly continuous semigroup generated by A_0, let $A_1(t) \in \mathcal{L}(X)$ for $t \geq 0$, and define $A(t) = A_0 + A_1(t)$. Then even if $t \mapsto A_1(t)$ is continuous, the Cauchy problem (2.1) may not have a differentiable solution for all initial conditions $x(0) = x \in \mathrm{Dom}(A) = \mathrm{Dom}(A_0)$ (see, e.g., [31]). Therefore we will want our development to allow for equations with solutions that exist only in the following mild sense.

Let $\{U(t,\tau)\}_{t\geq\tau}$ be an evolution family of operators corresponding to a solution of (2.1), and consider the nonautonomous inhomogeneous equation

$$\dot{x}(t) = A(t)x(t) + f(t), \qquad t \geq 0, \tag{3.11}$$

where f is a locally integrable X-valued function on $\mathbb{R}_+$. A function $x(\cdot)$ is a mild solution of (3.11) with initial value $x(\theta) = x_\theta \in \mathrm{Dom}(A(\theta))$ if

$$x(t) = U(t,\theta)x_\theta + \int_\theta^t U(t,\tau)f(\tau)\,d\tau, \quad t \geq \theta.$$

Given operators $D(t)$, the existence of mild solutions to an additively perturbed (3.10) corresponds to the existence of an evolution family $\{U_1(t,\tau)\}_{t\geq\tau}$ satisfying

$$U_1(t,\theta)x = U(t,\theta)x + \int_\theta^t U(t,\tau)D(\tau)U_1(\tau,\theta)x\,d\tau \tag{3.12}$$

for all $x \in X$. It will be assumed that the perturbation operators $D(t)$ are strongly measurable and essentially bounded functions of t. In view of this, we use the notation $\mathcal{L}_s(X)$ to denote the set $\mathcal{L}(X)$ endowed with the strong operator topology, and we use $L^\infty(\mathbb{R}_+, \mathcal{L}_s(X))$ to denote the set of bounded, strongly measurable $\mathcal{L}(X)$-valued functions on $\mathbb{R}_+$. A function $D(\cdot) \in L^\infty(\mathbb{R}_+, \mathcal{L}_s(X))$ induces a multiplication operator $\mathcal{D}$ defined by $\mathcal{D}x(t) = D(t)x(t)$ for $x(\cdot) \in L^p(\mathbb{R}_+, X)$. In fact, $\mathcal{D}$ is a bounded operator on $L^p(\mathbb{R}_+, X)$ with $\|\mathcal{D}\| \leq \|D(\cdot)\|_\infty := \operatorname{ess\,sup}_{t\in\mathbb{R}_+} \|D(t)\|$.

Evolution semigroups induced by an evolution family as in (3.1) have been studied by several authors who have characterized such semigroups in terms of their generators on general Banach function spaces of X-valued functions (see [36, 43] and the bibliographies therein). The sets $\mathcal{F} = L^p(\mathbb{R}_+, X)$ or $\mathcal{F} = C_{00}(\mathbb{R}_+, X)$ considered here are examples of more general "Banach function spaces." In the development that follows we use a theorem of Schnaubelt [43] (see also Räbiger et al. [35, 36]) which shows exactly when a strongly continuous semigroup on $\mathcal{F}$ arises from a strongly continuous evolution family on X. We state a version of this result which will be used below; a more general version is proven in [36]. The set $C_c^1(\mathbb{R}_+)$ consists of differentiable functions on $\mathbb{R}_+$ that have compact support.

THEOREM 3.7. *Let $\{T^t\}_{t\geq 0}$ be a strongly continuous semigroup generated by Γ on $\mathcal{F}$. The following are equivalent.*

(i) *$\{T^t\}_{t\geq 0}$ is an evolution semigroup; i.e., there exists an exponentially bounded evolution family so that T^t is defined as in* (3.1).

(ii) *There exists a core, $\mathcal{C}$, of Γ such that for all $\varphi \in C_c^1(\mathbb{R}_+)$ and $f \in \mathcal{C}$, it follows that $\varphi f \in \mathrm{Dom}(\Gamma)$ and $\Gamma(\varphi f) = -\varphi' f + \varphi\Gamma f$. Moreover, there exists $\lambda \in \rho(\Gamma)$ such that $R(\lambda,\Gamma) : \mathcal{F} \to C_{00}(\mathbb{R}_+, X)$ is continuous with dense range.*

Now let $\{U(t,\tau)\}_{t\geq\tau}$ be an evolution family on X, and let Γ be the generator of the corresponding evolution semigroup, $\{E^t\}_{t\geq 0}$, as in (3.1). If $D(\cdot)\in L^\infty(\mathbb{R}_+,\mathcal{L}_s(X))$, then the multiplication operator $\mathcal{D}$ is a bounded operator on $\mathcal{F}=L^p(\mathbb{R}_+,X)$. Since a bounded perturbation of a generator of a strongly continuous semigroup is itself such a generator, the operator $\Gamma_1=\Gamma+\mathcal{D}$ generates a strongly continuous semigroup, $\{E_1^t\}_{t\geq 0}$ on $\mathcal{F}$ (see, e.g., [30]). In fact, Γ_1 generates an *evolution* semigroup; see [35, 43].

PROPOSITION 3.8. *Let $D(\cdot)\in L^\infty(\mathbb{R}_+,\mathcal{L}_s(X))$, and let $\{U(t,\tau)\}_{t\geq\tau}$ be an exponentially bounded evolution family. Then there exists a unique evolution family $\mathcal{U}_1=\{U_1(t,\tau)\}_{t\geq\tau}$ which solves the integral equation* (3.12)*. Moreover, $\mathcal{U}_1$ is exponentially stable if and only if $\Gamma+\mathcal{D}$ is invertible.*

Proof. As already observed, $\Gamma_1=\Gamma+\mathcal{D}$ generates a strongly continuous semigroup, $\{E_1^t\}_{t\geq 0}$ on $\mathcal{F}$. To see that this is in fact an evolution semigroup, note that for $\lambda\in\rho(\Gamma)\cap\rho(\Gamma_1)$,

$$\mathrm{Range}(R(\lambda,\Gamma))=\mathrm{Dom}(\Gamma)=\mathrm{Dom}(\Gamma+\mathcal{D})=\mathrm{Range}(R(\lambda,\Gamma_1))$$

is dense in $C_{00}(\mathbb{R}_+,X)$. Also, if $\mathcal{C}$ is a core for Γ, then it is a core for Γ_1, and so for $\varphi\in C_c^1(\mathbb{R})$, $f\in\mathcal{C}$,

$$\Gamma_1(\varphi f)=\Gamma(\varphi f)+\mathcal{D}(\varphi f)=-\varphi' f+\varphi\Gamma f+\varphi\mathcal{D}f=-\varphi' f+\varphi(\Gamma+\mathcal{D})f.$$

Consequently, Theorem 3.7 shows that $\{E_1^t\}_{t\geq 0}$ corresponds to an evolutionary family, $\{U_1(t,\tau)\}_{t\geq\tau}$. Moreover, $x(t)=U_1(t,\tau)x(\tau)$ is seen to define a mild solution to (3.10). Indeed,

$$E_1^t f=E^t f+\int_0^t E^{(t-\tau)}\mathcal{D}E_1^\tau f\,d\tau \tag{3.13}$$

holds for all $f\in F$. In particular, for $x\in X$ and any $\varphi\in C_c^1(\mathbb{R})$, setting $f=\varphi\otimes x$ in (3.13), where $\varphi\otimes x(t)=\varphi(t)x$, and using a change of variables leads to

$$\varphi(\theta)U_1(t,\theta)x=\varphi(\theta)U(t,\theta)x+\varphi(\theta)\int_\theta^t U(t,\tau)D(\tau)U_1(\tau,\theta)x\,d\tau.$$

Therefore, (3.12) holds for all $x\in X$.

Finally, Corollary 3.6 shows that $\mathcal{U}_1$ is exponentially stable if and only if Γ_1 is invertible. ∎

The existence of mild solutions under bounded perturbations of this type is well known (see, e.g., [10]), but an immediate consequence of the approach given here is the property of robustness for the stability of $\{U(t,\tau)\}_{t\geq\tau}$. Indeed, by continuity properties of the spectrum of an operator Γ, there exists $\epsilon>0$ such that Γ_1 is invertible whenever $\|\Gamma_1-\Gamma\|<\epsilon$; that is, $\{U_1(t,\tau)\}_{t\geq\tau}$ is exponentially stable whenever $\|D(\cdot)\|_\infty<\epsilon$. Also, the type of proof presented here can be extended to address the case of unbounded perturbations. For an example of this, we refer to [36]. Finally, and most important to the present paper, is the fact that this approach provides insight into the concept of the stability radius. This topic is studied next.

4. Stability radius. The goal of this section is to use the previous development to study the (complex) stability radius of an exponentially stable system. Loosely speaking, this is a measurement on the size of the smallest operator under which the additively perturbed system loses exponential stability. This is an important concept

for linear systems theory and was introduced by Hinrichsen and Pritchard as the basis for a state-space approach to studying robustness of linear time-invariant [17] and time-varying systems [16, 18, 32]. A systematic study of various stability radii in the spirit of the current paper has recently be given by Fischer and van Neerven [14].

4.1. General estimates. In this subsection we give estimates for the stability radius of general nonautonomous systems on Banach spaces. The perturbations considered here are additive "structured" perturbations of output feedback type. That is, let U and Y be Banach spaces, and let $\Delta(t) : Y \to U$ denote an unknown disturbance operator. The operators $B(t) : U \to X$ and $C(t) : X \to Y$ describe the structure of the perturbation in the following (formal) sense: if $u(t) = \Delta(t)y(t)$ is viewed as a feedback for the system

$$\begin{aligned} &\dot{x}(t) = A(t)x(t) + B(t)u(t), \qquad x(s) = x_s \in \mathrm{Dom}(A(s)),\\ &y(t) = C(t)x(t), \quad t \geq s \geq 0, \end{aligned} \tag{4.1}$$

then the nominal system $\dot{x}(t) = A(t)x(t)$ is subject to the structured perturbation

$$\dot{x}(t) = (A(t) + B(t)\Delta(t)C(t))x(t), \quad t \geq 0. \tag{4.2}$$

In this section B and C do not represent input and output operators; rather, they describe the structure of the uncertainty of the system. Also, systems considered throughout this paper are not assumed to have differentiable solutions, and so (4.2) is to be interpreted in the mild sense as described in (3.12), where $D(t) = B(t)\Delta(t)C(t)$. Similarly, (4.1) is interpreted in the mild sense; that is, there exists a strongly continuous exponentially bounded evolution family $\{U(t,\tau)\}_{t\geq\tau}$ on a Banach space X which satisfies

$$\begin{aligned} &x(t) = U(t,s)x(s) + \int_s^t U(t,\tau)B(\tau)u(\tau)\,d\tau,\\ &y(t) = C(t)x(t), \qquad t \geq s \geq 0. \end{aligned} \tag{4.3}$$

In the case of time-invariant systems, (4.3) takes the form

$$\begin{aligned} &x(t) = e^{tA}x_0 + \int_0^t e^{(t-\tau)A}Bu(\tau)\,d\tau,\\ &y(t) = Cx(t), \qquad t \geq 0, \end{aligned} \tag{4.4}$$

where $\{e^{tA}\}_{t\geq 0}$ is a strongly continuous semigroup on X generated by A, $x(0) = x_0 \in \mathrm{Dom}(A)$.

It should be emphasized that we will not address questions concerning the existence of solutions for a perturbed system (4.2) beyond the point already discussed in Proposition 3.8. In view of that proposition, we make the following assumptions: B, C, and Δ are strongly measurable and essentially bounded functions of t; i.e., $B(\cdot) \in L^\infty(\mathbb{R}_+, \mathcal{L}_s(U,X))$, $C(\cdot) \in L^\infty(\mathbb{R}_+, \mathcal{L}_s(X,Y))$, and $\Delta(\cdot) \in L^\infty(\mathbb{R}_+, \mathcal{L}_s(Y,U))$. As such, they induce bounded multiplication operators, $\mathcal{B}$, $\mathcal{C}$, and $\tilde{\Delta}$ acting on the spaces $L^p(\mathbb{R}_+, U)$, $L^p(\mathbb{R}_+, X)$, and $L^p(\mathbb{R}_+, Y)$, respectively.

Next, for an exponentially bounded evolution family $\{U(t,\tau)\}_{t\geq\tau}$, define the "input-output" operator $\mathbb{L}$ on functions $u : \mathbb{R}_+ \to U$ by the rule

$$(\mathbb{L}u)(t) = C(t)\int_0^t U(t,\tau)B(\tau)u(\tau)\,d\tau.$$

Using the above notation, note that $\mathbb{L} = \mathcal{C}\mathbb{G}\mathcal{B}$. Much of the stability analysis that follows is based on this observation in combination with Theorem 3.5 which shows that the operator $\mathbb{G}$ completely characterizes stability of the corresponding evolution family.

We now turn to the definition of the stability radius. For this let $\mathcal{U} = \{U(t,\tau)\}_{t\geq\tau}$ be an exponentially stable evolution family on X. Set $\mathcal{D} = \mathcal{B}\tilde{\Delta}\mathcal{C}$, and let $\mathcal{U}_\Delta = \{U_\Delta(t,\tau)\}_{t\geq\tau}$ denote the evolution family corresponding to solutions of the perturbed equation (3.10). That is, $\mathcal{U}_\Delta$ satisfies

$$U_\Delta(t,s)x = U(t,s)x + \int_s^t U(t,\tau)B(\tau)\Delta(\tau)C(\tau)U_\Delta(\tau,s)x\,d\tau, \qquad x \in X.$$

Define the (complex) *stability radius* for $\mathcal{U}$ with respect to the perturbation structure $(B(\cdot), C(\cdot))$ as the quantity

$$r_{stab}(\mathcal{U}, B, C) = \sup\{r \geq 0 : \|\Delta(\cdot)\|_\infty \leq r \Rightarrow \mathcal{U}_\Delta \text{ is exponentially stable}\}.$$

This definition applies to both nonautonomous and autonomous systems, though in the latter case the notation $r_{stab}(\{e^{tA}\}, B, C))$ will be used to distinguish the case where all the operators except $\Delta(t)$ are independent of t. We will have occasion to consider the *constant stability radius* which is defined for the case in which $\Delta(t) \equiv \Delta$ is constant; this will be denoted by $rc_{stab}(\{e^{tA}\}, B, C))$ or $rc_{stab}(\mathcal{U}, B, C))$, depending on the context. The above remarks concerning $\Gamma + \mathcal{D}$ (see Proposition 3.8), when combined with Theorem 3.2, make it clear that

$$r_{stab}(\mathcal{U}, B, C) = \sup\{r \geq 0 : \|\Delta(\cdot)\|_\infty \leq r \Rightarrow \Gamma + \mathcal{B}\tilde{\Delta}\mathcal{C} \text{ is invertible}\}. \tag{4.5}$$

It is well known that for autonomous systems in which U and Y are Hilbert spaces and $p = 2$, the stability radius may be expressed in terms of the norm of the input-output operator or the transfer function

$$\frac{1}{\|\mathbb{L}\|_{\mathcal{L}(L^2)}} = r_{stab}(\{e^{tA}\}, B, C) = \frac{1}{\sup_{s\in\mathbb{R}} \|C(A - is)^{-1}B\|}; \tag{4.6}$$

see, e.g., [18, Thm. 3.5]. For nonautonomous equations, a scalar example given in Example 4.4 of [16] shows that, in general, a strict inequality $1/\|\mathbb{L}\| < r_{stab}(\mathcal{U}, B, C)$ may hold. Moreover, even for autonomous systems, when Banach spaces are allowed or when $p \neq 2$, Example 4.13 and Example 4.15 below will show that neither of the equalities in (4.6) necessarily hold. Subsection 4.3 focuses on autonomous equations, and a primary objective there is to prove the following result.

THEOREM 4.1. *For the general autonomous systems,*

$$\frac{1}{\|\mathbb{L}\|_{\mathcal{L}(L^p)}} \leq r_{stab}(\{e^{tA}\}, B, C) \leq \frac{1}{\sup_{s\in\mathbb{R}} \|C(A - is)^{-1}B\|}, \quad 1 \leq p < \infty. \tag{4.7}$$

As seen next, the lower bound here holds for general nonautonomous systems and may be proven in a very direct way using the make-up of the operator $\mathbb{L} = \mathcal{C}\mathbb{G}\mathcal{B}$. This lower bound is also proven in [18, Thm. 3.2] using a completely different approach.

THEOREM 4.2. *Assume $\mathcal{U}$ is an exponentially stable evolution family, and let Γ denote the generator of the corresponding evolution semigroup. If*

$$B(\cdot) \in L^\infty(\mathbb{R}_+, \mathcal{L}_s(U, X)) \quad \textit{and} \quad C(\cdot) \in L^\infty(\mathbb{R}_+, \mathcal{L}_s(X, Y)),$$

then $\mathbb{L}$ is a bounded operator from $L^p(\mathbb{R}_+, U)$ to $L^p(\mathbb{R}_+, Y)$, $1 \le p < \infty$, the formula

$$\mathbb{L} = \mathcal{C}\mathbb{G}\mathcal{B} = -\mathcal{C}\Gamma^{-1}\mathcal{B}$$

holds, and

$$\frac{1}{\|\mathbb{L}\|} \le r_{stab}(\mathcal{U}, B, C). \tag{4.8}$$

In the "unstructured" case, where $U = Y = X$ and $B = C = I$, one has

$$\mathbb{L} = -\Gamma^{-1}, \quad \text{and} \quad \frac{1}{\|\Gamma^{-1}\|} \le r_{stab}(\mathcal{U}, I, I) \le \frac{1}{r(\Gamma^{-1})},$$

where $r(\cdot)$ denotes the spectral radius.

Proof. Since $\mathcal{U}$ is exponentially stable, Γ is invertible and $\Gamma^{-1} = -\mathbb{G}$. The required formula for $\mathbb{L}$ follows from (3.6).

Set $\mathcal{H} := \Gamma^{-1}\mathcal{B}\tilde{\Delta}$. To prove (4.8), let $\Delta(\cdot) \in L^\infty(\mathbb{R}_+, \mathcal{L}_s(Y, U))$ and suppose that $\|\Delta(\cdot)\|_\infty < 1/\|\mathbb{L}\|$. Then $\|\mathbb{L}\tilde{\Delta}\| < 1$, and hence $I - \mathbb{L}\tilde{\Delta} = I + \mathcal{C}\Gamma^{-1}\mathcal{B}\tilde{\Delta}$ is invertible on $L^p(\mathbb{R}_+, Y)$. That is, $I + \mathcal{C}\mathcal{H}$ is invertible on $L^p(\mathbb{R}_+, Y)$, and hence $I + \mathcal{H}\mathcal{C}$ is invertible on $L^p(\mathbb{R}_+, X)$ (with inverse $(I - \mathcal{H}(I + \mathcal{C}\mathcal{H})^{-1}\mathcal{C})$). Now

$$\Gamma + \mathcal{B}\tilde{\Delta}\mathcal{C} = \Gamma(I + \Gamma^{-1}\mathcal{B}\tilde{\Delta}\mathcal{C}) = \Gamma(I + \mathcal{H}\mathcal{C}),$$

and so $\Gamma + \mathcal{B}\tilde{\Delta}\mathcal{C}$ is invertible. It follows from the expression (4.5) that $1/\|\mathbb{L}\| \le r_{stab}(\mathcal{U}, B, C)$.

For the last assertion, suppose that $r_{stab}(\mathcal{U}, I, I) > 1/r(\Gamma^{-1})$. Then there exists λ such that $|\lambda| = r(\Gamma^{-1})$ and $\lambda + \Gamma^{-1}$ is not invertible. But then setting $\tilde{\Delta} \equiv \frac{1}{\lambda}$ gives $\|\tilde{\Delta}\| = \frac{1}{|\lambda|} < r_{stab}(\mathcal{U}, I, I)$, and so $\Gamma + \tilde{\Delta} = \tilde{\Delta}(\lambda + \Gamma^{-1})\Gamma$ is invertible, which is a contradiction. ∎

4.2. The transfer function for nonautonomous systems. In this subsection we consider a time-varying version of (4.6) and then observe that the concept of a transfer function, or frequency-response function, arises naturally from these ideas. For this we assume in this subsection that X, U, and Y are Hilbert spaces and $p = 2$.

Let $\{U(t, \tau)\}_{t \ge \tau}$ be a uniformly exponentially stable evolution family, and let $\{E^t\}_{t \ge 0}$ be the induced evolution semigroup with generator Γ on $L^2(\mathbb{R}_+, X)$. Recall that $\mathcal{B}$ and $\mathcal{C}$ denote multiplication operators, with respective multipliers $B(\cdot)$ and $C(\cdot)$, that act on the spaces $L^2(\mathbb{R}_+, U)$ and $L^2(\mathbb{R}_+, X)$, respectively. Let $\tilde{\mathcal{B}}$ and $\tilde{\mathcal{C}}$ denote operators of multiplication induced by $\mathcal{B}$ and $\mathcal{C}$, respectively; e.g., $(\tilde{\mathcal{B}}\mathbf{u})(t) = \mathcal{B}(\mathbf{u}(t))$ for $\mathbf{u} : \mathbb{R}_+ \to L^2(\mathbb{R}_+, U)$. Now consider the operator $\mathbb{G}_*$ as defined in (3.8) and note that operator $\mathbb{L}_* := \tilde{\mathcal{C}}\mathbb{G}_*\tilde{\mathcal{B}}$ may be viewed (formally) as an input-output operator for the "autonomized" system $\dot{f} = \Gamma f + \mathcal{B}u, \ \ g = \mathcal{C}f,$ where the state space is $L^2(\mathbb{R}_+, X)$. It follows from the known Hilbert-space equalities in (4.6) that

$$\frac{1}{\|\mathbb{L}_*\|} = r_{stab}(\{E^t\}, \mathcal{B}, \mathcal{C}) = \frac{1}{\sup_{s \in \mathbb{R}} \|\mathcal{C}(\Gamma - is)^{-1}\mathcal{B}\|}.$$

Note, however, that the rescaling identities (3.3) for Γ imply that

$$\|\mathbb{L}_*\| = \|\mathcal{C}(\Gamma - is)^{-1}\mathcal{B}\| = \|\mathcal{C}\Gamma^{-1}\mathcal{B}\| = \|\mathbb{L}\|,$$

and so the stability radius for the evolution semigroup is also $1/\|\mathbb{L}\|$. In view of the above-mentioned nonautonomous scalar example for which $1/\|\mathbb{L}\| < r_{stab}(\mathcal{U}, B, C)$, we

see that even though the evolution semigroup (or its generator) completely determines the exponential stability of a system, it does not provide a formula for the stability radius.

However, the operator $\mathcal{C}(\Gamma - is)^{-1}\mathcal{B}$ appearing above suggests that the transfer function for time-varying systems arises naturally when viewed in the context of evolution semigroups. Several authors have considered the concept of a transfer function for nonautonomous systems, but the work of Ball, Gohberg, and Kaashoek [4] seems to be the most comprehensive in providing a system-theoretic input-output interpretation for the value of such a transfer function at a point. Their interpretation justifies the term *frequency-response* function for time-varying finite-dimensional systems with "time-varying complex exponential inputs." Our remarks concerning the frequency response for time-varying infinite-dimensional systems will be restricted to inputs of the form $u(t) = u_0 e^{\lambda t}$.

For motivation, consider the input-output operator $\mathbb{L}$ associated with an autonomous system (4.4) where the nominal system is exponentially stable. The *transfer function* of $\mathbb{L}$ is the unique bounded analytic $\mathcal{L}(U, Y)$-valued function H, defined on $\mathbb{C}_+ = \{\lambda \in \mathbb{C} : \operatorname{Re} \lambda > 0\}$ such that for any $u \in L^2(\mathbb{R}_+, U)$,

$$(\widehat{\mathbb{L}u})(\lambda) = H(\lambda)\hat{u}(\lambda), \qquad \lambda \in \mathbb{C}_+,$$

where $\widehat{\ }$ denotes the Laplace transform (see, e.g., [44]). In this autonomous setting, A generates a uniformly exponentially stable strongly continuous semigroup, and $\mathbb{L} = \mathcal{C}\mathbb{G}\mathcal{B}$, where $\mathbb{G}$ is the operator of convolution with the semigroup operators e^{tA} (see (3.5)). Standard arguments show that $(\widehat{\mathbb{L}u})(\lambda) = C(\lambda - A)^{-1}B\hat{u}(\lambda)$; that is, $H(\lambda) = C(\lambda - A)^{-1}B$.

Now let $\mathbb{L}$ be the input-output operator for the nonautonomous system (4.3). We wish to identify the transfer function of $\mathbb{L}$ as the Laplace transform of the appropriate operator. We are guided by the fact that, just as $(\lambda - A)^{-1}$ may be expressed as the Laplace transform of the semigroup generated by A, the operator $(\lambda - \Gamma)^{-1}$ is the Laplace transform of the evolution semigroup. For nonautonomous systems, $\mathbb{L}$ is again given by $\mathcal{C}\mathbb{G}\mathcal{B}$, although now $\mathbb{G}$ from (3.6) is not, generally, a convolution operator. So instead recall the operator $\mathbb{G}_*$ from (3.8) which *is* the operator of convolution with the evolution semigroup $\{E^t\}_{t\geq 0}$. As noted above, the operator $\mathbb{L}_* := \tilde{\mathcal{C}}\mathbb{G}_*\tilde{\mathcal{B}}$ may be viewed as an input-output operator for an autonomous system (where the state space is $L^2(\mathbb{R}_+, X)$). Therefore, the autonomous theory applies directly to show that, for $\mathbf{u} \in L^2(\mathbb{R}_+, L^2(\mathbb{R}_+, U))$,

$$(\widehat{\mathbb{L}_*\mathbf{u}})(\lambda) = \mathcal{C}(\lambda - \Gamma)^{-1}\mathcal{B}\hat{\mathbf{u}}(\lambda). \tag{4.9}$$

In other words, the transfer function for $\mathbb{L}_*$ is $\mathcal{C}(\lambda - \Gamma)^{-1}\mathcal{B}$, where

$$\mathcal{C}(\lambda - \Gamma)^{-1}\mathcal{B}u = \mathcal{C}\int_0^\infty e^{-\lambda\tau}E^\tau\mathcal{B}u\,d\tau, \qquad u \in L^2(\mathbb{R}_+, U).$$

Evaluating these expressions at $t \in \mathbb{R}_+$ gives

$$[\mathcal{C}(\lambda - \Gamma)^{-1}\mathcal{B}u](t) = \int_0^t C(t)U(t,\tau)B(\tau)u(\tau)e^{-\lambda(t-\tau)}\,d\tau. \tag{4.10}$$

It is natural to call $\mathcal{C}(\lambda - \Gamma)^{-1}\mathcal{B}$ the transfer function for the nonautonomous system. Moreover, the following remarks show that, by looking at the right-hand side of

(4.10), this gives a natural frequency-response function for nonautonomous systems. To see this, we first consider autonomous systems and note that the definition of the transfer function for an autonomous system can be extended to allow for a class of "Laplace transformable" functions that are in $L^2_{loc}(\mathbb{R}_+, U)$ (see, e.g., [44]). This class includes constant functions of the form $v_0(t) = u_0$, $t \geq 0$, for a given $u_0 \in U$. If a periodic input signal of the form $u(t) = u_0 e^{i\omega t}$, $t \geq 0$, (for some $u_0 \in U$ and $\omega \in \mathbb{R}$) is fed into an autonomous system with initial condition $x(0) = x_0$, then, by definition of the input-output operator, we have

$$(\mathbb{L}u)(t) = C(i\omega - A)^{-1}Bu_0 \cdot e^{i\omega t} - Ce^{tA}x_0, \quad (\mathbb{L}u)(t) = C\int_0^t e^{(t-s)A}Bu(s)\,ds.$$

Thus the output

$$y(t; u(\cdot), x_0) = (\mathbb{L}u)(t) + Ce^{tA}x_0 = C(i\omega - A)^{-1}Bu_0 \cdot e^{i\omega t}$$

has the same frequency as the input. In view of this, the function $C(i\omega - A)^{-1}B$ is sometimes called the frequency-response function. Now recall that the semigroup $\{e^{tA}\}_{t\geq 0}$ is stable and so $\lim_{t\to\infty} \|Ce^{tA}x_0\| = 0$. On the other hand, consider $v(t) = u_0$ and (formally) apply $\mathcal{C}(i\omega - \Gamma)^{-1}\mathcal{B}$ to this v. For $x_0 = (i\omega - A)^{-1}Bu_0$, a calculation based on the Laplace transform formula for the resolvent of the generator (applied to the evolution semigroup $\{E^t\}_{t\geq 0}$) yields the identity

$$\left[\mathcal{C}(i\omega - \Gamma)^{-1}\mathcal{B}u_0\right](t) = C(i\omega - A)^{-1}Bu_0 - Ce^{tA}x_0 \cdot e^{-i\omega t}.$$

Let us consider this expression $[\mathcal{C}(i\omega - \Gamma)^{-1}\mathcal{B}u_0](t)$ in the nonautonomous case. By (4.10), this coincides with the frequency-response function for time-varying systems which is defined in [4, Cor. 3.2] by the formula

$$\int_0^t C(t)U(t,\tau)B(\tau)u_0 e^{i\omega(\tau - t)}\,d\tau.$$

Also, as noted in this reference, the result of our derivation agrees with the Arveson frequency-response function as it appears in [42]. We recover it here explicitly as the Laplace transform of an input-output operator (see (4.9)).

4.3. Autonomous systems. In this subsection we give the proof of (4.7) when X, U, and Y are Banach spaces. In the process, however, we also consider two other "stability radii": a pointwise stability radius and a dichotomy radius.

First, we give a generalization to Banach spaces of Theorem 2.1 (cf. [23]). Here, $\mathcal{F}_{per}$ denotes the Banach space $L^p([0, 2\pi], X)$, $1 \leq p < \infty$. If $\{e^{tA}\}_{t\geq 0}$ is a strongly continuous semigroup on X, $\{E^t_{per}\}_{t\geq 0}$ will denote the evolution semigroup defined on $\mathcal{F}_{per}$ by the rule $E^t_{per}f(s) = e^{tA}f([s - t](\mathrm{mod}\, 2\pi))$; its generator will be denoted by Γ_{per}. The symbol Λ will be used to denote the set of all finite sequences $\{v_k\}_{k=-N}^N$ in X or $\mathcal{D}(A)$, or $\{u_k\}_{k=-N}^N$ in U.

THEOREM 4.3. *Let A generate a C_0 semigroup $\{e^{tA}\}_{t\geq 0}$ on X. Let B and C be as above, and $\Delta \in (Y, U)$. Let $\{e^{t(A+B\Delta C)}\}_{t\geq 0}$ be the strongly continuous semigroup generated by $A + B\Delta C$. Then the following are equivalent.*

(i) $1 \in \rho(e^{2\pi(A+B\Delta C)})$.

(ii) $i\mathbb{Z} \subset \rho(A + B\Delta C)$ *and*

$$\sup_{\{v_k\}\in\Lambda} \frac{\|\sum_k (A - ik + B\Delta C)^{-1}v_k e^{ik(\cdot)}\|_{\mathcal{F}_{per}}}{\|\sum_k v_k e^{ik(\cdot)}\|_{\mathcal{F}_{per}}} < \infty.$$

(iii) $i\mathbb{Z} \subset \rho(A + B\Delta C)$ *and*

$$\inf_{\{v_k\}\in\Lambda} \frac{\|\sum_k (A - ik + B\Delta C)v_k e^{ik(\cdot)}\|_{\mathcal{F}_{per}}}{\|\sum_k v_k e^{ik(\cdot)}\|_{\mathcal{F}_{per}}} > 0.$$

Further, if Γ_{per} *denotes the generator of the evolution semigroup on* $\mathcal{F}_{per}$*, as above, and if* $1 \in \rho(e^{2\pi A})$*, then* Γ_{per} *is invertible and*

$$\|\mathcal{C}\Gamma_{per}^{-1}\mathcal{B}\| = \sup_{\{u_k\}\in\Lambda} \frac{\|\sum_k C(A - ik)^{-1}Bu_k e^{ik(\cdot)}\|_{L^p([0,2\pi],Y)}}{\|\sum_k u_k e^{ik(\cdot)}\|_{L^p([0,2\pi],U)}}, \tag{4.11}$$

where $\mathcal{C}\Gamma_{per}^{-1}\mathcal{B} \in \mathcal{L}(L^p([0,2\pi],U), L^p([0,2\pi],Y))$.

Proof. The equivalence of (i)–(iii) follows as in Theorem 2.3 of [23]. For the last statement, let $\{u_k\}$ be a finite set in U, and consider functions f and g of the form

$$f(s) = \sum_k (A - ik)^{-1}Bu_k e^{iks} \quad \text{and} \quad g(s) = \sum_k Bu_k e^{iks}.$$

Then $f = \Gamma_{per}^{-1}g$. For,

$$\begin{aligned}(\Gamma_{per}f)(s) &= \frac{d}{dt}\Big|_{t=0} e^{tA}f([s-t]\mathrm{mod}2\pi)\\ &= \sum_k [A(A - ik)^{-1}Bu_k e^{iks} - ik(A - ik)^{-1}Bu_k e^{iks}] = g(s).\end{aligned}$$

For functions of the form $h(s) = \sum_k u_k e^{iks}$, where $\{u_k\}_k$ is a finite set in U, we have $\mathcal{C}\Gamma^{-1}\mathcal{B}h = \sum_k C(A - ik)^{-1}Bu_k e^{ik(\cdot)}$. Taking the supremum over all such functions gives

$$\begin{aligned}\|\mathcal{C}\Gamma_{per}^{-1}\mathcal{B}\| &= \sup_h \frac{\|\mathcal{C}\Gamma_{per}^{-1}\mathcal{B}h\|}{\|h\|}\\ &= \sup_{\{u_k\}\in\Lambda} \frac{\|\sum_k C(A - ik)^{-1}Bu_k e^{ik(\cdot)}\|_{L^p([0,2\pi],Y)}}{\|\sum_k u_k e^{ik(\cdot)}\|_{L^p([0,2\pi],U)}}.\end{aligned}$$

□

In view of these facts we introduce a "pointwise" variant of the constant stability radius: for $t_0 > 0$ and $\lambda \in \rho(e^{t_0 A})$, define the *pointwise stability radius*

$$rc_{stab}^{\lambda}(e^{t_0 A}, B, C) := \sup\{r > 0 : \|\Delta\|_{\mathcal{L}(Y,U)} \le r \Rightarrow \lambda \in \rho(e^{t_0(A+B\Delta C)})\}.$$

By rescaling, the study of this quantity can be reduced to the case of $\lambda = 1$ and $t_0 = 2\pi$. Indeed,

$$rc_{stab}^{\lambda}(e^{t_0 A}, B, C) = \frac{2\pi}{t_0} rc_{stab}^{\lambda}(e^{2\pi A'}, B, C) \quad \text{where} \quad A' = \frac{t_0}{2\pi}A.$$

Also, after writing $\lambda = |\lambda|e^{i\theta}$ $(\theta \in \mathbb{R})$, note that

$$rc_{stab}^{\lambda}(e^{2\pi A}, B, C) = rc_{stab}^{1}(e^{2\pi A''}, B, C) \quad \text{for} \quad A'' = A - \frac{1}{2\pi}(\ln|\lambda| + i\theta).$$

Therefore,

$$rc_{stab}^{\lambda}(e^{t_0 A}, B, C) = \frac{2\pi}{t_0} rc_{stab}^{1}(e^{2\pi A'''}, B, C)$$

for

$$A''' = \frac{1}{2\pi}(t_0 A - \ln|\lambda| - i\theta).$$

In the following theorem we estimate $rc^1_{stab}(e^{2\pi A}, B, C)$. The idea for the proof goes back to [18]. See also further developments in [14].

THEOREM 4.4. *Let $\{e^{tA}\}_{t\geq 0}$ be a strongly continuous semigroup generated by A on X, and assume $1 \in \rho(e^{2\pi A})$. Let Γ_{per} denote the generator of the induced evolution semigroup on $\mathcal{F}_{per}$. Let $B \in \mathcal{L}(U, X)$ and $C \in \mathcal{L}(X, Y)$. Then*

$$\frac{1}{\|\mathcal{C}\Gamma^{-1}_{per}\mathcal{B}\|} \leq rc^1_{stab}(e^{2\pi A}, B, C) \leq \frac{1}{\sup_{k\in\mathbb{Z}} \|C(A-ik)^{-1}B\|}. \tag{4.12}$$

If U and Y are Hilbert spaces and $p = 2$, then equalities hold in (4.12).

Proof. The first inequality follows from an argument as in Theorem 4.2. For the second inequality, let $\epsilon > 0$, and choose $\bar{u} \in U$ with $\|\bar{u}\| = 1$ and $k_0 \in \mathbb{Z}$ such that

$$\|C(A - ik_0)^{-1}B\bar{u}\|_Y \geq \sup_{k\in\mathbb{Z}} \|C(A-ik)^{-1}B\| - \epsilon > 0.$$

Using the Hahn–Banach theorem, choose $y^* \in Y^*$ with $\|y^*\| \leq 1$ such that

$$\left\langle y^*, \frac{C(A-ik_0)^{-1}B\bar{u}}{\|C(A-ik_0)^{-1}B\bar{u}\|_Y} \right\rangle = 1.$$

Define $\Delta \in \mathcal{L}(Y, U)$ by

$$\Delta y = -\frac{\langle y^*, y\rangle}{\|C(A-ik_0)^{-1}B\bar{u}\|_Y}\bar{u}, \quad y \in Y.$$

We note that

$$\Delta C(A-ik_0)^{-1}B\bar{u} = -\frac{\langle y^*, C(A-ik_0)^{-1}B\bar{u}\rangle}{\|C(A-ik_0)^{-1}B\bar{u}\|_Y}\bar{u} = -\bar{u}, \tag{4.13}$$

and

$$\|\Delta\| \leq \frac{1}{\|C(A-ik_0)^{-1}B\bar{u}\|_Y} \leq \frac{1}{\sup_{k\in\mathbb{Z}} \|C(A-ik_0)^{-1}B\bar{u}\|_Y - \epsilon}. \tag{4.14}$$

Now set $\bar{v} := (A - ik_0)^{-1}B\bar{u}$ in X. By (4.13), $\Delta C\bar{v} = -\bar{u}$, and so

$$(A - ik_0 + B\Delta C)\bar{v} = (A-ik_0)\bar{v} + B\Delta C\bar{v} = B\bar{u} - B\bar{u} = 0.$$

Therefore,

$$\inf_{\{v_k\}\in\Lambda} \frac{\|\sum_k (A - ik + B\Delta C)v_k e^{ik(\cdot)}\|_{\mathcal{F}_{per}}}{\|\sum_k u_k e^{ik(\cdot)}\|_{\mathcal{F}_{per}}} \leq \frac{\|(A-ik_0+B\Delta C)\bar{v}e^{ik_0(\cdot)}\|_{\mathcal{F}_{per}}}{\|\bar{v}e^{ik_0(\cdot)}\|_{\mathcal{F}_{per}}} = 0.$$

By Theorem 4.3, $1 \notin \rho(e^{2\pi(A+B\Delta C)})$. This shows that $rc^1_{stab}(e^{2\pi A}, B, C) \leq \|\Delta\|$.

To finish the proof, suppose that $rc^1_{stab}(e^{2\pi A}, B, C) > (\sup_{k\in\mathbb{Z}} \|C(A-ik)^{-1}B\|)^{-1}$. Then with $r := (\sup_{k\in\mathbb{Z}} \|C(A-ik)^{-1}B\bar{u}\|_Y - \epsilon)^{-1}$, and $\epsilon > 0$ chosen to be sufficiently small, one has

$$\frac{1}{\sup_{k\in\mathbb{Z}} \|C(A-ik)^{-1}B\bar{u}\|_Y} < r < rc^1_{stab}(e^{2\pi A}, B, C).$$

But then by (4.14), $\|\Delta\| \le r < rc^1_{stab}(e^{2\pi A}, B, C)$, which is a contradiction.

For the last statement of the theorem, note that Parseval's formula applied to (4.11) gives

$$\|\mathcal{C}\Gamma^{-1}_{per}\mathcal{B}\| = \sup_{\{u_k\}\in\Lambda} \frac{\left(\sum_k \|C(A-ik)^{-1}Bu_k\|^2_Y\right)^{1/2}}{\left(\sum_k \|u_k\|^2_U\right)^{1/2}} \le \sup_{k\in\mathbb{Z}} \|C(A-ik)^{-1}B\|. \tag{4.15}$$

Therefore,

$$\frac{1}{\|\mathcal{C}\Gamma^{-1}_{per}\mathcal{B}\|} \ge \frac{1}{\sup_{k\in\mathbb{Z}} \|C(A-ik)^{-1}B\|},$$

and hence equalities hold in (4.12). ∎

Next we consider the following "hyperbolic" variant of the constant stability radius. Recall that a strongly continuous semigroup $\{e^{tA}\}_{t\ge 0}$ on X is called *hyperbolic* if

$$\sigma(e^{tA}) \cap \mathbb{T} = \emptyset, \quad \text{where } \mathbb{T} = \{z \in \mathbb{C} : |z| = 1\}$$

for some (and, hence, for all) $t > 0$ (see, e.g., [29]). The hyperbolic semigroups are those for which the differential equation $\dot{x} = Ax$ has exponential dichotomy (see, e.g., [12]) with the dichotomy projection P being the Riesz projection corresponding to the part of spectrum of e^A that lies in the open unit disc.

For a given hyperbolic semigroup $\{e^{tA}\}_{t\ge 0}$ and operators B, C we define the *constant dichotomy radius* as

$$\begin{aligned} rc_{dich}(\{e^{tA}\}, B, C) := \sup\{r \ge 0 : \|\Delta\|_{\mathcal{L}(Y,U)} \le r \text{ implies} \\ \sigma(e^{t(A+B\Delta C)}) \cap \mathbb{T} = \emptyset \text{ for all } t > 0\}. \end{aligned}$$

The dichotomy radius measures the size of the smallest $\Delta \in \mathcal{L}(Y, U)$ for which the perturbed equation $\dot{x} = [A + B\Delta C]x$ loses the exponential dichotomy.

Now for any $\xi \in [0, 1]$, consider the rescaled semigroup generated by $A_\xi := A - i\xi$ consisting of operators $e^{tA_\xi} = e^{-i\xi t}e^{tA}$, $t \ge 0$. The pointwise stability radius can be related to the dichotomy radius as follows.

LEMMA 4.5. *Let $\{e^{tA}\}_{t\ge 0}$ be a hyperbolic semigroup. Then*

$$rc_{dich}(\{e^{tA}\}, B, C) = \inf_{\xi\in[0,1]} rc^1_{stab}(e^{2\pi A_\xi}, B, C).$$

Proof. Denote the left-hand side by α and the right-hand side by β. First fix $r < \beta$. Let $\xi \in [0, 1]$. If $\|\Delta\| \le r$, then $1 \in \rho(e^{2\pi(A_\xi + B\Delta C)})$ and so $e^{i\xi 2\pi} \in \rho(e^{2\pi(A+B\Delta C)})$ for all $\xi \in [0, 1]$. That is, $e^{is} \in \rho(e^{2\pi(A+B\Delta C)})$ for all $s \in \mathbb{R}$, and so $\sigma(e^{2\pi(A+B\Delta C)}) \cap \mathbb{T} = \emptyset$. This shows that $r \le \alpha$, and so $\beta \le \alpha$.

Now suppose $r < \alpha$. If $\|\Delta\| \le r$, then $\sigma(\{e^{t(A+B\Delta C)}\}) \cap \mathbb{T} = \emptyset$, and so $e^{i\xi t} \in \rho(e^{t(A+B\Delta C)})$ for all $\xi \in [0, 1], t \in \mathbb{R}$. That is, $1 \in \rho(e^{t(A_\xi + B\Delta C)})$. This says $r \le \beta$ and so $\alpha \le \beta$. ∎

Under the additional assumption that the semigroup $\{e^{tA}\}_{t\ge 0}$ is exponentially stable (that is, hyperbolic with a trivial dichotomy projection $P = I$), Lemma 4.5 gives, in fact, a formula for the constant *stability* radius. Indeed, the following simple proposition holds.

PROPOSITION 4.6. *Let $\{e^{tA}\}_{t\ge 0}$ be an exponentially stable semigroup. Then*

$$rc_{dich}(\{e^{tA}\}, B, C) = rc_{stab}(\{e^{tA}\}, B, C).$$

Proof. Denote the left-hand side by α and the right-hand side by β. Take $r < \beta$ and any Δ with $\|\Delta\| \le r$. By definition of the constant stability radius, $\omega_0(\{e^{t(A+B\Delta C)}\}) < 0$. In particular, $\sigma(e^{t(A+B\Delta C)}) \cap \mathbb{T} = \emptyset$, and $r \le \alpha$ shows that $\beta \le \alpha$.

Suppose that $\beta < r < \alpha$ for some r. By the definition of the stability radius β, there exists a Δ with $\|\Delta\| \in (\beta, r)$ such that the semigroup $\{e^{t(A+B\Delta C)}\}_{t\ge 0}$ is not stable.

For any $\tau \in [0,1]$ one has $\|\tau\Delta\| \le r < \alpha$. By the definition of the dichotomy radius α it follows that the semigroup $\{e^{t(A+\tau B\Delta C)}\}_{t\ge 0}$ is hyperbolic for each $\tau \in [0,1]$. Now consider its dichotomy projection

$$P(\tau) = (2\pi i)^{-1} \int_{\mathbb{T}} \left(\lambda - e^{A+\tau B\Delta C}\right)^{-1} d\lambda,$$

which is the Riesz projection corresponding to the part of $\sigma(e^{A+\tau B\Delta C})$ located inside of the open unit disk. The function $\tau \mapsto P(\tau)$ is norm continuous. Indeed, since the bounded perturbation $\tau B\Delta C$ of the generator A is continuous in τ, the operators $e^{t(A+\tau B\Delta C)}$, $t \ge 0$, depend on τ continuously (see, e.g., [30, Cor. 3.1.3]); this implies the continuity of $P(\cdot)$ (see, e.g., [12, Thm. I.2.2]).

By assumption $\{e^{tA}\}_{t\ge 0}$ is exponentially stable, so $P(0) = I$. Also, $P(1) \ne I$ since the semigroup $\{e^{t(A+B\Delta C)}\}_{t\ge 0}$ with $\|\Delta\| \le r < \alpha$ is hyperbolic but not stable. Since either $\|I - P(\tau)\| = 0$ or $\|I - P(\tau)\| \ge 1$, this contradicts the continuity of $\|P(\cdot)\|$. ∎

A review of the above development shows that the inequality claimed in (4.7) of Theorem 4.1 can now be proved.

Proof of Theorem 4.1. Indeed, $r_{stab}(\{e^{tA}\}, B, C) \le rc_{stab}(\{e^{tA}\}, B, C)$, and so

$$\begin{aligned}
\frac{1}{\|\mathbb{L}\|} \le r_{stab}(\{e^{tA}\}, B, C) &\le rc_{stab}(\{e^{tA}\}, B, C) && \text{(Theorem 4.2)}\\
&\le rc_{dich}(\{e^{tA}\}, B, C) && \text{(Proposition 4.6)}\\
&\le \inf_{\xi\in[0,1]} rc^1_{stab}(e^{2\pi A_\xi}, B, C) && \text{(Lemma 4.5)}\\
&\le \inf_{\xi\in[0,1]} \frac{1}{\sup_{k\in\mathbb{Z}} \|C(A_\xi - ik)^{-1}B\|} && \text{(Theorem 4.4)}\\
&= \frac{1}{\sup_{s\in\mathbb{R}} \|C(A - is)^{-1}B\|}. && \blacksquare
\end{aligned}$$

We will need below the following simple corollary that holds for *bounded* generators A. (In fact, as shown in [14, Cor. 2.5], formula (4.16) below holds provided A generates a semigroup $\{e^{tA}\}_{t\ge 0}$ that is *uniformly* continuous just for $t > 0$.)

COROLLARY 4.7. *Assume $A \in \mathcal{L}(X)$ generates a (uniformly continuous) stable semigroup on a Banach space X. Then*

$$rc_{stab}(\{e^{tA}\}, B, C) = \frac{1}{\sup_{s\in\mathbb{R}} \|C(A - is)^{-1}B\|}. \tag{4.16}$$

Proof. By Theorem 4.1, it remains to prove only the inequality "$\ge$." Fix Δ with $\|\Delta\|$ strictly less than the right-hand side of (4.16). Since $A + B\Delta C \in \mathcal{L}(X)$, it suffices to show that $A + B\Delta C - \lambda = (A - \lambda)(I + (A - \lambda)^{-1}B\Delta C)$ is invertible for

each λ with $\operatorname{Re}\lambda \geq 0$. By the analyticity of resolvent, $\sup_{\operatorname{Re}\lambda\geq 0}\|C(A-\lambda)^{-1}B\| \leq \sup_{s\in\mathbb{R}}\|C(A-is)^{-1}B\|$. Thus

$$\begin{aligned}\|\Delta\| &< \frac{1}{\sup_{s\in\mathbb{R}}\|C(A-is)^{-1}B\|}\\ &\leq \frac{1}{\sup_{\operatorname{Re}\lambda\geq 0}\|C(A-\lambda)^{-1}B\|} \leq \frac{1}{\|C(A-\lambda)^{-1}B\|}, \quad \operatorname{Re}\lambda \geq 0,\end{aligned}$$

implies that $I+C(A-\lambda)^{-1}B\Delta$ is invertible. Therefore (cf. the proof of Theorem 4.2), $I+(A-\lambda)^{-1}B\Delta C$ is invertible. ∎

4.4. The norm of the input-output operator. Since the lower bound on the stability radius is given by the norm of the input-output operator, which is defined by way of the solution operators, it is of interest to express this quantity in terms of the operators A, B, and C. In this subsection it is shown that for autonomous systems this quantity can, in fact, be expressed explicitly in terms of the transfer function

$$\|\mathbb{L}\| = \sup_{u\in\mathcal{S}(\mathbb{R},U)} \frac{\|\int_{\mathbb{R}} C(A-is)^{-1}Bu(s)e^{is(\cdot)}\,ds\|_{L^p(\mathbb{R},Y)}}{\|\int_{\mathbb{R}} u(s)e^{is(\cdot)}\,ds\|_{L^p(\mathbb{R},U)}}. \tag{4.17}$$

Here we use $\mathcal{S}(\mathbb{R},X)$ to denote the Schwartz class of rapidly decreasing X-valued functions defined on $\mathbb{R}$: $\{v:\mathbb{R}\to X \,\big|\, \sup_{s\in\mathbb{R}}\|s^m v^{(n)}(s)\|<\infty; n,m\in\mathbb{N}\}$. As noted in (4.6), $\|\mathbb{L}\|$ equals $\sup_{s\in\mathbb{R}}\|C(A-is)^{-1}B\|$ if U and Y are Hilbert spaces and $p=2$. The section concludes by providing a similar expression, involving sums, which serves as a lower bound for the constant stability radius.

The current focus is on autonomous systems, so let $\{e^{tA}\}_{t\geq 0}$ be a strongly continuous semigroup generated by A and consider the evolution semigroups $\{E^t_{\mathbb{R}}\}_{t\geq 0}$ defined on functions on the entire real line as in (2.2), and $\{E^t\}_{t\geq 0}$ defined for functions on the half-line as in (3.4). As before, $\Gamma_{\mathbb{R}}$ and Γ will denote the generators of these semigroups on $L^p(\mathbb{R},X)$ and $L^p(\mathbb{R}_+,X)$, respectively. Both semigroups will be used as we first show that $\|\mathcal{C}\Gamma_{\mathbb{R}}^{-1}\mathcal{B}\|$ equals the expression in (4.17) and then check that $\|\mathbb{L}\| \equiv \|\mathcal{C}\Gamma^{-1}\mathcal{B}\| = \|\mathcal{C}\Gamma_{\mathbb{R}}^{-1}\mathcal{B}\|$.

Given $v\in\mathcal{S}(\mathbb{R},X)$, let g_v denote the function

$$g_v(\tau) = \frac{1}{2\pi}\int_{\mathbb{R}} v(s)e^{i\tau s}\,ds, \quad \tau\in\mathbb{R},$$

and set $\mathfrak{G} = \{g_v : v\in\mathcal{S}(\mathbb{R},X)\}$. Assuming $\sup_{s\in\mathbb{R}}\|(A-is)^{-1}\|<\infty$, define, for a given $v\in\mathcal{S}(\mathbb{R},X)$, the function

$$f_v(\tau) = \frac{1}{2\pi}\int_{\mathbb{R}} (A-is)^{-1}v(s)e^{i\tau s}\,ds, \quad \tau\in\mathbb{R},$$

and set $\mathfrak{F} = \{f_v : v\in\mathcal{S}(\mathbb{R},X)\}$.

PROPOSITION 4.8. *Assume* $\sup_{s\in\mathbb{R}}\|(A-is)^{-1}\|<\infty$. *Then*

(i) $\mathfrak{G}$ *consists of differentiable functions, and is dense in* $L^p(\mathbb{R},X)$;

(ii) $\mathfrak{F}$ *is dense in* $\operatorname{Dom}(\Gamma_{\mathbb{R}})$;

(iii) *if* $v\in\mathcal{S}(\mathbb{R},X)$, *then* $\Gamma_{\mathbb{R}}f_v = g_v$.

Proof. For $g\in L^1(\mathbb{R},X)$, denote the Fourier transform by

$$\hat{g}(\tau) = \frac{1}{2\pi}\int_{\mathbb{R}} e^{-is\tau}g(s)\,ds.$$

Note that $\mathfrak{G} = \{g : \mathbb{R} \to X : \exists\ v \in \mathcal{S}(\mathbb{R}, X) \text{ so that } \hat{g} = v\}$, and so $\mathfrak{G}$ contains the set $\{g \in L^1(\mathbb{R}, X) : \hat{g} \in \mathcal{S}(\mathbb{R}, X)\}$. Since the latter set is dense in $L^p(\mathbb{R}, X)$, property (i) follows.

$\mathfrak{G}$ consists of differentiable functions since for $v \in \mathcal{S}(\mathbb{R}, X)$, the integral defining g_v converges absolutely. Moreover, for $v \in \mathcal{S}(\mathbb{R}, X)$, the function $w(s) = (A - is)^{-1}v(s)$, $s \in \mathbb{R}$, is also in $\mathcal{S}(\mathbb{R}, X)$, since $\sup_{s\in\mathbb{R}} \|(A - is)^{-1}\| < \infty$. Hence f_v is differentiable with derivative

$$f_v'(\tau) = \frac{1}{2\pi}\int_{\mathbb{R}} is(A - is)^{-1}v(s)e^{i\tau s}\,ds = \frac{1}{2\pi}\int_{\mathbb{R}} isw(s)e^{i\tau s}\,ds.$$

So $f_v' \in L^p(\mathbb{R}, X)$, and hence $\mathfrak{F}$ is dense in $\mathrm{Dom}(-d/dt + \mathcal{A})$.

Property (iii) follows from the following calculation:

$$\begin{aligned}(\Gamma f_v)(\tau) &= \frac{1}{2\pi}\int_{\mathbb{R}} [-is(A - is)^{-1}v(s)e^{is\tau} + A(A - is)^{-1}v(s)e^{is\tau}]\,ds \\ &= \frac{1}{2\pi}\int_{\mathbb{R}} (A - is)(A - is)^{-1}v(s)e^{is\tau}\,ds = g_v(\tau). \qquad \square\end{aligned}$$

Set $\Lambda_{\mathcal{S}} = \{v \in \mathcal{S}(\mathbb{R}, X) : v(s) \in \mathrm{Dom}(A) \text{ for } s \in \mathbb{R},\ Av \in \mathcal{S}(\mathbb{R}, X)\}$.

PROPOSITION 4.9. *Let $\{e^{tA}\}_{t\geq 0}$ be a strongly continuous semigroup generated by A. Let Γ and $\Gamma_{\mathbb{R}}$ be the generators of the evolution semigroups on $L^p(\mathbb{R}_+, X)$ and $L^p(\mathbb{R}, X)$, as defined in (3.4) and (2.2), respectively. Then the following assertions hold.*

(i) *If $\sigma(A) \cap i\mathbb{R} = \emptyset$ and $\sup_{s\in\mathbb{R}} \|(A - is)^{-1}\| < \infty$, then*

$$\|\Gamma_{\mathbb{R}}\|_{\bullet, L^p(\mathbb{R},X)} = \inf_{v\in\Lambda_{\mathcal{S}}} \frac{\|\int_{\mathbb{R}}(A - is)v(s)e^{is(\cdot)}\,ds\|_{L^p(\mathbb{R},X)}}{\|\int_{\mathbb{R}} v(s)e^{is(\cdot)}\,ds\|_{L^p(\mathbb{R},X)}}.$$

(ii) *If $\Gamma_{\mathbb{R}}$ is invertible on $L^p(\mathbb{R}, X)$, then $\{e^{tA}\}_{t\geq 0}$ is hyperbolic and*

$$\|\Gamma_{\mathbb{R}}^{-1}\|_{\mathcal{L}(L^p(\mathbb{R},X))} = \sup_{v\in\mathcal{S}(\mathbb{R},X)} \frac{\|\int_{\mathbb{R}}(A - is)^{-1}v(s)e^{is(\cdot)}\,ds\|_{L^p(\mathbb{R},X)}}{\|\int_{\mathbb{R}} v(s)e^{is(\cdot)}\,ds\|_{L^p(\mathbb{R},X)}}.$$

(iii) *If Γ is invertible on $L^p(\mathbb{R}_+, X)$, then $\{e^{tA}\}_{t\geq 0}$ is exponentially stable and*

$$\|\Gamma^{-1}\|_{\mathcal{L}(L^p(\mathbb{R}_+,X))} = \|\Gamma_{\mathbb{R}}^{-1}\|_{\mathcal{L}(L^p(\mathbb{R},X))}.$$

Proof. To show (i), let $v \in \mathcal{S}(\mathbb{R}, X)$. Since $\sup_{s\in\mathbb{R}} \|(A - is)^{-1}\| < \infty$, the formula $w(s) = (A - is)^{-1}v(s)$, $s \in \mathbb{R}$, defines a function w in $\Lambda_{\mathcal{S}}$. Now

$$g_v(\tau) = \frac{1}{2\pi}\int_{\mathbb{R}}(A - is)(A - is)^{-1}v(s)e^{is\tau}\,ds = \frac{1}{2\pi}\int_{\mathbb{R}}(A - is)w(s)e^{is\tau}\,ds$$

and

$$f_v(\tau) = \frac{1}{2\pi}\int_{\mathbb{R}} w(s)e^{is\tau}\,ds.$$

However, from Proposition 4.8,

$$\|\Gamma_{\mathbb{R}}\|_{\bullet} = \inf_{f_v\in\mathfrak{F}} \frac{\|\Gamma_{\mathbb{R}} f_v\|}{\|f_v\|} = \inf_{v\in\mathcal{S}(\mathbb{R},X)} \frac{\|g_v\|}{\|f_v\|} = \inf_{w\in\Lambda_{\mathcal{S}}} \frac{\|\int_{\mathbb{R}}(A - is)w(s)e^{is(\cdot)}\,ds\|}{\|\int_{\mathbb{R}} w(s)e^{is(\cdot)}\,ds\|}.$$

To see (ii), note that

$$\|\Gamma_{\mathbb{R}}^{-1}\| = \|\Gamma_{\mathbb{R}}\|_\bullet^{-1} = \left[\inf_{v\in\mathcal{S}(\mathbb{R},X)} \frac{\|\Gamma_{\mathbb{R}} f_v\|}{\|f_v\|}\right]^{-1} = \sup_{v\in\mathcal{S}(\mathbb{R},X)} \frac{\|f_v\|}{\|g_v\|}.$$

For (iii) note that $\|\Gamma_{\mathbb{R}}\|_{\bullet,L^p(\mathbb{R},X)} \le \|\Gamma\|_{\bullet,L^p(\mathbb{R}_+,X)}$. Indeed, let $f \in L^p(\mathbb{R},X)$ with $\operatorname{supp} f \subseteq \mathbb{R}_+$. If $f \in \operatorname{Dom}(-d/dt+\mathcal{A})$, then $\operatorname{supp}\Gamma_{\mathbb{R}} f \subseteq \mathbb{R}_+$ and $\|\Gamma_{\mathbb{R}} f\|_{L^p(\mathbb{R},X)} = \|\Gamma f\|_{L^p(\mathbb{R}_+,X)}$. To see that $\|\Gamma_{\mathbb{R}}\|_\bullet \ge \|\Gamma\|_\bullet$, let $\epsilon > 0$ and choose $f \in \operatorname{Dom}(-d/dt+\mathcal{A})$ with compact support such that $\|f\|_{L^p(\mathbb{R},X)} = 1$ and $\|\Gamma_{\mathbb{R}}\|_\bullet \ge \|\Gamma_{\mathbb{R}} f\| - \epsilon$. Now choose $\tau \in \mathbb{R}$ such that $f_\tau(s) := f(s-\tau)$, $s \in \mathbb{R}$, defines a function, $f_\tau \in L^p(\mathbb{R},X)$, with $\operatorname{supp} f_\tau \subseteq \mathbb{R}_+$. Let $\bar{f}_\tau$ denote the element of $L^p(\mathbb{R}_+,X)$ which coincides with f_τ on $\mathbb{R}_+$. Then $\|f_\tau\| = \|\bar{f}_\tau\|$ and $\Gamma\bar{f}_\tau = -d/dt\, f(\cdot-\tau) + Af(\cdot-\tau) = (\Gamma_{\mathbb{R}} f)_\tau$. Therefore, $\|\Gamma_{\mathbb{R}}\|_\bullet \ge \|\Gamma_{\mathbb{R}} f\| - \epsilon = \|(\Gamma_{\mathbb{R}} f)_\tau\| - \epsilon = \|\Gamma_{\mathbb{R}}\bar{f}_\tau\| - \epsilon \ge \|\Gamma\|_\bullet - \epsilon$. ∎

PROPOSITION 4.10. *The set* $\mathfrak{G}_U = \{g_u : u \in \mathcal{S}(\mathbb{R},U)\}$ *is dense in* $L^p(\mathbb{R},U)$. *If* $u \in \mathcal{S}(\mathbb{R},U)$ *and* $B \in \mathcal{L}(U,X)$, *then* $Bu \in \mathcal{S}(\mathbb{R},X)$ *and* $\Gamma_{\mathbb{R}} f_{Bu} = \mathcal{B} g_u$.

Proof. The first statement is clear, as in Proposition 4.8. The second follows from the properties of Schwartz functions, and from the calculation

$$\Gamma_{\mathbb{R}} f_{Bu} = g_{Bu}(\tau) = \frac{1}{2\pi}\int_{\mathbb{R}} Bu(s)e^{is\tau}\,ds = B\frac{1}{2\pi}\int_{\mathbb{R}} u(s)e^{is\tau}\,ds. \qquad ∎$$

Recall (see Remarks 2.4 and Theorem 3.2) that $\{e^{tA}\}_{t\ge0}$ is hyperbolic (respectively, stable) if and only if $\Gamma_{\mathbb{R}}$ (respectively, Γ) is invertible on $L^p(\mathbb{R},X)$ (respectively, $L^p(\mathbb{R}_+,X)$).

THEOREM 4.11. *If* $\Gamma_{\mathbb{R}}$ *is invertible on* $L^p(\mathbb{R},X)$, *then*

$$\|\mathcal{C}\Gamma_{\mathbb{R}}^{-1}\mathcal{B}\| = \sup_{u\in\mathcal{S}(\mathbb{R},U)} \frac{\|\int_{\mathbb{R}} C(A-is)^{-1}Bu(s)e^{is(\cdot)}\,ds\|_{L^p(\mathbb{R},Y)}}{\|\int_{\mathbb{R}} u(s)e^{is(\cdot)}\,ds\|_{L^p(\mathbb{R},U)}}. \tag{4.18}$$

If Γ *is invertible on* $L^p(\mathbb{R}_+,X)$, *then the norm of* $\mathbb{L} = \mathcal{C}\Gamma^{-1}\mathcal{B}$, *as an operator from* $L^p(\mathbb{R}_+,U)$ *to* $L^p(\mathbb{R}_+,Y)$, *is given by the above formula:*

$$\|\mathbb{L}\| = \|\mathcal{C}\Gamma_{\mathbb{R}}^{-1}\mathcal{B}\|. \tag{4.19}$$

If, in addition, U *and* Y *are Hilbert spaces and* $p = 2$, *then*

$$\|\mathbb{L}\| = \sup_{s\in\mathbb{R}} \|C(A-is)^{-1}B\|_{\mathcal{L}(U,Y)}. \tag{4.20}$$

Proof. For $u \in \mathcal{S}(\mathbb{R},U)$, consider functions f_{Bu} and g_u. Proposition 4.10 gives $f_{Bu} = \Gamma_{\mathbb{R}}^{-1}\mathcal{B} g_u$ and

$$\begin{aligned}\|\mathcal{C}\Gamma_{\mathbb{R}}^{-1}\mathcal{B}\| &= \sup_{g_u\in\mathfrak{G}_U} \frac{\|\mathcal{C}\Gamma_{\mathbb{R}}^{-1}\mathcal{B} g_u\|_{L^p(\mathbb{R},Y)}}{\|g_u\|_{L^p(\mathbb{R},U)}} = \sup_{g_u\in\mathfrak{G}_U} \frac{\|\mathcal{C} f_{Bu}\|}{\|g_u\|} \\ &= \sup_{u\in\mathcal{S}(\mathbb{R},U)} \frac{\|\int_{\mathbb{R}} C(A-is)^{-1}Bu(s)e^{is(\cdot)}\,ds\|_{L^p(\mathbb{R},Y)}}{\|\int_{\mathbb{R}} u(s)e^{is(\cdot)}\,ds\|_{L^p(\mathbb{R},U)}},\end{aligned}$$

which proves (4.18).

Now if Γ is invertible on $L^p(\mathbb{R}_+,X)$, then $\{e^{tA}\}_{t\ge0}$ is exponentially stable by Corollary 3.6. Hence $\Gamma_{\mathbb{R}}$ is invertible on $L^p(\mathbb{R},X)$. Moreover, for the case of the *stable* semigroup $\{e^{tA}\}_{t\ge0}$, the formula for $\Gamma_{\mathbb{R}}^{-1}$ (see, e.g., [25]) takes the form

$$(\Gamma_{\mathbb{R}}^{-1} f)(t) = \int_0^\infty e^{sA} f(t-s)\,ds = \int_{-\infty}^t e^{(t-s)A} f(s)\,ds.$$

If $\operatorname{supp} f \subseteq (0,\infty)$, then

$$(\Gamma_{\mathbb{R}}^{-1} f)(t) = \int_{-\infty}^{t} e^{(t-s)A} f(s)\, ds = \int_{0}^{t} e^{(t-s)A} f(s)\, ds. \tag{4.21}$$

For a function $h \in L^p(\mathbb{R}_+, X)$, define an extension $\tilde{h} \in L^p(\mathbb{R}, X)$ by $\tilde{h}(t) = h(t)$ for $t \geq 0$ and $\tilde{h}(t) = 0$ for $t < 0$. Then (4.21) shows that $\Gamma_{\mathbb{R}}^{-1}\tilde{h} = (\Gamma^{-1}h)^{\sim}$. In particular, for $u \in L^p(\mathbb{R}_+, U)$, $\widetilde{\mathbb{L}u} = \widetilde{\mathcal{C}\Gamma^{-1}\mathcal{B}u} = \mathcal{C}\Gamma_{\mathbb{R}}^{-1}\mathcal{B}\tilde{u}$. Therefore,

$$\begin{aligned}\|\mathbb{L}u\|_{L^p(\mathbb{R}_+,Y)} &= \|\widetilde{\mathbb{L}u}\|_{L^p(\mathbb{R},Y)} = \|\mathcal{C}\Gamma_{\mathbb{R}}^{-1}\mathcal{B}\tilde{u}\|_{L^p(\mathbb{R},Y)} \\ &\leq \|\mathcal{C}\Gamma_{\mathbb{R}}^{-1}\mathcal{B}\| \cdot \|\tilde{u}\|_{L^p(\mathbb{R},U)} = \|\mathcal{C}\Gamma_{\mathbb{R}}^{-1}\mathcal{B}\| \cdot \|u\|_{L^p(\mathbb{R}_+,U)}.\end{aligned}$$

This shows that $\|\mathbb{L}\| \leq \|\mathcal{C}\Gamma_{\mathbb{R}}^{-1}\mathcal{B}\|$.

To prove that equality holds in (4.19), let $\epsilon > 0$ and choose $u \in L^p(\mathbb{R}, U)$, $\|u\| = 1$, such that $\|\mathcal{C}\Gamma_{\mathbb{R}}^{-1}\mathcal{B}u\|_{L^p(\mathbb{R},Y)} \geq \|\mathcal{C}\Gamma_{\mathbb{R}}^{-1}\mathcal{B}\| - \epsilon$. Without loss of generality, u may be assumed to have compact support. Now choose r such that $\operatorname{supp} u(\cdot - r) \subseteq (0,\infty)$ and set $w(\cdot) := u(\cdot - r)$. Then $w \in L^p(\mathbb{R}, U)$ with $\operatorname{supp} w \subseteq (0,\infty)$. Let $\bar{w}$ denote the element of $L^p(\mathbb{R}_+, U)$ that coincides with w on $\mathbb{R}_+$. As in (4.21) we have

$$\mathcal{C}\Gamma_{\mathbb{R}}^{-1}\mathcal{B}w(t) = C\int_{0}^{t} e^{(t-s)A}Bw(s)\, ds = C\int_{-\infty}^{t} e^{(t-s)A}Bw(s)\, ds.$$

Since $\|\bar{w}\|_{L^p(\mathbb{R}_+,U)} = \|w\|_{L^p(\mathbb{R},U)} = \|u\|_{L^p(\mathbb{R},U)} = 1$, it follows that

$$\begin{aligned}\|\mathbb{L}\| &\geq \|\mathbb{L}\bar{w}\|_{L^p(\mathbb{R}_+,Y)} = \|\widetilde{\mathbb{L}\bar{w}}\|_{L^p(\mathbb{R},Y)} \\ &= \|\mathbb{L}\tilde{w}\|_{L^p(\mathbb{R},Y)} = \|\mathcal{C}\Gamma_{\mathbb{R}}^{-1}\mathcal{B}w\|_{L^p(\mathbb{R},Y)} \\ &= \|C\int_{-\infty}^{\cdot} e^{(\cdot-\tau)A}Bu(\tau)\, d\tau\|_{L^p(\mathbb{R},Y)} \\ &= \|\mathcal{C}\Gamma_{\mathbb{R}}^{-1}\mathcal{B}u\|_{L^p(\mathbb{R},Y)} \geq \|\mathcal{C}\Gamma_{\mathbb{R}}^{-1}\mathcal{B}\| - \epsilon.\end{aligned}$$

This confirms (4.19). Parseval's formula and (4.7) give (4.20). □

If $\{e^{tA}\}_{t\geq 0}$ is exponentially stable, then the inequalities in (4.7) give lower and upper bounds on the stability radius in terms of $\mathbb{L}$ and $C(A - is)^{-1}B$, respectively. The previous theorem shows that $\|\mathbb{L}\|$ can be explicitly expressed in terms of an integral involving $C(A - is)^{-1}B$. We conclude by observing that a lower bound for the constant stability radius can be expressed by a similar formula involving a sum. For this, let $\xi \in [0,1]$ and set

$$S_\xi := \sup_{\{u_k\}\in\Lambda} \frac{\|\sum_k C(A - i\xi - ik)^{-1}Bu_k e^{ik(\cdot)}\|_{L^p([0,2\pi],Y)}}{\|\sum_k u_k e^{ik(\cdot)}\|_{L^p([0,2\pi],U)}}.$$

We note that S_ξ is computed as in (4.11) with A replaced by $A_\xi = A - i\xi$.

COROLLARY 4.12. *Let $\{e^{tA}\}_{t\geq 0}$ be an exponentially stable semigroup generated by A. Then*

$$\frac{1}{\sup_{\xi\in[0,1]} S_\xi} \leq rc_{stab}(\{e^{tA}\}, B, C) \leq \frac{1}{\sup_{s\in\mathbb{R}} \|C(A - is)^{-1}B\|}.$$

Proof. Fix $\xi \in [0,1]$, and let $\Gamma_{per,\xi}$ denote the generator on $L^p([0,2\pi],X)$ of the evolution semigroup induced by $\{e^{tA_\xi}\}_{t\geq 0}$. By Theorem 4.3, $\|\mathcal{C}\Gamma_{per,\xi}^{-1}\mathcal{B}\| = S_\xi$, and so by Theorem 4.4,

$$\frac{1}{S_\xi} \leq rc^1_{stab}(e^{2\pi A_\xi}, B, C) \leq \frac{1}{\sup_{k\in\mathbb{Z}} \|C(A_\xi - ik)^{-1}B\|}.$$

By Proposition 4.6, taking the infimum over $\xi \in [0,1]$ gives

$$\begin{aligned}\frac{1}{\sup_{\xi\in[0,1]} S_\xi} &\leq \inf_{\xi\in[0,1]} rc^1_{stab}(e^{2\pi A_\xi}, B, C) = rc_{stab}(\{e^{tA}\}, B, C)\\ &\leq \inf_{\xi\in[0,1]} \frac{1}{\sup_{k\in\mathbb{Z}} \|C(A_\xi - ik)^{-1}B\|}\\ &= \frac{1}{\sup_{s\in\mathbb{R}} \|C(A - is)^{-1}B\|}. \qquad \square\end{aligned}$$

4.5. Two counterexamples. In contrast to the Hilbert-space setting, the following Banach-space examples show that either inequality in (4.7) may be strict. We start with the example where the second inequality in (4.7) is strict.

Example 4.13. An example due to Arendt (see, e.g., [29], Ex. 1.4.5) exhibits a (positive) strongly continuous semigroup $\{e^{tA}\}_{t\geq 0}$ on a Banach space X with the property that $s_0(A) < \omega_0(A) < 0$ for the abscissa of uniform boundedness of the resolvent and the growth bound. Now, for α such that $0 \leq \alpha \leq -\omega_0(A)$, consider a rescaled semigroup generated by $A + \alpha$, and denote by $\Gamma_{A+\alpha}$ the generator of the induced evolution semigroup on $L^p(\mathbb{R}_+, X)$. The following relationships hold:

$$\text{for } 0 \leq \alpha < -\omega_0(A), \quad s_0(A+\alpha) = s_0(A) + \alpha < \omega_0(A) + \alpha = \omega_0(A+\alpha) < 0;$$

$$\text{for } \alpha_0 := -\omega_0(A), \quad s_0(A+\alpha_0) < \omega_0(A+\alpha_0) = 0.$$

This says that $s_0(A+\alpha) < 0$ for all $\alpha \in [0,\alpha_0]$ and hence

$$M := \sup_{\alpha\in[0,\alpha_0]} \sup_{s\in\mathbb{R}} \|(A + \alpha - is)^{-1}\| < \infty.$$

Now note (see Corollary 3.6) that $\omega_0(A+\alpha) < 0$ if and only if $\|\Gamma^{-1}_{A+\alpha}\| < \infty$. Since $\omega_0(A+\alpha) \to 0$ as $\alpha \to \alpha_0$, we conclude that $\|\Gamma^{-1}_{A+\alpha}\| \to \infty$ as $\alpha \to \alpha_0$. Since $\alpha \mapsto \|\Gamma^{-1}_{A+\alpha}\|$ is a continuous function of α on $[0,\alpha_0)$, there exists $\alpha_1 \in [0,\alpha_0)$ such that $\|\Gamma^{-1}_{A+\alpha_1}\| > M$, and so the following inequality is strict:

$$\frac{1}{\|\Gamma^{-1}_{A+\alpha_1}\|} < \frac{1}{\sup_{s\in\mathbb{R}} \|(A + \alpha_1 - is)^{-1}\|}.$$

Also, we claim that there exists $\alpha_2 \in [0,\alpha_0)$ such that the following inequality is strict:

$$rc_{stab}(\{e^{t(A+\alpha_2)}\}, I, I) < \frac{1}{\sup_{s\in\mathbb{R}} \|(A + \alpha_2 - is)^{-1}\|}.$$

To see this, let us suppose that for each $\alpha \in [0,\alpha_0)$ one has $rc_{stab}(\{e^{t(A+\alpha)}\}, I, I) \geq 1/(2M)$. Again, using that $\omega_0(A+\alpha) \to 0$ as $\alpha \to \alpha_0$, find $\alpha \in [0,\alpha_0)$ such that

$|\omega_0(A+\alpha)| < 1/(2M)$. Let $\Delta = \omega_0(A+\alpha)I$. Since $\|\Delta\| = |\omega_0(A+\alpha)|$, by the definition of stability radius one has

$$0 > \omega_0(A+\alpha+\Delta) = \omega_0(A+\alpha) - \omega_0(A+\alpha) = 0,$$

which is a contradiction. Thus there exists $\alpha_2 \in [0, \alpha_0)$ such that

$$rc_{stab}(\{e^{t(A+\alpha_2)}\}, I, I) \le \frac{1}{2M} < \frac{1}{M} \le \frac{1}{\sup_{s\in\mathbb{R}} \|(A+\alpha_2 - is)^{-1}\|},$$

as claimed. ∎

This example shows that the second inequality in (4.7) can be strict due to the Banach-space pathologies related to the failure of Gearhart's Theorem 2.1. Another example, given below, shows that the first inequality in (4.7) could be strict due to the lack of Parseval's formula (see (4.15) in the proof of Theorem 4.4). That is, the choice of $p = 2$ in (4.6) is as important as the fact that X in (4.6) is a *Hilbert* space. First, we need a formula for the norm of the input-output operator on $L^1(\mathbb{R}_+, X)$.

PROPOSITION 4.14. *Assume $\{e^{tA}\}_{t\ge 0}$ is an exponentially stable C_0 semigroup on a Banach space X. The norm of the operator $\mathbb{L} = \Gamma^{-1}$ on $L^1(\mathbb{R}_+, X)$ is*

$$\|\Gamma^{-1}\|_{\mathcal{L}(L^1(\mathbb{R}_+,X))} = \sup_{\|x\|=1} \int_0^\infty \|e^{tA}x\|\, dt. \tag{4.22}$$

Proof. Recall (see (3.5)) that

$$\Gamma^{-1} f(t) = -\int_0^t e^{\tau A} f(t-\tau)\, d\tau \quad t \in \mathbb{R}_+, \quad f \in L^1(\mathbb{R}_+, X)$$

is the convolution operator. Choose positive $\delta_n \in L^1(\mathbb{R}_+, \mathbb{R})$ with $\|\delta_n\|_{L^1} = 1$ such that

$$\|g * \delta_n - g\|_{L^1(\mathbb{R}_+,X)} \to 0 \quad \text{as} \quad n \to \infty \quad \text{for each} \quad g \in L^1(\mathbb{R}_+, X).$$

Fix $x \in X$, $\|x\| = 1$, let $f = \delta_n x \in L^1(\mathbb{R}_+, X)$, and note that

$$\Gamma^{-1} f(t) = -\int_0^t e^{\tau A} x\, \delta_n(t-\tau)\, d\tau = -(g * \delta_n)(t) \quad \text{for} \quad g(t) = e^{tA}x, \quad t \in \mathbb{R}_+.$$

This implies "$\ge$" in (4.22). To see "$\le$," take $f = \sum_{i=1}^N \alpha_i x_i$ with $\alpha_i \in L^1(\mathbb{R}_+, \mathbb{R})$ having disjoint supports and $\|x_i\| = 1$, $i = 1, \dots, N$. Now $\|f\|_{L^1(\mathbb{R}_+,X)} = \sum_i \|\alpha_i\|_{L^1}$, and for $f_i(t) = e^{tA}x_i$ one has

$$\Gamma^{-1} f(t) = -\int_0^t \sum_i e^{\tau A} x_i \alpha_i(t-\tau)\, d\tau = -\sum_i (f_i * \alpha_i)(t).$$

Using Young's inequality,

$$\begin{aligned}\|\Gamma^{-1} f\|_{L^1(\mathbb{R}_+,X)} &\le \sum_i \|f_i * \alpha_i\|_{L^1(\mathbb{R}_+,X)} \\ &\le \sum_i \|f_i\|_{L^1(\mathbb{R}_+,X)} \|\alpha_i\|_{L^1} \le \sup_{\|x\|=1} \int_0^\infty \|e^{tA}x\|\, dt \sum_i \|\alpha_i\|_{L^1}. \qquad \blacksquare\end{aligned}$$

Example 4.15. Take $X = \mathbb{C}^2$ with the ℓ_1 norm. Let

$$A = \begin{pmatrix} -1 & 1 \\ -1 & -1 \end{pmatrix} \quad \text{so that} \quad e^{tA} = \begin{pmatrix} e^{-t}\cos(t) & e^{-t}\sin(t) \\ -e^{-t}\sin(t) & e^{-t}\cos(t) \end{pmatrix},$$

and

$$(A - is)^{-1} = \frac{1}{(1+is)^2 + 1} \begin{pmatrix} -1 - is & -1 \\ 1 & -1 - is \end{pmatrix}.$$

Since the extreme points of X are $e^{i\theta}e_1$ and $e^{i\theta}e_2$ ($\theta \in \mathbb{R}$), where e_1 and e_2 are the unit vectors of $\mathbb{C}^2$, we see that

$$\|(A - is)^{-1}\| = \frac{|1 + is| + 1}{|(1+is)^2 + 1|}.$$

It may be numerically established that

$$\sup_{s \in \mathbb{R}} \|(A - isI)^{-1}\| \approx 1.087494476.$$

By Corollary 4.7, the reciprocal to the last expression is equal to $rc_{stab}(\{e^{tA}\}, I, I)$. On the other hand, using Proposition 4.14,

$$\begin{aligned} \|\mathbb{L}\| = \|\Gamma_A^{-1}\| &= \sup_{\|x\|=1} \int_0^\infty \|e^{tA}x\|\, dt \\ &= \int_0^\infty |e^{-t}\cos(t)| + |e^{-t}\sin(t)|\, dt \approx 1.262434309. \end{aligned}$$

Therefore, the first inequality in (4.7) may be strict. □

The following example shows that the norm of the input-output operator depends on p.

Example 4.16. Let

$$A = \begin{pmatrix} 9/2 & -5/2 \\ 25/2 & -13/2 \end{pmatrix},$$

acting on $\mathbb{C}^2$ with the Euclidean norm. Thus

$$e^{tA} = e^{-t} \begin{pmatrix} \cos t + (11/2)\sin t & -(5/2)\sin t \\ (25/2)\sin t & \cos t - (11/2)\sin t \end{pmatrix}.$$

Then

$$\|\Gamma^{-1}\|_{L_1 \to L_1} \geq \int_0^\infty \|e^{tA}e_1\|\, dt \approx 7.748310791,$$

whereas

$$\|\Gamma^{-1}\|_{L_2 \to L_2} = \sup_{s \in R} \|(A - is)^{-1}\| \approx 2.732492852. \qquad \square$$

5. Internal and external stability. Work aimed at properties of stability and robustness of linear time-invariant systems is often based on transform techniques. More specifically, if the transfer function $H(\lambda) = C(A-\lambda)^{-1}B$ is a bounded analytic function of λ in the right half-plane $\mathbb{C}_+ = \{\lambda \in \mathbb{C} : \mathrm{Re}\lambda > 0\}$, then the autonomous system (4.4) is said to be *externally stable*. This property is often used to deduce *internal stability* of the system, i.e., the uniform exponential stability of the nominal system $\dot{x} = Ax$. The relationship between internal and external stability has been studied extensively; see, e.g., [3, 9, 8, 21, 26, 39, 40] and the references therein. In this section we examine the extent to which these techniques apply to Banach-space settings and time-varying systems. For this, *input-output stability* of the system (4.3) will refer to the property that the input-output operator $\mathbb{L}$ is bounded from $L^p(\mathbb{R}_+, U)$ to $L^p(\mathbb{R}_+, Y)$. If internal stability is assumed initially, then the inequalities in (4.7) exhibit a relationship between these concepts of stability. The next two theorems look at these relationships more closely and show, in particular, when internal stability may be deduced from one of the "external" stability conditions. Therefore, throughout this section $\{U(t,\tau)\}_{t\geq\tau}$ will denote a strongly continuous exponentially bounded evolution family that is *not* assumed to be exponentially stable.

5.1. The nonautonomous case. In this subsection we give a very short proof of the fact that for general nonautonomous systems on Banach spaces, internal stability is equivalent to stabilizability, detectability, and input-output stability. Before proceeding, it is worth reviewing some properties of time-invariant systems. For this, let $\{e^{tA}\}_{t\geq 0}$ be a strongly continuous semigroup generated by A on X, and let $H_+^\infty(\mathcal{L}(X))$ denote the space of operator-valued functions $G : \mathbb{C} \to \mathcal{L}(X)$ which are analytic on $\mathbb{C}_+$ and $\sup_{\lambda\in\mathbb{C}_+} \|G(\lambda)\| < \infty$. If X is a Hilbert space, it is well known that $\{e^{tA}\}_{t\geq 0}$ is exponentially stable if and only if $\lambda \mapsto (\lambda - A)^{-1}$ is an element of $H_+^\infty(\mathcal{L}(X))$; see, e.g., [11, Thm. 5.1.5]. This is a consequence of the fact that when X is a Hilbert space, $s_0(A) = \omega_0(e^{tA})$ (see [29] or Theorem 2.1). If X is a Banach space, then strict inequality $s_0(A) < \omega_0(e^{tA})$ can hold, and so exponential stability is no longer determined by the operator $G(\lambda) = (\lambda - A)^{-1}$. Extending these ideas to address systems (4.4), one considers $H(\lambda) = C(\lambda - A)^{-1}B$: it can be shown that if U and Y are Hilbert spaces, then (4.4) *is internally stable if and only if it is stabilizable, detectable, and externally stable* (i.e., $H(\cdot) \in H_+^\infty(\mathcal{L}(U,Y))$). See Rebarber [39] for a general result of this type. It should be pointed out that this work of Rebarber and others more recently allows for a certain degree of unboundedness of the operators B and C. Such "regular" systems (see [44]) and their time-varying generalizations might be addressed by combining the techniques of the present paper (including the characterization of generation of evolution semigroups as found in [36]) along with those of [18] and [20]. This will not be done here.

If one allows for Banach spaces, the conditions of stabilizability and detectability are *not* sufficient to ensure that external stability implies internal stability. Indeed, let A generate a semigroup for which $s_0(A) < \omega_0(e^{tA}) = 0$ (see Example 4.13). Then the system (4.4) with $B = I$ and $C = I$ is trivially stabilizable, detectable, and externally stable. But since $\omega_0(e^{tA}) = 0$, it is not internally stable.

Since the above italicized statement concerning external stability fails for Banach-space systems (4.4) and does not apply to time-varying systems (4.3), we aim to prove the following extension of this.

THEOREM 5.1. *The system* (4.3) *is internally stable if and only if it is stabilizable, detectable, and input-output stable.*

This theorem appears as part of Theorem 5.3. A version of it for finite-dimensional

time-varying systems was proven by Anderson in [3]. The fact that Theorem 5.1 actually *extends* the Hilbert-space statement above follows from the fact that the Banach-space inequality $\sup_{\lambda\in\mathbb{C}_+}\|H(\lambda)\|\le\|\mathbb{L}\|$ (see [45]) which relates the operators that define external and input-output stability is actually an *equality* for Hilbert-space systems (see also [44]).

In Theorems 5.1 and 5.3 the following definitions are used.

DEFINITION 5.2. *The nonautonomous system* (4.3) *is said to be*

(a) stabilizable *if there exists* $F(\cdot)\in L^\infty(\mathbb{R}_+,\mathcal{L}_s(X,U))$ *and a corresponding exponentially stable evolution family* $\{U_{BF}(t,\tau)\}_{t\ge\tau}$ *such that, for* $t\ge s$ *and* $x\in X$, *one has*

$$U_{BF}(t,s)x=U(t,s)x+\int_s^t U(t,\tau)B(\tau)F(\tau)U_{BF}(\tau,s)x\,d\tau; \tag{5.1}$$

(b) detectable *if there exists* $K(\cdot)\in L^\infty(\mathbb{R}_+,\mathcal{L}_s(Y,X))$ *and a corresponding exponentially stable evolution family* $\{U_{KC}(t,\tau)\}_{t\ge\tau}$ *such that, for* $t\ge s$ *and* $x\in X$, *one has*

$$U_{KC}(t,s)x=U(t,s)x+\int_s^t U_{KC}(t,\tau)K(\tau)C(\tau)U(\tau,s)x\,d\tau. \tag{5.2}$$

An autonomous control system is called stabilizable *if there is an operator* $F\in\mathcal{L}(X,U)$ *such that* $A+BF$ *generates a uniformly exponentially stable semigroup; that is,* $\omega_0(A+BF)<0$. *Such a system is* detectable *if there is an operator* $K\in\mathcal{L}(Y,X)$ *such that* $A+KC$ *generates a uniformly exponentially stable semigroup.*

Using Theorem 3.5 to characterize exponential stability in terms of the operator $\mathbb{G}$ as in (3.6) makes the proof of the following theorem a straightforward manipulation of the appropriate operators.

THEOREM 5.3. *The following are equivalent for a strongly continuous exponentially bounded evolution family of operators* $\mathcal{U}=\{U(t,\tau)\}_{t\ge\tau}$ *on a Banach space* X.

(i) $\mathcal{U}$ *is exponentially stable on* X.
(ii) $\mathbb{G}$ *is a bounded operator on* $L^p(\mathbb{R}_+,X)$.
(iii) *System* (4.3) *is stabilizable and* $\mathbb{G}\mathcal{B}$ *is a bounded operator from* $L^p(\mathbb{R}_+,U)$ *to* $L^p(\mathbb{R}_+,X)$.
(iv) *System* (4.3) *is detectable and* $\mathcal{C}\mathbb{G}$ *is a bounded operator from* $L^p(\mathbb{R}_+,X)$ *to* $L^p(\mathbb{R}_+,Y)$.
(v) *System* (4.3) *is stabilizable and detectable and* $\mathbb{L}=\mathcal{C}\mathbb{G}\mathcal{B}$ *is a bounded operator from* $L^p(\mathbb{R}_+,U)$ *to* $L^p(\mathbb{R}_+,Y)$.

Proof. The equivalence of (i) and (ii) is the equivalence of (i) and (ii) in Theorem 3.5.

To see that (ii) implies (iii), (iv), and (v), note that $\mathcal{B}$ and $\mathcal{C}$ are bounded, and thus $\mathbb{L}$ is bounded when $\mathbb{G}$ is bounded. So when (ii) holds, the exponential stability of $\mathcal{U}$ together with the boundedness of $B(\cdot)$, $C(\cdot)$, $F(\cdot)$, and $K(\cdot)$ assure the existence of the evolution families $\{U_{BF}(t,\tau)\}_{t\ge\tau}$ and $\{U_{KC}(t,\tau)\}_{t\ge\tau}$ as solutions of the integral equations in Definition 5.2, thereby showing that (iii), (iv), and (v) hold.

To see that (iii) $\Rightarrow$ (ii), first note that the assumption of stabilizability assures the existence of an exponentially stable evolution family $\mathcal{U}_{BF}=\{U_{BF}(t,\tau)\}_{t\ge\tau}$ satisfying (5.1) for some $F(\cdot)\in L^\infty(\mathbb{R}_+,\mathcal{L}_s(X,U))$. Given this exponentially stable family, we define the operator $\mathbb{G}_{BF}$ by

$$\mathbb{G}_{BF}f(s):=\int_0^s U_{BF}(s,\tau)f(\tau)\,d\tau=\int_0^\infty (E^\tau_{BF}f)(s)\,d\tau, \tag{5.3}$$

where $\{E^\tau_{BF} f\}_{t\geq 0}$ is the semigroup induced by the evolution family $\mathcal{U}_{BF}$ as described in (3.1). $\mathbb{G}_{BF}$ is a bounded operator on $L^p(\mathbb{R}_+, X)$ by the equivalence of (i) and (ii).

For $f(\cdot) \in L^p(\mathbb{R}_+, X)$ and $s \in \mathbb{R}_+$, take $x = f(s)$ in (5.1). Then let $\xi = \tau - s$ to obtain

$$U_{BF}(t,s)f(s) = U(t,s)f(s) + \int_0^{t-s} U(t,\xi+s)B(\xi+s)F(\xi+s)U_{BF}(\xi+s,s)f(s)\,d\xi.$$

From this equation and from the definition of the semigroups $\{E^t\}_{t\geq 0}$ and $\{E^t_{BF}\}_{t\geq 0}$ we obtain

$$(E^{t-s}_{BF} f)(t) = (E^{t-s} f)(t) + \int_0^{t-s} (E^{t-s-\xi}\mathcal{B}\mathcal{F}E^\xi_{BF} f)(t)\,d\xi$$

and hence for $0 \leq r$ and $0 \leq \sigma$ that

$$(E^r_{BF} f)(\sigma) = (E^r f)(\sigma) + \int_0^r (E^{r-\xi}\mathcal{B}\mathcal{F}E^\xi_{BF} f)(\sigma)\,d\xi.$$

Integrate from 0 to ∞ to obtain

$$(\mathbb{G}_{BF} f)(\sigma) = (\mathbb{G} f)(\sigma) + \int_0^\infty \int_0^r (E^{r-\xi}\mathcal{B}\mathcal{F}E^\xi_{BF} f)(\sigma)\,d\xi\,dr.$$

Let $r = \zeta + \eta$ and $\xi = \eta$ to obtain

$$\begin{aligned}(\mathbb{G}_{BF} f)(\sigma) &= (\mathbb{G} f)(\sigma) + \int_0^\infty \int_0^\infty (E^\zeta \mathcal{B}\mathcal{F}E^\eta_{BF} f)(\sigma)\,d\eta\,d\zeta \\ &= (\mathbb{G} f)(\sigma) + (\mathbb{G}\mathcal{B}\mathcal{F}\mathbb{G}_{BF} f)(\sigma).\end{aligned} \tag{5.4}$$

That $\mathbb{G}$ is bounded now follows from (5.4), the boundedness of $\mathbb{G}\mathcal{B}$, and the boundedness of $\mathbb{G}_{BF}$ and $\mathcal{F}$.

To see that (iv) $\Rightarrow$ (ii), first note that the assumption of detectability assures the existence of an exponentially stable evolution family $\mathcal{U}_{KC} = \{U_{KC}(t,\tau)\}_{t\geq\tau}$ satisfying (5.2) for some $K(\cdot) \in L^\infty(\mathbb{R}_+, \mathcal{L}_s(Y,X))$. Given this exponentially stable family, the operator $\mathbb{G}_{KC}$, defined in a manner analogous to $\mathbb{G}_{BF}$ in (5.3), is a bounded operator on $L^p(\mathbb{R}_+, X)$. A derivation beginning with (5.2) and similar to that which gave (5.4) now gives $\mathbb{G}_{KC} = \mathbb{G} + \mathbb{G}_{KC}\mathcal{K}\mathcal{C}\mathbb{G}$. This equation, together with the assumed boundedness of $\mathbb{G}_{KC}$, $\mathcal{K}$, and $\mathcal{C}\mathbb{G}$, gives the boundedness of $\mathbb{G}$.

Finally, to see that (v) $\Rightarrow$ (ii), again note that the assumption of detectability yields an exponentially stable evolution family $\mathcal{U}_{KC}$ and an associated bounded operator $\mathbb{G}_{KC}$. For $u(\cdot) \in L^p(\mathbb{R}_+, U)$ and $s \in \mathbb{R}_+$ take $x = B(s)u(s)$ in (5.2). A calculation similar to that which gave (5.4) now gives $\mathbb{G}_{KC}\mathcal{B} = \mathbb{G}\mathcal{B} + \mathbb{G}_{KC}\mathcal{K}\mathcal{C}\mathbb{G}\mathcal{B}$. The assumed boundedness of $\mathbb{L} = \mathcal{C}\mathbb{G}\mathcal{B}$, $\mathcal{K}$, and $\mathbb{G}_{KC}$ now yields the boundedness of $\mathbb{G}\mathcal{B}$. The boundedness of $\mathbb{G}\mathcal{B}$ together with the assumption of stabilizability implies that $\mathbb{G}$ is bounded by the equivalence of (iii) and (ii). □

5.2. The autonomous case. The main result of this subsection is Theorem 5.4 which builds on Theorem 4.11 and parallels Theorem 5.3 for autonomous systems of the form (4.4). The main point is to provide explicit conditions, in terms of the operators A, B, and C, which imply internal stability.

Let $A_\alpha := A - \alpha I$ denote the generator of the rescaled semigroup $\{e^{-\alpha t}e^{tA}\}_{t\geq 0}$.

THEOREM 5.4. *Let $\{e^{tA}\}_{t\geq 0}$ be a strongly continuous semigroup on a Banach space X generated by A. Let U and Y be Banach spaces and assume $B \in \mathcal{L}(U,X)$ and $C \in \mathcal{L}(X,Y)$. Then the following are equivalent.*

(i) *$\{e^{tA}\}_{t\geq 0}$ is exponentially stable.*

(ii) *$\mathbb{G}$ is a bounded operator on $L^p(\mathbb{R}_+,X)$.*

(iii) *$\sigma(A)\cap\overline{\mathbb{C}}_+ = \emptyset$ and $\displaystyle\sup_{v\in\mathcal{S}(\mathbb{R},X)} \frac{\|\int_{\mathbb{R}}(A_\alpha - is)^{-1}v(s)e^{is(\cdot)}\,ds\|_{L^p(\mathbb{R},X)}}{\|\int_{\mathbb{R}} v(s)e^{is(\cdot)}\,ds\|_{L^p(\mathbb{R},X)}} < \infty$ for all $\alpha \geq 0$.*

(iv) *$\sigma(A)\cap\overline{\mathbb{C}}_+ = \emptyset$ and $\displaystyle\sup_{u\in\mathcal{S}(\mathbb{R},U)} \frac{\|\int_{\mathbb{R}}(A_\alpha - is)^{-1}Bu(s)e^{is(\cdot)}\,ds\|_{L^p(\mathbb{R},X)}}{\|\int_{\mathbb{R}} u(s)e^{is(\cdot)}\,ds\|_{L^p(\mathbb{R},U)}} < \infty$ for all $\alpha \geq 0$, and (4.4) is stabilizable.*

(v) *$\sigma(A)\cap\overline{\mathbb{C}}_+ = \emptyset$ and $\displaystyle\sup_{v\in\mathcal{S}(\mathbb{R},X)} \frac{\|\int_{\mathbb{R}} C(A_\alpha - is)^{-1}v(s)e^{is(\cdot)}\,ds\|_{L^p(\mathbb{R},Y)}}{\|\int_{\mathbb{R}} v(s)e^{is(\cdot)}\,ds\|_{L^p(\mathbb{R},X)}} < \infty$ for all $\alpha \geq 0$, and (4.4) is detectable.*

(vi) *$\sigma(A)\cap\overline{\mathbb{C}}_+ = \emptyset$ and $\displaystyle\sup_{u\in\mathcal{S}(\mathbb{R},U)} \frac{\|\int_{\mathbb{R}} C(A_\alpha - is)^{-1}Bu(s)e^{is(\cdot)}\,ds\|_{L^p(\mathbb{R},Y)}}{\|\int_{\mathbb{R}} u(s)e^{is(\cdot)}\,ds\|_{L^p(\mathbb{R},U)}} < \infty$ for all $\alpha \geq 0$, and (4.4) is both stabilizable and detectable.*

Moreover, if $\{e^{tA}\}_{t\geq 0}$ is exponentially stable, then the norm of the input-output operator $\mathbb{L} = \mathcal{C}\mathbb{G}\mathcal{B}$ is equal to

$$\sup_{u\in\mathcal{S}(\mathbb{R},U)} \frac{\|\int_{\mathbb{R}} C(A - is)^{-1}Bu(s)e^{is(\cdot)}\,ds\|_{L^p(\mathbb{R},Y)}}{\|\int_{\mathbb{R}} u(s)e^{is(\cdot)}\,ds\|_{L^p(\mathbb{R},U)}}.$$

Proof. First note that the equivalence of statements (i) and (ii) follows from Remarks 3.3. Also, the implication (i)⇒(vi), as well as the last statement of the theorem, follow from Theorem 4.11. The exponential stability of $\{e^{tA}\}_{t\geq 0}$ is equivalent to the invertibility of $\Gamma_{\mathbb{R}}$ (Corollary 3.6), and so (iii) follows from (i) by Proposition 4.9.

To see that (iii) implies (i), begin by setting $\alpha = 0$. We wish to use properties of $\mathfrak{F}$ as in Proposition 4.8. We begin by observing that if the expression in (iii) is finite, then $\sup_{s\in\mathbb{R}}\|(A - is)^{-1}\| < \infty$. Indeed, if this were not the case, then there would exist $s_n \in \mathbb{R}$ and $x_n \in \mathrm{Dom}(A)$ with $\|x_n\| = 1$ such that $\|(A - is_n)x_n\| \to 0$ as $n \to \infty$. Choose functions $\beta_n \in \mathcal{S}(\mathbb{R})$ with the property that

$$\lim_{n\to\infty} \frac{\|\int_{\mathbb{R}} \beta_n(s)(is_n - is)e^{is(\cdot)}\,ds\|_{L^p(\mathbb{R})}}{\|\int_{\mathbb{R}} \beta_n(s)e^{is(\cdot)}\,ds\|_{L^p(\mathbb{R})}} = 0. \tag{5.5}$$

Note: to construct such a sequence of functions β_n, one takes, without loss of generality, $s_n = 0$ in (5.5) and chooses a "bump" function $\beta_0(s)$ where $\beta_0(0) = 1$ and β_0 has support in $(-1,1)$. Then set $\beta_n(s) = n\beta_0(ns)$. If $\check{}$ denotes the inverse Fourier transform, then $\check{\beta}_n(\tau) = \check{\beta}_0(\tau/n)$. Also, for $\alpha_n(s) = s\beta_n(s)$, one has $\check{\alpha}_n(\tau) = \frac{1}{n}\check{\alpha}_0(\tau/n)$, and so

$$\frac{\|\int_{\mathbb{R}} \beta_n(s)se^{is(\cdot)}\,ds\|^p_{L^p(\mathbb{R})}}{\|\int_{\mathbb{R}} \beta_n(s)e^{is(\cdot)}\,ds\|^p_{L^p(\mathbb{R})}} = \frac{\|\check{\alpha}_n\|^p}{\|\check{\beta}_n\|^p} = \frac{\left(\frac{1}{n}\right)^p\|\check{\alpha}_0\|^p}{\|\check{\beta}_0\|^p} \to 0 \quad \text{as } n \to \infty.$$

Now, setting $v_n(s) := \beta_n(s)(A - is)x_n$ gives a function v_n in $\mathcal{S}(\mathbb{R}, X)$ with the properties that $(A - is)^{-1}v_n(s) = \beta_n(s)x_n$. Thus

$$\begin{aligned}
\frac{\| \int_{\mathbb{R}} (A - is)^{-1} v_n(s) e^{is(\cdot)}\, ds\|}{\| \int_{\mathbb{R}} v_n(s) e^{is(\cdot)}\, ds\|} &= \frac{\| \int_{\mathbb{R}} \beta_n(s) x_n e^{is(\cdot)}\, ds\|}{\| \int_{\mathbb{R}} \beta_n(s)(A - is)x_n e^{is(\cdot)}\, ds\|} \\
&= \frac{\| \int_{\mathbb{R}} \beta_n(s) x_n e^{is(\cdot)}\, ds\|}{\| \int_{\mathbb{R}} \beta_n(s)(A - is_n)x_n e^{is(\cdot)}\, ds + \beta_n(s)(is_n - is)x_n e^{is(\cdot)}\, ds\|} \\
&\geq \frac{\| \int_{\mathbb{R}} \beta_n(s) x_n e^{is(\cdot)}\, ds\|}{\|(A - is_n)x_n\| \, \| \int_{\mathbb{R}} \beta_n(s) e^{is(\cdot)}\, ds\| + \| \int_{\mathbb{R}} \beta_n(s)(is_n - is)x_n e^{is(\cdot)}\, ds\|} \\
&= \left(\|(A - is_n)x_n\| + \frac{\| \int_{\mathbb{R}} \beta_n(s)(is_n - is) e^{is(\cdot)}\, ds\|}{\| \int_{\mathbb{R}} \beta_n(s) e^{is(\cdot)}\, ds\|} \right)^{-1}.
\end{aligned}$$

By the choice of s_n, x_n, and β_n, this last expression goes to ∞ as $n \to \infty$, contradicting (iii). Hence if the expression in (iii) is finite for $\alpha = 0$, then $\sup_{s\in\mathbb{R}} \|(A - is)^{-1}\| < \infty$.

Now we may apply Proposition 4.8(ii) and Proposition 4.9 to obtain

$$\|\Gamma_{\mathbb{R}}\|_\bullet = \inf_{v\in\mathcal{S}(\mathbb{R},X)} \frac{\|\Gamma_{\mathbb{R}} f_v\|}{\|f_v\|} = \inf_{v\in\mathcal{S}(\mathbb{R},X)} \frac{\|g_v\|}{\|f_v\|} = \left(\sup_{v\in\mathcal{S}(\mathbb{R},X)} \frac{\|f_v\|}{\|g_v\|} \right)^{-1} > 0.$$

This shows that $0 \notin \sigma_{ap}(\Gamma_{\mathbb{R}})$ and so, by [23], it follows that $\sigma_{ap}(e^{tA}) \cap \mathbb{T} = \emptyset$. On the other hand, since $\sigma(A) \cap i\mathbb{R} = \emptyset$, it follows from the spectral mapping theorem for the residual spectrum $\sigma_r(e^{tA})$ that

$$\sigma(e^{tA}) \cap \mathbb{T} = \big[\sigma_{ap}(e^{tA}) \cup \sigma_r(e^{tA})\big] \cap \mathbb{T} = \emptyset.$$

The same argument holds for any $\alpha \geq 0$. As a result, $\{e^{tA_\alpha}\}_{t\geq 0}$ is hyperbolic for each $\alpha \geq 0$, and thus $\{e^{tA}\}_{t\geq 0}$ is exponentially stable.

So far it has been shown that the statements (i)–(iii) are equivalent, and that statement (i) implies (vi). By showing that (vi)⇒(iv)⇒(iii) and (vi)⇒(v)⇒(iii), we complete the proof.

To see that (vi) implies (iv), begin by setting $\alpha = 0$. Since (4.4) is detectable, there exists $K \in \mathcal{L}(Y, X)$ such that $A + KC$ generates an exponentially stable semigroup. By the implication (i)⇒(iii) for the semigroup $\{e^{t(A+KC)}\}$, it follows that

$$M_1 := \sup_{v\in\mathcal{S}(\mathbb{R},X)} \frac{\| \int_{\mathbb{R}} (A + KC - is)^{-1} v(s) e^{is(\cdot)}\, ds\|}{\| \int_{\mathbb{R}} v(s) e^{is(\cdot)}\, ds\|}$$

is finite. So

(5.6)

$$\sup_{u\in\mathcal{S}(\mathbb{R},U)} \frac{\| \int_{\mathbb{R}} (A + KC - is)^{-1} B u(s) e^{is(\cdot)}\, ds\|}{\| \int_{\mathbb{R}} u(s) e^{is(\cdot)}\, ds\|}$$

$$= \sup_{u\in\mathcal{S}(\mathbb{R},U)} \frac{\| \int_{\mathbb{R}} (A + KC - is)^{-1} B u(s) e^{is(\cdot)}\, ds\|}{\| \int_{\mathbb{R}} B u(s) e^{is(\cdot)}\, ds\|} \cdot \frac{\|B \int_{\mathbb{R}} u(s) e^{is(\cdot)}\, ds\|}{\| \int_{\mathbb{R}} u(s) e^{is(\cdot)}\, ds\|} \leq M_1 \|B\|.$$

By hypothesis in (vi),

$$M_2 := \sup_{u\in\mathcal{S}(\mathbb{R},U)} \frac{\| \int_{\mathbb{R}} C(A - is)^{-1} B u(s) e^{is(\cdot)}\, ds\|}{\| \int_{\mathbb{R}} u(s) e^{is(\cdot)}\, ds\|}$$

is finite. For $u \in \mathcal{S}(\mathbb{R}, U)$, let $w(s) = KC(A - is)^{-1}Bu(s),\ s \in \mathbb{R}$. Then

$$\frac{\| \int_{\mathbb{R}} (A + KC - is)^{-1} KC(A - is)^{-1} Bu(s) e^{is(\cdot)}\, ds\|}{\| \int_{\mathbb{R}} u(s) e^{is(\cdot)}\, ds\|} \tag{5.7}$$

$$= \frac{\| \int_{\mathbb{R}} (A + KC - is)^{-1} w(s) e^{is(\cdot)}\, ds\|}{\| \int_{\mathbb{R}} w(s) e^{is(\cdot)}\, ds\|} \cdot \frac{\|K \int_{\mathbb{R}} C(A - is)^{-1} Bu(s) e^{is(\cdot)}\, ds\|}{\| \int_{\mathbb{R}} u(s) e^{is(\cdot)}\, ds\|}$$

$$\leq M_1 \|K\| M_2.$$

Finally, since

$$(A - is)^{-1} B = (A + KC - is)^{-1} B + (A + KC - is)^{-1} KC(A - is)^{-1} B,$$

it follows from (5.6) and (5.7) that

$$\sup_{u \in \mathcal{S}(\mathbb{R}, U)} \frac{\| \int_{\mathbb{R}} (A - is)^{-1} Bu(s) e^{is(\cdot)}\, ds\|}{\| \int_{\mathbb{R}} u(s) e^{is(\cdot)}\, ds\|} \leq M_1 \|B\| + M_1 \|K\| M_2.$$

This argument holds for all $\alpha \geq 0$, so the implication (vi)⇒(iv) follows.

To see that (iv) implies (iii), we again argue only in the case $\alpha = 0$. Since (4.4) is stabilizable, there exists $F \in B(X, U)$ such that $A + BF$ generates an exponentially stable semigroup. By the implication (i)⇒(iii) for the semigroup $\{e^{t(A+BF)}\}$, it follows that

$$M_3 := \sup_{v \in \mathcal{S}(\mathbb{R}, X)} \frac{\| \int_{\mathbb{R}} (A + BF - is)^{-1} v(s) e^{is(\cdot)}\, ds\|}{\| \int_{\mathbb{R}} v(s) e^{is(\cdot)}\, ds\|}$$

is finite. By the hypotheses in (iv),

$$M_4 := \sup_{u \in \mathcal{S}(\mathbb{R}, U)} \frac{\| \int_{\mathbb{R}} (A - is)^{-1} Bu(s) e^{is(\cdot)}\, ds\|}{\| \int_{\mathbb{R}} u(s) e^{is(\cdot)}\, ds\|}$$

is finite. For $v \in \mathcal{S}(\mathbb{R}, X)$, set $w(s) = F(A + BF - is)^{-1} v(s),\ s \in \mathbb{R}$. Then

$$\sup_{v \in \mathcal{S}(\mathbb{R}, X)} \frac{\| \int_{\mathbb{R}} (A - is)^{-1} BF(A + BF - is)^{-1} v(s) e^{is(\cdot)}\, ds\|}{\| \int_{\mathbb{R}} v(s) e^{is(\cdot)}\, ds\|} \tag{5.8}$$

$$= \frac{\| \int_{\mathbb{R}} (A - is)^{-1} Bw(s) e^{is(\cdot)}\, ds\|}{\| \int_{\mathbb{R}} w(s) e^{is(\cdot)}\, ds\|} \cdot \frac{\|F \int_{\mathbb{R}} (A + BF - is)^{-1} v(s) e^{is(\cdot)}\, ds\|}{\| \int_{\mathbb{R}} v(s) e^{is(\cdot)}\, ds\|}$$

$$\leq M_4 \|F\| M_3.$$

Since

$$(A - is)^{-1} = (A + BF - is)^{-1} + (A - is)^{-1} BF(A + BF - is)^{-1},$$

it follows from (5.8) that

$$\sup_{v \in \mathcal{S}(\mathbb{R}, X)} \frac{\| \int_{\mathbb{R}} (A - is)^{-1} v(s) e^{is(\cdot)}\, ds\|}{\| \int_{\mathbb{R}} v(s) e^{is(\cdot)}\, ds\|} \leq M_3 + M_4 \|F\| M_3.$$

Thus (iii) follows from (iv). Similar arguments show that (vi)⇒(v) and that (v)⇒(iii). □

From the equivalence of statements (i) and (iii), it follows that the growth bound of a semigroup on a Banach space is given by

$$\omega_0(e^{tA}) = \inf\left\{\alpha \in \mathbb{R} : \sup_{v\in\mathcal{S}(\mathbb{R},X)} \frac{\|\int_{\mathbb{R}}(A_\alpha - is)^{-1}v(s)e^{is(\cdot)}\,ds\|}{\|\int_{\mathbb{R}} v(s)e^{is(\cdot)}\,ds\|} < \infty\right\}.$$

This is a natural generalization of the formula for the growth bound for a semigroup on a Hilbert space as provided by Gearhart's theorem (see [19, 2, 29, 33] and cf. Theorem 2.1):

$$\omega_0(e^{tA}) = s_0(A) = \inf\left\{\alpha \in \mathbb{R} : \sup_{\mathrm{Re}\lambda\geq\alpha} \|(A-\lambda)^{-1}\| < \infty\right\}.$$

THEOREM 5.5. *Let $\{e^{tA}\}_{t\geq 0}$ be a strongly continuous semigroup on a Banach space X with the property that $s_0(A) = \omega_0(e^{tA})$. Assume (4.4) is stabilizable and detectable. If $\overline{\mathbb{C}}_+ \subset \rho(A)$ and $M := \sup_{s\in\mathbb{R}} \|C(A-is)^{-1}B\| < \infty$, then $\{e^{tA}\}_{t\geq 0}$ is exponentially stable.*

Proof. Choose operators $F \in \mathcal{L}(X,U)$ and $K \in \mathcal{L}(Y,X)$ such that the semigroups generated by $A+BF$ and $A+KC$ are exponentially stable. Then $s_0(A+BF) < 0$ and $s_0(A+KC) < 0$, and so

$$M_1 := \sup_{s\in\mathbb{R}} \|(A+BF-is)^{-1}\| \quad \text{and} \quad M_2 := \sup_{s\in\mathbb{R}} \|(A+KC-is)^{-1}\|$$

are both finite. Since

$$(A-is)^{-1}B = (A+KC-is)^{-1}B + (A+KC-is)^{-1}KC(A-is)^{-1}B,$$

it follows that

$$M_3 := \sup_{s\in\mathbb{R}} \|(A-is)^{-1}B\| \leq M_2\|B\| + M_2\|K\|M.$$

Also,

$$(A-is)^{-1} = (A+BF-is)^{-1} + (A-is)^{-1}BF(A+BF-is)^{-1},$$

and so

$$\sup_{s\in\mathbb{R}} \|(A-is)^{-1}\| \leq M_1 + M_3\|F\|M_1.$$

Therefore, $\omega_0(e^{tA}) = s_0(A) < 0$. □

The following result, based on [22], describes a particular situation in which $s_0(A) = \omega_0(e^{tA})$.

COROLLARY 5.6. *Assume that for the generator A of a strongly continuous semigroup $\{e^{tA}\}_{t\geq 0}$ on a Banach space X there exists an $\omega > \omega_0(e^{tA})$ such that*

$$\int_{-\infty}^{\infty} \|(\omega + i\tau - A)^{-1}x\|_X^2\,d\tau < \infty \quad \textit{for all} \quad x \in X, \tag{5.9}$$

and

$$\int_{-\infty}^{\infty} \|(\omega + i\tau - A^*)^{-1} x^*\|_{X^*}^2 \, d\tau < \infty \quad \textit{for all} \quad x^* \in X^*, \tag{5.10}$$

where X^* is the adjoint space. Then system (4.4) *is internally stable if and only if it is stabilizable, detectable, and externally stable.*

Proof. According to [22] (see also [29, Cor. 4.6.12]), conditions (5.9)–(5.10) imply $s_0(A) = \omega_0(e^{tA})$. Now Theorem 5.5 gives the result. ∎

REFERENCES

[1] H. Amann, *Operator-valued Fourier multipliers, vector-valued Besov spaces, and applications*, Math. Nachr., 186 (1997), pp. 5–56.

[2] W. Arendt, A. Grabosch, G. Greiner, U. Groh, H. P. Lotz, U. Moustakas, R. Nagel, F. Neubrander, and U. Schlotterbeck, *One Parameter Semigroups of Positive Operators*, Lecture Notes in Math. 1184, Springer-Verlag, Berlin, 1986.

[3] B. D. O. Anderson, *Internal and external stability of linear time-varying systems*, SIAM J. Control Optim., 20 (1982), pp. 408–413.

[4] J. Ball, I. Gohberg, and M. A. Kaashoek, *A frequency response function for linear, time-varying systems*, Math. Control Signals Systems, 8 (1995), pp. 334–351.

[5] J. A. Burns and B. B. King, *A note on the regularity of solutions of infinite dimensional Riccati equations*, Appl. Math. Lett., 7 (1994), pp. 13–17.

[6] C. Buse, *On the Perron–Bellman theorem for evolutionary process with exponential growth in Banach spaces*, New Zealand J. Math., 27 (1998), pp. 183–190.

[7] C. Chicone and Y. Latushkin, *Evolution Semigroups in Dynamical Systems and Differential Equations*, Math. Surveys Monogr. 70, AMS, Providence, RI, 1999.

[8] R. Curtain, *Equivalence of input-output stability and exponential stability for infinite-dimensional systems*, Math. Systems Theory, 21 (1988), pp. 19–48.

[9] R. Curtain, *Equivalence of input-output stability and exponential stability*, Systems Control Lett., 12 (1989), pp. 235–239.

[10] R. Curtain and A. J. Pritchard, *Infinite Dimensional Linear System Theory*, Lecture Notes in Control and Inform. Sci. 8, Springer-Verlag, New York, 1978.

[11] R. Curtain and H. J. Zwart, *An Introduction to Infinite-Dimensional Linear Systems Theory*, Springer-Verlag, New York, 1995.

[12] J. Daleckij and M. Krein, *Stability of Differential Equations in Banach Space*, AMS, Providence, RI, 1974.

[13] R. Datko, *Uniform asymptotic stability of evolutionary processes in a Banach space*, SIAM J. Math. Anal., 3 (1972), pp. 428–445.

[14] A. Fischer and J. M. A. M. van Neerven, *Robust stability of C_0-semigroups and an application to stability of delay equations*, J. Math. Anal. Appl., 226 (1998), pp. 82–100.

[15] J. Hale, *Ordinary Differential Equations*, Wiley-Interscience, New York, 1969.

[16] D. Hinrichsen, A. Ilchmann, and A. J. Pritchard, *Robustness of stability of time-varying linear systems*, J. Differential Equations, 82 (1989), pp. 219–250.

[17] D. Hinrichsen and A. J. Pritchard, *Stability radius for structured perturbations and the algebraic Riccati equation*, Systems Control Lett., 8 (1986), pp. 105–113.

[18] D. Hinrichsen and A. J. Pritchard, *Robust stability of linear evolution operators on Banach spaces*, SIAM J. Control Optim., 32 (1994), pp. 1503–1541.

[19] F. Huang, *Characteristic conditions for exponential stability of linear dynamical systems in Hilbert spaces*, Ann. Differential Equations, 1 (1985), pp. 45–53.

[20] B. Jacob, V. Dragan, and A. J. Pritchard, *Infinite-dimensional time-varying systems with nonlinear output feedback*, Integral Equations Operator Theory, 22 (1995), pp. 440–462.

[21] C. A. Jacobson and C. N. Nett, *Linear state space systems in infinite-dimensional space: The role and characterization of joint stabilizability/detectability*, IEEE Trans. Automat. Control, 33 (1988), pp. 541–550.

[22] M. A. Kaashoek and S. M. Verduyn Lunel, *An integrability condition for hyperbolicity of the semigroup*, J. Differential Equations, 112 (1994), pp. 374–406.

[23] Y. Latushkin and S. Montgomery-Smith, *Evolutionary semigroups and Lyapunov theorems in Banach spaces*, J. Funct. Anal., 127 (1995), pp. 173–197.

[24] Y. Latushkin, S. Montgomery-Smith, and T. Randolph, *Evolutionary semigroups and di-*

chotomy of linear skew-product flows on locally compact spaces with Banach fibers, J. Differential Equations, 125 (1996), pp. 73–116.

[25] Y. Latushkin and T. Randolph, *Dichotomy of differential equations on Banach spaces and an algebra of weighted composition operators*, Integral Equations Operator Theory, 23 (1995), pp. 472–500.

[26] H. Logemann, *Stabilization and regulation of infinite-dimensional systems using coprime factorizations*, in Analysis and Optimization of Systems: State and Frequency Domain Approaches for Infinite-Dimensional Systems (Sophia-Antipolis, 1992), R. F. Curtain, A. Bensoussan, and J.-L. Lions, eds., Lecture Notes in Control and Inform. Sci. 185, Springer-Verlag, Berlin, 1993, pp. 102–139.

[27] N. van Minh, R. Räbiger, and R. Schnaubelt, *Exponential stability, exponential expansiveness, and exponential dichotomy of evolution equations on the half-line*, Integral Equations Operator Theory, 32 (1998), pp. 332–353.

[28] J. M. A. M. van Neerven, *Characterization of exponential stability of a semigroup of operators in terms of its action by convolution on vector-valued function spaces over* $\mathbb{R}_+$, J. Differential Equations, 124 (1996), pp. 324–342.

[29] J. M. A. M. van Neerven, *The Asymptotic Behavior of a Semigroup of Linear Operators*, Oper. Theory Adv. Appl. 88, Birkhäuser, Basel, 1996.

[30] A. Pazy, *Semigroups of Linear Operators and Applications to Partial Differential Equations*, Springer-Verlag, New York, Berlin, 1983.

[31] R. S. Phillips, *Perturbation theory for semi-groups of linear operators*, Trans. Amer. Math. Soc., 74 (1953), pp. 199–221.

[32] A. J. Pritchard and S. Townley, *Robustness of linear systems*, J. Differential Equations, 77 (1989), pp. 254–286.

[33] J. Prüss, *On the spectrum of* C_0*-semigroups*, Trans. Amer. Math. Soc., 284 (1984), pp. 847–857.

[34] F. Räbiger and R. Schnaubelt, *The spectral mapping theorem for evolution semigroups on spaces of vector-valued functions*, Semigroup Forum, 52 (1996), pp. 225–239.

[35] F. Räbiger, A. Rhandi, and R. Schnaubelt, *Perturbation and an abstract characterization of evolution semigroups*, J. Math. Anal. Appl., 198 (1996), pp. 516–533.

[36] F. Räbiger, A. Rhandi, R. Schnaubelt, and J. Voigt, *Non-autonomous Miyadera perturbations*, Differential Integral Equations, 13 (1999), pp 341–368.

[37] T. Randolph, Y. Latushkin, and S. Clark, *Evolution semigroups and stability of time-varying systems on Banach spaces*, in Proceedings of the 36th IEEE Conference on Decision and Control, San Diego, CA, 1997, pp. 3932–3937.

[38] R. Rau, *Hyperbolic evolution semigroups on vector valued function spaces*, Semigroup Forum, 48 (1994), pp. 107–118.

[39] R. Rebarber, *Conditions for the equivalence of internal and external stability for distributed parameter systems*, IEEE Trans. Automat. Control, 31 (1993), pp. 994–998.

[40] R. Rebarber, *Frequency domain methods for proving the uniform stability of vibrating systems*, in Analysis and Optimization of Systems: State and Frequency Domain Approaches for Infinite-Dimensional Systems (Sophia-Antipolis, 1992), R. F. Curtain, A. Bensoussan, and J.-L. Lions, eds., Lecture Notes in Control and Inform. Sci. 185, Springer-Verlag, Berlin, 1993, pp. 366–377.

[41] M. Renardy, *On the linear stability of hyperbolic PDEs and viscoelastic flows*, Z. Angew. Math. Phys., 45 (1994), pp. 854–865.

[42] R. Saeks and G. Knowles, *The Arveson frequency response and systems theory*, Internat. J. Control, 42 (1985), pp. 639–650.

[43] R. Schnaubelt, *Exponential Bounds and Hyperbolicity of Evolution Families*, Ph.D. Thesis, Eberhard-Karls-Univeristät, Tübingen, Germany, 1996.

[44] G. Weiss, *Transfer functions of regular linear systems, part* I: *Characterizations of regularity*, Trans. Amer. Math. Soc., 342 (1994), pp. 827–854.

[45] G. Weiss, *Representation of shift invariant operators on* L^2 *by* H^∞ *transfer functions: An elementary proof, a generalization to* L^p *and a counterexample for* L^∞, Math. Control Signals Systems, 4 (1991), pp. 193–203.

SIAM J. CONTROL OPTIM.
Vol. 38, No. 6, pp. 1794–1804

SOLVABLE STATES IN n-PLAYER STOCHASTIC GAMES*

NICOLAS VIEILLE†

Abstract. We prove that, in every stochastic game with finitely many states and actions, there exists at least one state, starting from which an equilibrium payoff exists. This is achieved by proving that there exists a solvable set. This generalizes to an arbitrary number of players a result due to Thuijsman and Vrieze in the case of two players.

Key words. stochastic games, nonzero-sum games, equilibrium payoff

AMS subject classifications. 91A06, 91A15

PII. S0363012998345937

1. Introduction.

We prove that, in every stochastic game with finitely many states and actions, there exists at least one state, starting from which an equilibrium payoff exists. An associated ϵ-equilibrium profile consists of playing essentially like some stationary profile x, sustained by appropriate threats. More precisely, the players cycle periodically over some ergodic sets of the Markov chain induced by x and travel between these sets by using small perturbations of x.

This result is a generalization to an arbitrary number of players of a result due to Thuijsman and Vrieze [6] (see also [8]) in the case of two players. For two players, the analog result turned out to be the first step in the proof of existence of equilibrium payoffs for general two-player games (cf. [9]).

The proof is much more involved than for two players. Usually, states are compared by means of the (undiscounted) min max values. This is not enough here. We order the states lexicographically, using *discounted* min max values. We argue that the set of states with highest min max values contains a solvable set. As far as I know, this is the first result on nonzero-sum games for which this refined comparison between states turns out to be important (a similar comparison was used by Tijs and Vrieze [7] for two-player, zero-sum games).

This has the consequence that in every stochastic game, there exists at least one initial state for which an equilibrium payoff does exist.

Section 2 is devoted to the model and the statement of the result. Section 3 contains some additional definitions and a refined statement of the result. The existence of a solvable set is established in section 5.

2. Model and results.

2.1. Rules of the games.

For every finite set S, $\Delta(S)$ is the simplex of probability distributions over S; for $\mu \in \Delta(S)$ and $s \in S$, the probability of s is $\mu(s)$. The set of players is $\mathcal{I} = \{1, \dots, I\}$, with $I \in \mathbf{N}$.

A stochastic game is described by a finite set of states K, the finite set A^i of actions available to player $i \in \mathcal{I}$ in each state k, a transition function $\mathbf{q} : K \times A \to \Delta(K)$, with $A = \prod_i A^i$, which describes the evolution of the state of the game, and a function $g : K \times A \to \mathbf{R}^I$ which describes the payoff received by the players in any stage.

*Received by the editors October 5, 1998; accepted for publication (in revised form) September 7, 1999; published electronically June 15, 2000.
http://www.siam.org/journals/sicon/38-6/34593.html

†Ecole Polytechnique, Laboratoire D'Économétrie, 1 rue Descartes, 75005 Paris, France (vieille@poly.polytechnique.fr).

The game is played as follows. The set of stages is the set $\mathbf{N}^*$ of positive integers. The initial state k_1 is given. In stage n, the current state k_n is announced to the players. Each player i chooses an action $a_n^i \in A^i$; the action combination $a_n = (a_n^i)_{i\in\mathcal{I}}$ is publicly announced, k_{n+1} is drawn according to $\mathbf{q}(\cdot|k_n, a_n)$, and the game proceeds to stage $n+1$.

We denote by $H_\infty = (K \times A)^{\mathbf{N}}$ the set of plays and by $H_n = (K \times A)^{n-1} \times K$ the set of histories up to stage n. A *strategy* of player i is a sequence $\sigma^i = (\sigma_n^i)_{n\in\mathbf{N}^*}$, where $\sigma_n^i : H_n \to \Delta(A^i)$ describes the behavior of player i in stage n. The set of strategies of player i is denoted by Σ^i, and we set $\Sigma = \prod_{i\in\mathcal{I}} \Sigma^i$.

We denote by $\mathcal{H}_n$ the algebra of cylinder sets over H_∞, induced by H_n, and we set $\mathcal{H}_\infty = \sigma(\mathcal{H}_n, n \in \mathbf{N}^*)$.

We denote by $\mathbf{P}_{k,\sigma}$ the probability distribution over $(H_\infty, \mathcal{H}_\infty)$, when the initial state is k, and the players use the strategies σ. Expectations with respect to $\mathbf{P}_{k,\sigma}$ are denoted by $\mathbf{E}_{k,\sigma}$.

The mixed extension of $\mathbf{q}$ to $K \times \prod_i \Delta(A^i)$ is still denoted by $\mathbf{q}$. We will follow standard use and denote by $-i$ the *coalition* of players other than i: we set $A^{-i} = \prod_{j\neq i} A^j$, $\Sigma^{-i} = \prod_{j\neq i} \Sigma^i$, and we use similar shortcuts whenever needed.

All norms used in the paper are *supremum* norms. We assume for simplicity that $\|g\| \leq 1$. The cardinality of a finite set A is denoted by $|A|$. Finally, for $P \in \Delta(K)$ and $f : K \to \mathbf{R}$, we denote either by Pf or $P[f]$ the expectation of f under P.

2.2. min max and equilibrium payoffs. We denote by $g_n = g(k_n, a_n) \in \mathbf{R}^I$ the payoff in stage n and by

$$\gamma_n(k, \sigma) = \mathbf{E}_{k,\sigma}\left[\frac{1}{n}\sum_{p=1}^{n} g_p\right]$$

the expected average payoff up to stage n, when the initial state is k, and the players use $\sigma \in \Sigma$. We set $\gamma_n(\sigma) = (\gamma_n(k, \sigma))_{k\in K}$, and we will use similar vector notations whenever convenient.

As often happens in stochastic games, punishment strategies will play an important role. Recall that $v^i \in \mathbf{R}^K$ is the min max for player i if (see [3]) the following hold.

1. Players $-i$ can guarantee v^i: for every $\epsilon > 0$, there exists $\sigma_\epsilon^{-i} \in \Sigma^{-i}$ and a stage $N_0 \in \mathbf{N}$, such that
$$\forall \sigma^i \in \Sigma^i, n \geq N_0, \gamma_n^i(\sigma^i, \sigma_\epsilon^{-i}) \leq v^i + \epsilon.$$
2. Player i can defend v^i: for every $\epsilon > 0$, there exists a stage $N_0 \in \mathbf{N}$, such that for every $\sigma^{-i} \in \Sigma^{-i}$,
$$\exists \sigma^i \in \Sigma^i, \text{ such that } \forall n \geq N_0, \gamma_n^i(\sigma^i, \sigma^{-i}) \geq v^i - \epsilon.$$

The first condition says that, by playing (σ_ϵ^{-i}), players other than i ensure that player i's average payoffs will not exceed v^i. The second says that player i always has a reply which keeps his average payoff above v^i. Therefore, $v^i(k)$ is the punishment level that players $-i$ may inflict on player i, starting from k.

The existence of the min max in this framework has been proven by Neyman [4], extending the proof of Mertens and Neyman [2].

As usual, we say that $\sigma_* \in \Sigma$ is an ϵ-equilibrium profile in the n-stage game if for every $k \in K$, $i \in \mathcal{I}$, and $\sigma^i \in \Sigma^i$,

$$\gamma_n^i(k, \sigma^i, \sigma_*^{-i}) \leq \gamma_n^i(k, \sigma_*) + \epsilon.$$

We say that σ_* is an ϵ-equilibrium profile if it is an ϵ-equilibrium profile in the n-stage game for every n large enough.

Finally, we define equilibrium payoffs.

DEFINITION 1. *Let $d \in (\boldsymbol{R}^I)^K$. d is an equilibrium payoff if, for every $\varepsilon > 0$, there exists an ε-equilibrium profile σ^* and an integer N_0 such that*

$$||d - \gamma_n(\sigma^*)|| < \varepsilon \quad \forall n \geq N_0.$$

The set of equilibrium payoffs of Γ is denoted by $E(\Gamma)$. It is clear how to specialize the above definitions to a *given* initial state. We denote $E_k(\Gamma)$ the set of equilibrium payoffs of the game Γ, with initial state k. Notice that $E_k(\Gamma) \subset \mathbf{R}^I$ and that $E(\Gamma) = \prod_{k\in K} E_k(\Gamma)$.

Our basic result is the following.

THEOREM 2. *There exists $k \in K$, with $E_k(\Gamma) \neq \emptyset$.*

We actually prove a more precise result, stated below (Proposition 9).

3. Preliminaries. In this section, we define solvable states and give in Proposition 7 an improved statement of Theorem 2.

3.1. Stationary strategies and communication. A strategy of player i is *stationary* if the behavior of player i depends only upon the *current* state. Thus a stationary strategy can be identified to a vector $x^i = (x^i_k)$, where $x^i_k \in \Delta(A^i)$ is the lottery used by player i whenever the current state is k. We denote by $\mathcal{S}^i (= \Delta(A^i)^K)$ the set of stationary strategies of player i and set $\mathcal{S} = \prod_i \mathcal{S}^i$.

For $x^i_k \in \Delta(A^i)$, $\mathrm{supp}(x^i_k) = \{a^i \in A^i, x^i_k(a^i) > 0\}$ is the support of x^i_k. For $x^i = (x^i_k)_{k\in K} \in \mathcal{S}^i$, we set $\mathrm{supp}(x^i) = \prod_k \mathrm{supp}(x^i_k)$.

For every $x \in \mathcal{S}$, the sequence $(k_n)_{n\in\mathbf{N}^*}$ is a Markov chain under $\mathbf{P}_{k,x}$. Hence $\gamma(x) = \lim_{n\to\infty} \gamma_n(x)$ exists. For simplicity, we shall speak of ergodic sets for x, instead of ergodic sets for the Markov chain induced by x, and we shall use similar shortcuts whenever convenient. If R is an ergodic set for x, then $\gamma(k, x)$ is independent of $k \in R$. We simply write $\gamma_R(x)$.

We define now *communicating* sets, which generalize *ergodic* sets. A *perturbation* of $x \in \mathcal{S}$ is a profile $\tilde{x} \in \mathcal{S}$ such that $\mathrm{supp}(x) \subseteq \mathrm{supp}(\tilde{x})$.

Let $x \in \mathcal{S}$, and let $k, k' \in K$. A finite sequence $(k_0 = k, k_1, \ldots, k_N = k')$ in K is a path from k to k' under x if, for any $n \in \{0, \ldots, N-1\}$, $\mathbf{q}(k_{n+1}|k_n, x_{k_n}) > 0$. k *leads to* k' under x if there exists a path from k to k'.

A subset E of K is closed under x if $\mathbf{q}(E|k, x_k) = 1$ for every $k \in E$.

DEFINITION 3. *Let $x \in \mathcal{S}$, $E \subseteq K$. E is a communicating set for x if, for any $k, k' \in E$, there exists a perturbation $\tilde{x}$ of x such that, under $\tilde{x}$, E is closed and k leads to k'.*

This captures the idea that the players are able to visit an infinite number of times any given state in E, without leaving E.

DEFINITION 4. *$\mathcal{C}(x)$ is the set of communicating sets for x.*

Remark. Of course, any set which is ergodic for x is a communicating set for x. More generally, let $x, x' \in \mathcal{S}$ be such that $\mathrm{supp}(x) \subseteq \mathrm{supp}(x')$. Every ergodic set for x' is a communicating set for x.

3.2. Solvable sets. Let $x \in \mathcal{S}$, and let $E \in \mathcal{C}(x)$. We denote by $\mathcal{R}_E(x)$ the set of subsets of E which are ergodic for x. Since a communicating set for x is closed under x, $\mathcal{R}_E(x) \neq \emptyset$.

DEFINITION 5. *Let $x \in \mathcal{S}$, $E \in \mathcal{C}(x)$, and $\mu \in \Delta(\mathcal{R}_E(x))$. The 3-tuple (E, x, μ) is solvable if for every $k \in E, i \in \mathcal{I}, a^i \in A^i$, one has*

$$\mathbf{q}(\cdot|k, a^i, x_k^{-i})v^i \leq \sum_{R \in \mathcal{R}_E(x)} \mu(R)\gamma_R^i(x).$$

It is convenient to say simply that E is solvable. We refer to the quantity

$$\sum_{R \in \mathcal{R}_E(x)} \mu(R)\gamma_R^i(x)$$

as the solvable payoff of E. It follows from the definition that the solvable payoff is at least $\max_{k \in E} v(k)$. We shall prove Propositions 6 and 7, stated below. They obviously imply Theorem 2.

PROPOSITION 6. *Let (E, x, μ) be a solvable set. For $k \in E, \sum_R \mu(R)\gamma_R(x) \in E_k(\Gamma)$.*

PROPOSITION 7. *There exists a solvable set.*

The notion of solvable sets that is used here for n-player games is the generalization of the one that proved most useful in the analysis of two-player games (see [9]). It turns out that the proof of Proposition 7 we give below establishes the existence of *easy initial sets.*

DEFINITION 8. *Let $E \subseteq K$. The set E is an easy initial set if, for every $\varepsilon > 0$, there exists a stationary profile x such that E is an ergodic set for x and, for every $k \in E, i \in \mathcal{I}, a^i \in A^i$, one has*

$$\mathbf{q}(\cdot|k, a^i, x_k^{-i})v^i - \varepsilon \leq \gamma_E^i(x).$$

This notion is due to Thuijsman and Vrieze [6]. It is not difficult to check that easy initial sets are solvable. Therefore, the existence of easy initial sets is a stronger result than the existence of solvable sets. Also, replacing solvable sets by easy initial sets in the statement of Proposition 6 (and modifying the rest of the statement accordingly) would allow for a slightly easier proof. We nevertheless deal here with the notion of solvable sets, since this notion is slightly easier to use in two-player games. This hints that it might also be the case for n-player games.

3.3. Reminder on discounted games. We recall here some known facts about discounted games. For $\lambda \in (0, 1], \sigma \in \Sigma$, and $k \in K$, we denote by

$$\gamma_\lambda(k, \sigma) = E_{k,\sigma}\left[\lambda \sum_{n=1}^{\infty} (1-\lambda)^{n-1} g_n\right]$$

the discounted payoff induced by σ, when the initial state is k.

For $i \in \mathcal{I}$, we also denote by $v_\lambda^i(k) = \min_{\Sigma^{-i}} \max_{\Sigma^i} \gamma_\lambda^i(k, \sigma^{-i}, \sigma^i)$ the discounted min max value for player i.

For $\lambda \in (0, 1]$, $i \in \mathcal{I}$, and $x \in \mathcal{S}$, we define an operator $U_{\lambda,x}^i$ over functions $u : K \to \mathbf{R}$ by

$$U_{\lambda,x}^i u(k) = \lambda g^i(k, x_k) + (1-\lambda)\mathbf{q}(\cdot|k, x_k)u.$$

Let $e \in \mathbf{R}^K$ be the vector $(1, 1, \ldots, 1)$. The following two facts are well known.

Fact 1. For $u : K \to \mathbf{R}, c \in \mathbf{R}, x \in \mathcal{S}$, and $\lambda \in (0, 1]$,

$$U_{\lambda,x}^i u \geq u - c \cdot e \Rightarrow \gamma_\lambda^i(x) \geq u - \frac{c}{\lambda}.$$

Indeed, $U^i_{\lambda,x}u \geq u - c\cdot e$ yields inductively $(U^i_{\lambda,x})^n u \geq u - c(\frac{1-(1-\lambda)^n}{\lambda})\cdot e$. The right-hand side property follows, since $\gamma^i_\lambda(x) = \lim_{n\to\infty}(U^i_{\lambda,x})^n u$.

Fact 2 (see Shapley [5]). For every $x^{-i} \in \mathcal{S}^{-i}$, there exists $x^i \in \mathcal{S}^i$, such that

$$U^i_{\lambda,(x^{-i},x^i)} v^i_\lambda \geq v^i_\lambda.$$

It is known (cf. Bewley and Kohlberg [1] and Neyman [4]) that $v^i = \min_{\lambda\to 0} v^i_\lambda$ and that, for each $i \in \mathcal{I}$ and $k \in K$, the function $\lambda \mapsto v^i_\lambda(k)$ has a Puiseux expansion in a neighborhood of 0: there exists $\lambda_0 > 0, M \in \mathbf{N}, N \in \mathbf{Z}$, and a sequence of real numbers $a^i_p(k)$, $p \in \mathbf{Z}$, such that

$$\forall \lambda < \lambda_0,\ v^i_\lambda(k) = \sum_{p=N}^{+\infty} a^i_p(k)\lambda^{\frac{p}{M}}.$$

Since payoffs are bounded, $a^i_p(k) = 0$ for $p < 0$.

4. Solvable payoffs are equilibrium payoffs. This lemma is proven in Vieille [9] in the case of two players. We briefly sketch the extension to an arbitrary number of players.

We describe ϵ-equilibrium profiles associated to $\sum_R \mu(R)\gamma_R(x)$. The players visit cyclically the elements of $\mathcal{R}_E(x)$ in a fixed order $R_1, \ldots, R_L$, without ever leaving E. In each visit to R_l, they stick to x during N_l stages. Going from one element R_l to the following one R_{l+1} is feasible using small perturbations of x, since E is a communicating set for x. For each l, we choose a stationary profile $x(l)$, such that E is closed under $x(l)$, and R_{l+1} is the only subset of E which is ergodic for $x(l)$. Moreover, we require that $x(l)$ be close to x. The length N_l of the visits to R_l is chosen to be proportional to $\mu(R_l)$ and sufficiently large so that (i) the payoff during one visit to R_l is close to $\gamma_{R_l}(x)$, and (ii) the expected time between two visits is small compared to the length of the visits.

This behavior is supported by tests, designed as follows. Players need to enforce not only that each player i plays roughly according to x^i during the visits to the different sets, but also that player i indeed plays $x^i(l)$ when the profile asks for going from one ergodic set R_l to the following one R_{l+1}. On the equilibrium path, there will be infinitely many cycles and hence infinitely many excursions from R_l to R_{l+1}. Denote by F_l the smallest subset of E, which contains R_l and is closed for $x(l)$. On the equilibrium path, only states in F_l can be reached during an excursion from R_l to R_{l+1}. Moreover, each of the states in F_l will be visited infinitely many times during these excursions. These observations allow for standard statistical checking of the play of each player, based on the empirical distribution of moves. Details are standard.

As soon as one of the statistical tests is failed by some player, say by player i, the others switch to a *punishment profile* which brings player i's future payoffs down to v^i.

Remark. If one did replace min max values by max min values in the definition of a solvable set, the proof of existence would be much simpler than the one below. However, the above scenario does not work, since there is no punishment profile which brings player i's payoffs down to the max min.

Remark. In the two-player case, simpler ε-equilibrium profiles can be designed. Checking of the opponent's play can be reduced to checking that its average payoff remains (after sufficiently many stages) close to its theoretical payoff. This is not true for more than two-player games, since it might then be in player i's interest to deviate

in such a way as to increase player j's payoff ($j \neq i$), in order to benefit from the punishment of player j.[1]

Remark. In the case of two players, Proposition 6 can be extended. Assume that perfect monitoring does not hold, and that, prior to playing in stage n, each player is only told the current state k_n and the payoff vector obtained in the previous stage g_{n-1}. There might exist profitable deviations from the above scenario. However, this is not the case (hence the scenario describes an ϵ-equilibrium profile) if the following additional assumptions are satisfied.

1. v is constant over E.
2. For each $R \in \mathcal{R}_E(x)$, the perturbation $\tilde{x}$ of x which is used to reach R can be chosen in such a way that for every $k \in E, i \in \mathcal{I}$, and $a^i \in A^i$,

$$\mathbf{q}(\cdot|k, a^i, \tilde{x}_k^{-i})v^i \leq \sum_{R \in \mathcal{R}_E(x)} \mu(R)\gamma_R^i(x).$$

(The first test in the scenario can no longer be performed; it has to be replaced by: check that the current state is in E.) The solvable sets which we define below will enjoy these additional properties by construction. Therefore, Theorem 2 still holds.

5. Existence of a solvable set.

5.1. Preliminary discussion. A few preliminary explanations are in order at this point. We prove below the existence of a solvable set within the set of states with *highest* min max value (ranked according to some lexicographic order).

The basic idea is the following. Intuitively, each player i might prevent the play from leaving the states with highest min max value: otherwise, players $-i$ could bring players i's average payoffs below this highest value (simply by playing in such a way that the play reaches a set with lower min max values, and by playing optimally from then on). The obvious idea is then to order the states lexicographically, according to their min max values, to consider the set F of states with maximal min max values, and to find a stationary equilibrium in the discounted constrained game, in which each player is restricted to the actions that prevent the play from leaving F (this is a bit loose, since the corresponding sets of actions have to be defined iteratively; precise statements will follow). One then would prove that the payoff induced by this constrained equilibrium is individually rational, and one would deduce that some ergodic set for this equilibrium is solvable.

There are difficulties in implementing this program. If one deals with the *undiscounted* min max value, it is not true that constrained discounted equilibria yield individually rational payoffs (even approximately), as is shown by the two-player, zero-sum example below.

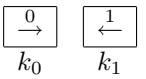
$$\begin{matrix} 0 & 1^* \\ 1^* & 0^* \end{matrix}$$

This game has one nonabsorbing state k, which is described by the previous matrix, and two absorbing states k_0 and k_1, with payoffs 0 and 1. As usual, the meaning of a starred entry is that, if the entry is played in state k, the play moves to the absorbing state with the corresponding payoff. Since an absorbing state is always solvable, the existence of solvable sets is here trivial. However, observe that the undiscounted value in the nonabsorbing state is 1 (the stationary strategy which puts a weight of ε to the

[1] I wish to thank a referee for pointing this out.

bottom row guarantees $1-\varepsilon$). The states with maximal undiscounted value to player 1 are therefore k and k_1. Set $F=\{k,k_1\}$. In the nonabsorbing state, the only action of player 1 that prevents the play from leaving F is the top row. Given that player 1 is playing the top row, both actions of player 2 prevent the play from leaving F. In the constrained game, where player 1 is restricted to the top row, the unique discounted equilibrium is clearly obtained when player 2 plays the left row, which results in a payoff of zero, which is not individually rational. Thus the above program fails (of course, one might argue that, for every player, there is always one initial state such that the corresponding payoff is individually rational; the fact that we deal with more than two players would lead to substantial difficulties).

Therefore, some care is needed in the comparison of states. We instead deal with the *discounted* min max value. But it is not true that a player can prevent the play from leaving the set of states with highest discounted min max value, as is shown by the next trivial one-player game.



This game has two states: the player has only one action in each state, and the play bounces back and forth between k_0 and k_1 (arrays indicate transitions, and numbers indicate current payoffs). The discounted payoff starting from k_1 is $\frac{1}{2-\lambda}$, which is higher than the payoff $\frac{1-\lambda}{2-\lambda}$ starting from k_0. Notice that the spread $\frac{\lambda}{2-\lambda}$ is some $O(\lambda)$. This is why the correct notion of comparison will allow for this $O(\lambda)$.

5.2. States with high discounted min max value. Let $f:(0,1]\to\mathbf{R}$ be a function with a Puiseux expansion $f(\lambda)=\sum_{p=0}^{\infty}a_p\lambda^{\frac{p}{M}}$ in the neighborhood of 0. We use the Landau notation and write $f=O(\lambda)$ if the function $\lambda\mapsto\frac{f(\lambda)}{\lambda}$ is bounded in a neighborhood of 0. This amounts to $\frac{1}{M}\inf\{p\in\mathbf{N},a_p\neq 0\}\geq 1$.

Let $F\subseteq K$. Since F is finite, it is clear that $\lambda\mapsto\max_F v^i_\lambda$ has a Puiseux expansion in the neighborhood of 0. We set

$$B_i(F)=\{k\in F,\ \max_F v^i_\lambda - v^i_\lambda(k)=O(\lambda)\}:$$

this is the (nonempty) subset of states of (almost) highest value for player i.

For $k\in B_i(F)$, there exist $c_0\geq 0$ and $\lambda_0>0$ such that

$$\forall\lambda<\lambda_0,\ \max_F v^i_\lambda - v^i_\lambda(k)\leq c_0\lambda. \tag{1}$$

For $k\notin B_i(F)$, there exists $c_1>0$, $\lambda_0>0$, and $\beta<1$, such that

$$\forall\lambda<\lambda_0,\ \max_F v^i_\lambda - v^i_\lambda(k)>c_1\lambda^\beta. \tag{2}$$

Since K and $\mathcal{I}$ are finite, we may choose c_0,c_1,λ_0, and β independently of F,k, and i.

Set then $F_0=K$ and, for $i=1,\dots,I$, $F_i=B_i(F_{i-1})$, and $F=F_I$. We shall prove the following improvement of Proposition 7.

PROPOSITION 9. *There exists a solvable set (E,x,μ), with $E\subseteq F$.*

Remark. The order $1,\dots,I$ on the players is obviously arbitrary: if one constructs the sets (F_i) using any other order on the players, the conclusion of Proposition 9 still holds.

For $k \in F$, we define inductively subsets $\tilde{A}_k^i$ of A^i for $i = 1, \dots, I$. The definitions of $\tilde{A}_k^1$ and of $\tilde{A}_k^i$ are similar. We give only the latter. For $k \in K$, set $B_k^j = \tilde{A}_k^j$ if $j < i$ and $B_k^j = A_k^j$ if $j > i$.

For $k \in F_i$, we set

$$\tilde{A}_k^i = \{a^i \in A^i \ \forall a^{-i} \in B_k^{-i},\ \mathbf{q}(F_i|k, a^i, a^{-i}) = 1\} :$$

this set contains the actions of player i that prevent the play from leaving F_i, if players $-i$ play actions in B_k^{-i}. For $k \notin F_i$, we set $\tilde{A}_k^i = A^i$.

Leaving aside for a moment the issue of nonemptiness of $\tilde{A}_k^i$, we briefly explain the motivation and the agenda.

By induction, given that players $-i$ use only actions with B_k^{-i}, the play, once in F_{i-1}, stays in F_{i-1}. F_i is to be thought of as those states in F_{i-1} with maximal min max value to player i. For $k \in F_i$, $\tilde{A}_k^i$ is the set of actions which force the play to remain within F_i. We shall consider discounted games in which each player i is restricted to actions in $\tilde{A}_k^i$. These games do have a stationary equilibrium x_λ (standard proof). We shall prove that any subset E of F which is ergodic for some x_λ is solvable, with payoff $\lim_{\lambda \to 0} \gamma_\lambda(k, x_\lambda)$, where $k \in E$. In order to establish this fact, we need to prove three kinds of results.

- A communication property: it will be trivial, given the construction of the set.
- No unilateral deviation of a player i can increase the current min max level: $\forall a^i, \mathbf{q}(\cdot|k, a^i, x_k^{-i})v^i \le v^i(k)$. This will be straightforward, given the definitions of the sets (F_i) and $(\tilde{A}^i)$.
- Somewhat surprisingly, the only nontrivial issue is that of individual rationality: is it true that $\lim_{\lambda \to 0} \gamma_\lambda^i(k, x_\lambda) \ge v^i(k)$? The positive answer will be deduced from Lemma 12.

5.3. Properties of $\tilde{A}_k^i$. For $k \in F_i$ and $\lambda < \lambda_0$, the following hold by definition:

1. For each $a^i \in A^i$ and $a^{-i} \in B_k^{-i}$, one has $\mathbf{q}(F_{i-1}|k, a^i, a^{-i}) = 1$; hence, using (1),

$$\mathbf{q}(\cdot|k, a^i, a^{-i})v_\lambda^i \le v_\lambda^i(k) + c_0\lambda;$$

therefore,

$$U_{\lambda, a^i, a^{-i}}^i v_\lambda^i(k) \le v_\lambda^i(k) + C_0\lambda \tag{3}$$

for some C_0 which depends only upon the data of the game.

2. For each $a^i \notin \tilde{A}_k^i$, there exists $a^{-i} \in B_k^{-i}$, such that $\mathbf{q}(F_i|k, a^i, a^{-i}) < 1$. Since $\mathbf{q}(F_i|k, a^i, a^{-i}) = 1$, this implies, using (2), that

$$\mathbf{q}(\cdot|k, a^i, a^{-i})v_\lambda^i < v_\lambda^i(k) - \eta c_1 \lambda^\beta + c_0\lambda,$$

where $\eta > 0$ depends only upon $\mathbf{q}$. Therefore, there exists $\lambda_1 < \lambda_0$ such that

$$\forall \lambda < \lambda_1, U_{\lambda, a^i, a^{-i}}^i v_\lambda^i(k) < v_\lambda^i(k) - c_2\lambda^\beta \tag{4}$$

for some $c_2 > 0$ which depends upon the data of the game.

LEMMA 10. $\tilde{A}^i_k \neq \emptyset \ \forall i,\ k \in F_i$.

Proof. We proceed by induction over i. Fix x^{-i} with $\mathrm{supp}(x^{-i}) = B^{-i}$. For each $\lambda > 0$, there exists $a^i = (a^i_k) \in (A^i)^K$ such that $U^i_{\lambda,(a^i,x^{-i})} v^i_\lambda \geq v^i_\lambda$ (Fact 2). Since $(A^i)^K$ is finite, there exist $a^i \in (A^i)^K$ and a sequence (λ_n) decreasing to 0, such that

$$U^i_{\lambda_n,a^i,x^{-i}} v^i_{\lambda_n} \geq v^i_{\lambda_n} \text{for every } n.$$

From (1) and (2), one deduces that $a^i_k \in \tilde{A}^i_k$ for every k. □

We shall need the following property.

LEMMA 11. *Let Ω be a finite set, $P_1, P_2 \in \Delta(\Omega)$, $u : \Omega \to \mathbf{R}$. Then*

$$|P_1 u - P_2 u| \leq \|P_1 - P_2\| \max_{k,k' \in \Omega} |u(k) - u(k')|.$$

Proof. Clearly, $|P_1 u - P_2 u|$ is unchanged when a constant is added to u. Replace u by $u - \min_\Omega u$, so that $\max_{k,k' \in \Omega} |u(k) - u(k')| = \max_\Omega u$. Then

$$|P_1 u - P_2 u| \leq \sum_{k \in \Omega} |P_1(k) - P_2(k)| u(k) \leq \|P_1 - P_2\| \max_\Omega u. \qquad \square$$

The next lemma essentially entails that provided players $-i$ use some profile x^{-i} with $\mathrm{supp}(x^{-i}) \subseteq B^{-i}$, there is a reply x^i of player i with $\mathrm{supp}(x^i) \subseteq \tilde{A}^i$, such that $\gamma^i_\lambda(x^{-i}, x^i) \geq v^i_\lambda - o(1)$.

Choose a real number $\gamma > 0$, such that $\beta + \gamma < 1$.

LEMMA 12. *Let $\lambda < \min\{\lambda_1, (\frac{c_2}{C_0})^{\frac{1}{1-\beta-\gamma}}\}$, and let $x^{-i} \in \mathcal{S}^{-i}$. If $\mathrm{supp}(x^{-i}) \subseteq B^{-i}$, there exists $a^i \in \tilde{A}^i$, such that*

$$U^i_{\lambda,a^i,x^{-i}} v^i_\lambda \geq v^i_\lambda - c_3 \lambda^{1+\gamma},$$

with $c_3 = |A^{-i}|(1 + 2C_0)$.

Proof. Choose $\tilde{x}^{-i}$, such that

$$\begin{cases} \mathrm{supp}(\tilde{x}^{-i}) = B^{-i}, \\ \|\tilde{x}^{-i} - x^{-i}\| \leq |A^{-i}| \lambda^\gamma, \\ \tilde{x}^j_k(a^j_k) \geq \lambda^\gamma \ \forall k, j, a^j_k \in \tilde{A}^j_k. \end{cases}$$

($\tilde{x}^{-i}$ is to be interpreted as a λ^γ-perturbation of x^{-i}.)

Choose $a^i = (a^i_k) \in (A^i)^K$, such that $U^i_{\lambda,a^i,\tilde{x}^{-i}} v^i_\lambda \geq v^i_\lambda$. We argue now that $a^i_k \in \tilde{A}^i_k$ for every k. Since $\tilde{A}^i_k = A^i$ for $k \notin F_i$, this may only fail for some $k \in F_i$. In that case, one would have

$$U^i_{\lambda,a^i,a^{-i}} v^i_\lambda(k) \leq v^i_\lambda(k) + C_0 \lambda$$

for each $a^{-i}_k \in B^{-i}_k$, and

$$U^i_{\lambda,a^i,a^{-i}} v^i_\lambda(k) < v^i_\lambda(k) - c_2 \lambda^\beta$$

for at least one $a^{-i}_k \in B^{-i}_k$. By summation, one would get

$$U^i_{\lambda,a^i,\tilde{x}^{-i}} v^i_\lambda(k) < v^i_\lambda(k) - c_2 \lambda^{\beta+\gamma} + C_0 \lambda,$$

which would contradict the choice of a^i. Hence $a^i_k \in \tilde{A}^i_k$ for every $k \in K$.

Apply now Lemma 11, with $P_1 = \mathbf{q}(\cdot|k, a_k^i, x_k^{-i})$, $P_2 = \mathbf{q}(\cdot|k, a_k^i, \tilde{x}_k^{-i})$, and $u = v_\lambda^i$. This yields

$$\begin{aligned}|U^i_{\lambda,a^i,\tilde{x}^i} v^i_\lambda(k) - U^i_{\lambda,a^i,x^{-i}} v^i_\lambda(k)| &\le \lambda \|\tilde{x}_k^i - x_k^i\| \max|g^i| \\ &\quad + \|\tilde{x}_k^i - x_k^i\| \max_{k_1,k_2\in F} |v_\lambda^i(k_1) - v_\lambda^i(k_2)| \\ &\le |A^{-i}|\lambda^{1+\gamma}(1 + 2C_0).\end{aligned}$$

The result follows. □

5.4. Constrained games and solvable sets. Let $\tilde{\mathcal{S}}^i = \{x^i \in \mathcal{S}^i, \mathrm{supp}(x^i) \subseteq \tilde{A}^i\}$, and set $\tilde{\mathcal{S}} = \prod_i \tilde{\mathcal{S}}^i$. Let $\lambda > 0$. We define a best-reply map $\mathbf{B}_\lambda : \tilde{\mathcal{S}} \to \tilde{\mathcal{S}}$. For $i \in I, x \in \tilde{\mathcal{S}}$, and $k \in K$, set

$$\mathbf{B}^i_{\lambda,k}(x^{-i}) = \mathrm{argmax}_{\Delta(\tilde{A}^i_k)} U^i_{\lambda,.,x^{-i}} v^i_\lambda(k),$$

and

$$\mathbf{B}^i_\lambda(x^{-i}) = \prod_k \mathbf{B}^i_{\lambda,k}(x^{-i}).$$

Finally, we set $\mathbf{B}_\lambda(x) = \prod_i \mathbf{B}^i_\lambda(x^{-i})$.

Clearly, $\mathbf{B}_\lambda$ is upperhemicontinuous and convex-valued. Therefore, by Kakutani's theorem, it has a fixed point x_λ. By compactness, there exists a sequence $(\lambda_n)_n$ decreasing to 0, such that $\overline{x} = \lim_{n\to\infty} x_{\lambda_n}$ and $\gamma = \lim_{n\to\infty} \gamma_{\lambda_n}(x_{\lambda_n})$ exist.

Let $x \in \tilde{\mathcal{S}}$ such that $\mathrm{supp}(x) = \tilde{A}$. By construction, F is stable under x. Choose a set $E \subseteq F$, ergodic for x.

LEMMA 13. *There exists a distribution μ over $\mathcal{R}_E(\overline{x})$ such that $(E, \overline{x}, \mu)$ is solvable.*

Proof. To avoid cumbersome notations, we write λ instead of λ_n. All limits below are taken along the sequence (λ_n).

First, notice that $\mathrm{supp}(\overline{x}) \subseteq \mathrm{supp}(x)$; since E is ergodic for x, $E \in \mathcal{C}(\overline{x})$. Also, since $\mathrm{supp}(x_\lambda) \subseteq \mathrm{supp}(x)$, E is stable under x_λ.

Using Lemma 12 and the fixed-point property of x_λ, one has

$$U^i_{\lambda,x_\lambda} v^i_\lambda \ge v^i_\lambda - c_3\lambda^{1+\gamma}$$

for each player i. Hence $\gamma^i_\lambda(x_\lambda) \ge v^i_\lambda - c_3\lambda^\gamma$, by Fact 1.

For $k_0 \in E$, it is well known (see Thuijsman and Vrieze [6]) that $\lim_{\lambda\to 0} \gamma_\lambda(k_0, x_\lambda)$ belongs to the convex hull of $\{\gamma_R(\overline{x}), R \in \mathcal{R}_E(\overline{x})\}$. Hence there exists a distribution μ over $\mathcal{R}_E(\overline{x})$, such that

$$\lim_{\lambda\to 0} \gamma_\lambda(k_0, x_\lambda) = \sum_{R\in\mathcal{R}_E(\overline{x})} \mu(R)\gamma_R(\overline{x}) \ge v(k_0).$$

To complete the proof, we prove that

$$\mathbf{q}(\cdot|k, a^i, \overline{x}_k^{-i}) v^i \le \max_E v^i \tag{5}$$

for each i, $k \in E$, and $a^i \in A^i$.

By construction, $v^i(k) = \max_{F_{i-1}} v^i$ for every player i and $k \in E$. In particular, v is constant over E.

Since $\mathrm{supp}(\overline{x}^j_k) \subseteq B^j_k$ for each $j \ne i$, one has $\mathbf{q}(F_{i-1}|k, a^i, \overline{x}_k^{-i}) = 1$. This implies (5). □

REFERENCES

[1] T. Bewley and E. Kolhberg, *The asymptotic theory of stochastic games*, Math. Oper. Res., 1 (1976), pp. 197–208.
[2] J. F. Mertens and A. Neyman, *Stochastic games*, Internat. J. Game Theory, 10 (1981), pp. 53–66.
[3] J. F. Mertens, S. Sorin, and S. Zamir, *Repeated Games, Parts A, B, and C*, CORE Discussion Paper, Université Catholique de Louvain, Belgium, 1994.
[4] A. Neyman, *Stochastic Games*, preprint, 1988.
[5] L. S. Shapley, *Stochastic games*, Proc. Nat. Acad. Sci. U.S.A., 39 (1953), pp. 1095–1100.
[6] F. Thuijsman and O. J. Vrieze, *Easy initial states in stochastic games*, in Stochastic Games and Related Topics, in Honor of L. S. Shapley, Kluwer Academic Publishers, Dordrecht, The Netherlands, 1991, pp. 85–100.
[7] S. Tijs and K. Vrieze, *On the existence of easy initial states for undiscounted stochastic games*, Math. Oper. Res., 11 (1986), pp. 506–513.
[8] N. Vieille, *Solvable states in stochastic games*, Internat. J. Game Theory, 21 (1993), pp. 395–404.
[9] N. Vieille, 2-*player stochastic games* I: *A reduction*, Cahiers du CEREMADE, 9745, Université Paris 9-Dauphine, Paris, France, 1997.

SIAM J. CONTROL OPTIM.
Vol. 38, No. 6, pp. 1805–1829

MAXIMALLY ROBUST CONTROLLERS FOR MULTIVARIABLE SYSTEMS*

S. K. GUNGAH†, G. D. HALIKIAS‡, AND I. M. JAIMOUKHA§

Abstract. The set of all optimal controllers which maximize a robust stability radius for unstructured additive perturbations may be obtained using standard Hankel-norm approximation methods. These controllers guarantee robust stability for *all* perturbations which lie inside an open ball in the uncertainty space (say, of radius r_1). Necessary and sufficient conditions are obtained for a perturbation lying on the boundary of this ball to be destabilizing for *all* maximally robust controllers. It is thus shown that a "worst-case direction" exists along which all boundary perturbations are destabilizing. By imposing a parametric constraint such that the permissible perturbations cannot have a "projection" of magnitude larger than $(1-\delta)r_1$, $0 < \delta \leq 1$, in the most critical direction, the uncertainty region guaranteed to be stabilized by a *subset* of all maximally robust controllers can be *extended* beyond the ball of radius r_1. The choice of the *"best"* maximally robust controller—in the sense that the uncertainty region guaranteed to be stabilized becomes as large as possible—is associated with the solution of a superoptimal approximation problem. Expressions for the improved stability radius are obtained and some interesting links with μ-analysis are pursued.

Key words. robust control, μ-analysis, stability radius, superoptimization, Nehari problem

AMS subject classifications. 93B28, 93B40, 93B36

PII. S0363012999350559

1. Notation. $\mathcal{R}, \mathcal{R}_+$, and $\mathcal{C}$ denote the sets of real, nonnegative, and complex numbers, respectively. $\mathcal{C}_+$ ($\bar{\mathcal{C}}_+$), $\mathcal{C}_-$ ($\bar{\mathcal{C}}_-$) denote the open (closed) right half-plane and the open (closed) left half-plane, respectively. For a complex matrix A, A^T denotes the transpose while A' denotes the complex-conjugate transpose. $\sigma_i(A)$ denotes the ith largest singular value. The smallest singular value is denoted by $\underline{\sigma}(A)$ and the largest singular value is denoted by $\bar{\sigma}(A)$. The norm of A is defined as $\|A\| = \bar{\sigma}(A)$. For a square A, $\lambda(A)$ is the spectrum of A and $\lambda_{\max}(A)$ is the largest eigenvalue.

$\mathcal{L}_\infty^{p\times m}$ denotes the space of all $p \times m$ matrix functions with entries uniformly bounded on the $j\omega$-axis. $\mathcal{H}_\infty^{p\times m}$ and $\mathcal{H}_\infty^{-p\times m}$ denote the subspaces of $\mathcal{L}_\infty^{p\times m}$ consisting of all matrix functions whose elements are analytic in $\bar{\mathcal{C}}_+$ and $\bar{\mathcal{C}}_-$, respectively. $\|.\|_\infty$ denotes the $\mathcal{L}_\infty$ norm of matrices in $\mathcal{L}_\infty$ or the $\mathcal{H}_\infty$ norm of matrices in $\mathcal{H}_\infty$, depending on context. $\gamma\mathcal{B}\mathcal{H}_\infty^{p\times m} = \{G \in \mathcal{H}_\infty^{p\times m} : \|G\|_\infty \leq \gamma\}$ is the γ ball of $\mathcal{H}_\infty^{p\times m}$. The prefix $\mathcal{R}$ before a set symbol means that the elements of the set are restricted to be real-rational. Matrix dimensions of spaces will be occasionally suppressed.

$G(s)^\sim := G'(-\bar{s})$ denotes the *para-hermitian conjugate* of $G(s)$. $G(s)^{-\sim}$ stands for $(G(s)^\sim)^{-1}$. The *Hankel operator* with symbol $G \in \mathcal{H}_\infty$ is denoted by Γ_G while $\sigma_i(\Gamma_G)$ denotes the ith largest Hankel singular value of G. The Hankel norm of G, $\sigma_1(\Gamma_G)$, is also written as $\|\Gamma_G\|$ and the smallest Hankel singular value as $\underline{\sigma}(\Gamma_G)$.

A real-rational function $G(s)$ is called *stable* if it has no poles in $\bar{\mathcal{C}}_+$. If $G(s)$ has no poles in $\bar{\mathcal{C}}_-$, it is called *antistable*. Matrix (scalar and vector) transfer functions will

*Received by the editors January 4, 1999; accepted for publication (in revised form) October 18, 1999; published electronically June 20, 2000.

http://www.siam.org/journals/sicon/38-6/35055.html

†Interdisciplinary Research Center for Process Systems Engineering, Imperial College of Science, Technology and Medicine, London SW7-2BY, UK (s.gungah@ps.ic.ac.uk).

‡Department of Electronic and Electrical Engineering, University of Leeds, Leeds LS2 9JT, UK (een6gdh@sun.leeds.ac.uk).

§Department of Electrical and Electronic Engineering, Imperial College of Science, Technology and Medicine, London SW7-2BT, UK (i.jaimouka@ic.ac.uk).

be represented by uppercase (lowercase) boldface letters and with the dependence on s mostly suppressed. If $\boldsymbol{G}^{-1} = \gamma^{-2}\boldsymbol{G}^{\sim}$, then $\boldsymbol{G}$ is called γ-*allpass* (or simply *allpass* if $\gamma = 1$) and satisfies $\boldsymbol{G}\boldsymbol{G}^{\sim} = \boldsymbol{G}^{\sim}\boldsymbol{G} = \gamma^2 I$. A matrix function $\boldsymbol{G} \in \mathcal{RH}_\infty$ which satisfies $\boldsymbol{G}^{\sim}\boldsymbol{G} = I$ is called *inner*. A matrix function $\boldsymbol{G}(s) \in \mathcal{RH}_\infty$ which has full column rank for all $s \in \bar{\mathcal{C}}_+$ is called *outer*. If $\boldsymbol{U} \in \mathcal{L}_\infty^{l\times q}$ and

$$\boldsymbol{H} = \begin{bmatrix} \boldsymbol{H}_{11} & \boldsymbol{H}_{12} \\ \boldsymbol{H}_{21} & \boldsymbol{H}_{22} \end{bmatrix} \in \mathcal{L}_\infty^{(p+q)\times(m+l)}$$

with $\boldsymbol{H}_{11} \in \mathcal{L}_\infty^{p\times m}$, we define the lower linear fractional map $\mathcal{F}_l(\boldsymbol{H},\boldsymbol{U}) = \boldsymbol{H}_{11} + \boldsymbol{H}_{12}\boldsymbol{U}(I - \boldsymbol{H}_{22}\boldsymbol{U})^{-1}\boldsymbol{H}_{21}$, provided that $I - \boldsymbol{H}_{22}(\infty)\boldsymbol{U}(\infty)$ is invertible. If $\boldsymbol{U} \in \mathcal{L}_\infty^{m\times p}$, we define the upper linear fractional map $\mathcal{F}_u(\boldsymbol{H},\boldsymbol{U}) = \boldsymbol{H}_{22} + \boldsymbol{H}_{21}\boldsymbol{U}(I - \boldsymbol{H}_{11}\boldsymbol{U})^{-1}\boldsymbol{H}_{12}$, provided that $I - \boldsymbol{H}_{11}(\infty)\boldsymbol{U}(\infty)$ is invertible. If $\mathcal{U}$ is a set, then $\mathcal{F}_l(\boldsymbol{H},\mathcal{U})$ denotes the set $\{\mathcal{F}_l(\boldsymbol{H},\boldsymbol{U}) : \boldsymbol{U} \in \mathcal{U}\}$ and if $\boldsymbol{G}_1, \boldsymbol{G}_2, \boldsymbol{G}_3 \in \mathcal{L}_\infty$ have appropriate dimensions, then $\boldsymbol{G}_1 + \boldsymbol{G}_2\mathcal{U}\boldsymbol{G}_3$ denotes the set $\{\boldsymbol{G}_1 + \boldsymbol{G}_2\boldsymbol{U}\boldsymbol{G}_3 \ : \ \boldsymbol{U} \in \mathcal{U}\}$.

If $\boldsymbol{G} \in \mathcal{L}_\infty$, we define, for each i, $s_i^\infty(\boldsymbol{G}) = \sup_{\omega\in\mathcal{R}} \sigma_i(\boldsymbol{G}(j\omega))$. Clearly, $s_1^\infty(\boldsymbol{G}) = \|\boldsymbol{G}\|_\infty$. Suppose that $\mathcal{T}$ is a set of matrix functions. $\boldsymbol{T} \in \mathcal{T}$ is called a kth level *superoptimal* function if it minimizes the sequence $\{s_1^\infty(\boldsymbol{T}), s_2^\infty(\boldsymbol{T}), \ldots, s_k^\infty(\boldsymbol{T})\}$ with respect to lexicographic ordering among all $\boldsymbol{T} \in \mathcal{T}$. The minimized sequence is denoted by $\{s_1(\mathcal{T}), \ldots, s_k(\mathcal{T})\}$, and the $s_i(\mathcal{T})$s are called the superoptimal levels of $\mathcal{T}$.

2. Introduction. The work presented in this paper is related to the problem of maximizing the robust stability radius for systems subject to unstructured additive perturbations [25], [6], [23], [24]. In [6] it was shown that this problem is equivalent to a Nehari approximation. Moreover, an explicit state-space parametrization was obtained for all controllers which guarantee a robust stabilization radius $r < r_1$. A parametrization of all maximally robust controllers ($r = r_1$) is implicit in [6] and may be obtained from [4], [5]. The theory of optimal interpolation is used in [23] to give a solution for single input/single output systems.

In the multi-input/single output or single input/multi-output case, the optimal controller is unique. In the matrix case, however, a continuum of optimal controllers typically exists. It is therefore natural to ask whether a subset of these controllers offers improved robust stability properties, in the sense that it guarantees closed-loop stability for a larger class of uncertainties, compared to those offered by the optimal solution set considered in total. More specifically, we seek to identify the set of all controllers which guarantees robust stability for the largest possible region of the uncertainty space containing the open ball of radius r_1 as a subset. Clearly, this can only be achieved by imposing a structure on the set of admissible uncertainties.

Our approach is as follows: From the work in [25], [24], and [6] the maximum robust stability radius r_1 is the inverse of the smallest achievable $\mathcal{H}_\infty$ norm among all interpolating functions $\mathcal{T} = \{\boldsymbol{K}(I - \boldsymbol{G}\boldsymbol{K})^{-1}\}$, as $\boldsymbol{K}$ varies over the set of all internally stabilizing compensators of $\boldsymbol{G}$. Using an allpass dilation technique, the set of all optimal interpolating functions $\mathcal{T}_1 = \{\boldsymbol{T} \in \mathcal{T} : \|\boldsymbol{T}\|_\infty = r_1^{-1}\} \subseteq \mathcal{T}$ has the form $\mathcal{T}_1 = \boldsymbol{Y}\mathrm{diag}(r_1^{-1}\boldsymbol{a}, \hat{\boldsymbol{R}} + \mathcal{Q})\boldsymbol{X}$, where $\hat{\boldsymbol{R}} \in \mathcal{RH}_\infty^-$, $\boldsymbol{X}$ and $\boldsymbol{Y}$ are square inner matrices, $\boldsymbol{a}$ is a scalar allpass function, and $\mathcal{Q}$ is the set of all r_1^{-1} suboptimal Nehari extensions of $\hat{\boldsymbol{R}}$, i.e., $\mathcal{Q} = \{\boldsymbol{Q} \in \mathcal{H}_\infty : \|\hat{\boldsymbol{R}} + \boldsymbol{Q}\|_\infty \le r_1^{-1}\}$.

Every optimal controller corresponding to an interpolating function in $\mathcal{T}_1$ stabilizes *all* perturbations which lie inside the open ball $\mathcal{D}_{r_1} = \{\boldsymbol{\Delta} \in \mathcal{L}_\infty : \|\boldsymbol{\Delta}\|_\infty < r_1, \eta(\boldsymbol{G} + \boldsymbol{\Delta}) = \eta(\boldsymbol{G})\}$, where $\eta(\cdot)$ denotes number of poles in $\mathcal{C}_+$. Next, we consider perturbations $\boldsymbol{\Delta}$ which lie on the boundary of $\mathcal{D}_{r_1}$. It is shown that such boundary

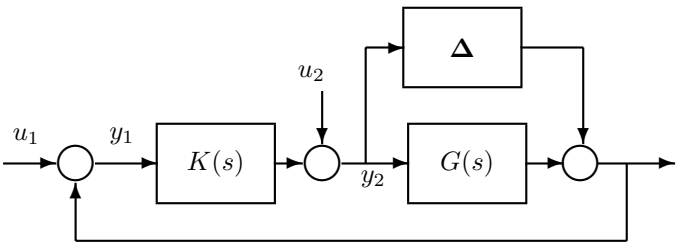


FIG. 1. *Closed-loop system.*

perturbations are *uniformly* destabilizing (i.e., they destabilize the closed-loop system for every optimal controller) if and only if $|\boldsymbol{x}^T(j\omega)\boldsymbol{\Delta}(j\omega)\boldsymbol{y}(j\omega)| = r_1$ for some frequency ω, and where $\boldsymbol{x}^T$ and $\boldsymbol{y}$ are the first row and column of $\boldsymbol{X}$ and $\boldsymbol{Y}$, respectively. Moreover, all frequencies ω are equally critical, in the sense that destabilizing boundary perturbations can be constructed for every $\omega \in \mathcal{R}$. This shows that it is futile to attempt to extend the uncertainty set guaranteed to be stabilized by a subset of all optimal controllers in the (frequency-dependent) direction defined by vectors $\boldsymbol{x}^T$ and $\boldsymbol{y}$. By imposing a parametric constraint (uniform in ω) such that the permissible perturbations cannot have a "projection" of magnitude larger than $(1-\delta)r_1$ $(0 < \delta \leq 1)$ in this direction, the uncertainty region guaranteed to be stabilized by a subset of all optimal controllers can be extended beyond $\mathcal{D}_{r_1}$. Using a result in [15], it is shown that for each $\delta \in (0, 1]$ the corresponding constrained robust stability radius is maximized by the set of controllers which minimize the first two superoptimal levels of $\mathcal{T}$. A closed-form expression of the improved stability radius is also obtained which involves δ and the first two superoptimal levels of $\mathcal{T}$. This work is related to the results presented in [18] which also uses superoptimization concepts to give an extension of the allowable perturbation set.

An alternative interpretation of our results leads to interesting connections with the problem of robust stabilization of systems subject to structured perturbations and μ-synthesis in general [19]. By suitably defining δ, robust stabilization problems for a number of uncertainty structures can be formulated in our setting, and bounds on the achievable robust-stability radius can be obtained. An upper bound on μ for the constant complex case is derived in the last section.

The layout of the paper is as follows: Section 3 outlines a number of known results in the area of robust stabilization of systems subject to unstructured additive perturbations. The maximum robust stability radius is obtained by solving a Nehari approximation problem [6] and leads to a parametrization of all optimal interpolating functions, using the results of [4], [5]. An alternative parametrization of this set is obtained in section 4, using an allpass dilation technique [7], [5]. A recursive application of this method leads to the solution of the superoptimal approximation problem [26], [22], [16], [13], [14], [12], [20], [21]. In our context, this parametrization of the set of all optimal interpolating functions is used to characterize all uniformly destabilizing boundary perturbations. This analysis is carried out in section 4, which also includes our main result (Theorem 4.8) whose proof is based on a result from [15] (see Theorems 4.6 and 4.7). In section 5 a new (upper) bound on the structured singular value μ is obtained (for the constant problem). Finally, section 6 contains the conclusions.

3. Robust stabilization for unstructured additive perturbations. Let $\boldsymbol{G} \in \mathcal{RL}_\infty$. When $\boldsymbol{\Delta} = 0$, the closed-loop system in Figure 1 is internally stable

if and only if it is well-posed, i.e., $\det(I - \boldsymbol{G}(\infty)\boldsymbol{K}(\infty)) \neq 0$ and the four transfer functions $(u_1, u_2) \to (y_1, y_2)$ are in $\mathcal{H}_\infty$. In this case, we write $(\boldsymbol{G}, \boldsymbol{K}) \in \mathcal{S}$ and $\boldsymbol{K} \in \mathcal{K}$. Consider the set of additively perturbed systems $\boldsymbol{G} + \boldsymbol{\Delta}, \boldsymbol{\Delta} \in \mathcal{D}_r(\boldsymbol{G}, \boldsymbol{w})$, where $\boldsymbol{w}$ is a scalar outer $\mathcal{RH}_\infty$ (weighting) function and

$$\mathcal{D}_r(\boldsymbol{G}, \boldsymbol{w}) = \left\{ \boldsymbol{\Delta} \in \mathcal{L}_\infty : \|\boldsymbol{w}^{-1}\boldsymbol{\Delta}\|_\infty < r, \eta(\boldsymbol{G}) = \eta(\boldsymbol{G} + \boldsymbol{\Delta}) \right\},$$

in which $\eta(\cdot)$ denotes the number of poles in $\mathcal{C}_+$, counted in a MacMillan degree sense. The system $(\boldsymbol{G}, \boldsymbol{K})$ is said to be $(r, \boldsymbol{w})$ robustly stable if $(\boldsymbol{G} + \boldsymbol{\Delta}, \boldsymbol{K}) \in \mathcal{S}$ for all $\boldsymbol{\Delta} \in \mathcal{D}_r(\boldsymbol{G}, \boldsymbol{w})$. $(\boldsymbol{G}, \boldsymbol{K})$ is said to be maximally robustly stable if (i) $(\boldsymbol{G}, \boldsymbol{K})$ is $(r_1, \boldsymbol{w})$ robustly stable and (ii) there exists $\boldsymbol{\Delta} \in \partial\mathcal{D}_{r_1}(\boldsymbol{G}, \boldsymbol{w}) = \{\boldsymbol{\Delta} \in \mathcal{L}_\infty : \|\boldsymbol{w}^{-1}\boldsymbol{\Delta}\|_\infty = r_1, \eta(\boldsymbol{G}) = \eta(\boldsymbol{G} + \boldsymbol{\Delta})\}$, such that $(\boldsymbol{G} + \boldsymbol{\Delta}, \boldsymbol{K}) \notin \mathcal{S}$. The following theorem [25] gives necessary and sufficient conditions for robust stabilization in the presence of additive unstructured perturbations.

THEOREM 3.1 (see [25], [6], [24]). *Let $\boldsymbol{G} \in \mathcal{RL}_\infty$ and suppose that $(\boldsymbol{G}, \boldsymbol{K}) \in \mathcal{S}$. Then $(\boldsymbol{G}, \boldsymbol{K})$ is $(r, \boldsymbol{w})$ robustly stable if and only if $\|\boldsymbol{w}\boldsymbol{K}(I - \boldsymbol{G}\boldsymbol{K})^{-1}\|_\infty \leq r^{-1}$.*

The following result [6], [1] shows that, without loss of generality, $\boldsymbol{G}$ can be assumed to be antistable and $\boldsymbol{w}$ can be taken to be equal to one.

THEOREM 3.2 (see [6], [1]). *Let $\boldsymbol{w}^{-1}\boldsymbol{G}$ have a decomposition $\boldsymbol{w}^{-1}\boldsymbol{G} = \boldsymbol{G}_1 + \boldsymbol{G}_2$ with $\boldsymbol{G}_1^\sim, \boldsymbol{G}_2 \in \mathcal{RH}_\infty$, and $\boldsymbol{G}_1(\infty) = 0$. Then there exists a $\boldsymbol{K} \in \mathcal{L}_\infty$ such that $(\boldsymbol{G}, \boldsymbol{K})$ is $(r, \boldsymbol{w})$ robustly stable if and only if $\boldsymbol{K} = \boldsymbol{w}^{-1}\boldsymbol{K}_1(I + \boldsymbol{G}_2\boldsymbol{K}_1)^{-1}$ for some $\boldsymbol{K}_1$ such that $(\boldsymbol{G}_1, \boldsymbol{K}_1)$ is $(r, 1)$ robustly stable and $\det\{(I + \boldsymbol{G}_2\boldsymbol{K}_1)(\infty)\} \neq 0$.*

Remark 3.1. We assume for simplicity that $\boldsymbol{G}^\sim \in \mathcal{RH}_\infty$, $\boldsymbol{G}(\infty) = 0$, and $\boldsymbol{w} = 1$. We also use the simplified notation

$$\mathcal{D}_r(\boldsymbol{G}) := \mathcal{D}_r(\boldsymbol{G}, 1) = \{\boldsymbol{\Delta} \in \mathcal{L}_\infty : \|\boldsymbol{\Delta}\|_\infty < r, \eta(\boldsymbol{G}) = \eta(\boldsymbol{G} + \boldsymbol{\Delta})\}, \tag{1}$$

$$\partial\mathcal{D}_r(\boldsymbol{G}) := \partial\mathcal{D}_r(\boldsymbol{G}, 1) = \{\boldsymbol{\Delta} \in \mathcal{L}_\infty : \|\boldsymbol{\Delta}\|_\infty = r, \eta(\boldsymbol{G}) = \eta(\boldsymbol{G} + \boldsymbol{\Delta})\}. \tag{2}$$

Let $\boldsymbol{G}$ have left and right coprime factorizations $\boldsymbol{G} = \boldsymbol{N}\boldsymbol{M}^{-1} = \tilde{\boldsymbol{M}}^{-1}\tilde{\boldsymbol{N}}$, respectively, with $\boldsymbol{N}, \boldsymbol{M}, \tilde{\boldsymbol{N}}, \tilde{\boldsymbol{M}} \in \mathcal{RH}_\infty$ and let $\boldsymbol{U}, \boldsymbol{V}, \tilde{\boldsymbol{U}}, \tilde{\boldsymbol{V}} \in \mathcal{RH}_\infty$ satisfy the Bezout identities $\tilde{\boldsymbol{V}}\boldsymbol{M} - \tilde{\boldsymbol{U}}\boldsymbol{N} = I$ and $\tilde{\boldsymbol{M}}\boldsymbol{V} - \tilde{\boldsymbol{N}}\boldsymbol{U} = I$. Then the set of all stabilizing controllers of $\boldsymbol{G}$ is

$$\mathcal{K} = \{(\boldsymbol{U} + \boldsymbol{M}\boldsymbol{Q})(\boldsymbol{V} + \boldsymbol{N}\boldsymbol{Q})^{-1} : \boldsymbol{Q} \in \mathcal{H}_\infty\}. \tag{3}$$

Let $\mathcal{T} = \{\boldsymbol{K}(I - \boldsymbol{G}\boldsymbol{K})^{-1} : \boldsymbol{K} \in \mathcal{K}\}$. We refer to $\mathcal{T}$ as the set of all *interpolating functions*. Using the parametrization of $\mathcal{K}$ in (3) gives the alternative characterization of $\mathcal{T}$ as

$$\mathcal{T} = \{(\boldsymbol{U} + \boldsymbol{M}\boldsymbol{Q})\tilde{\boldsymbol{M}} \ : \ \boldsymbol{Q} \in \mathcal{H}_\infty\}. \tag{4}$$

Let $\boldsymbol{G}$ have a minimal balanced realization $\boldsymbol{G}(s) = C(sI - A)^{-1}B + D$, such that $Re\ \lambda_i(A) > 0$ for all i and with $A\Sigma + \Sigma A' = BB'$, $A'\Sigma + \Sigma A = C'C$, $\Sigma = \operatorname{diag}(\sigma_1, \sigma_2, \ldots, \sigma_n)$, $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_n > 0$. Let $F = B'\Sigma^{-1}$ and $H = \Sigma^{-1}C'$. The coprime factors $\boldsymbol{N}, \boldsymbol{M}, \tilde{\boldsymbol{N}}, \tilde{\boldsymbol{M}}$ can now be defined (together with $\boldsymbol{U}, \boldsymbol{V}, \tilde{\boldsymbol{U}}, \tilde{\boldsymbol{V}}$) as

$$\begin{bmatrix} \boldsymbol{M}(s) & \boldsymbol{U}(s) \\ \boldsymbol{N}(s) & \boldsymbol{V}(s) \end{bmatrix} = \begin{bmatrix} -F \\ C \end{bmatrix} (sI - A + BF)^{-1} \begin{bmatrix} B & H \end{bmatrix} + \begin{bmatrix} I & 0 \\ 0 & I \end{bmatrix},$$

$$\begin{bmatrix} \tilde{\boldsymbol{V}}(s) & -\tilde{\boldsymbol{U}}(s) \\ -\tilde{\boldsymbol{N}}(s) & \tilde{\boldsymbol{M}}(s) \end{bmatrix} = \begin{bmatrix} F \\ -C \end{bmatrix} (sI - A + HC)^{-1} \begin{bmatrix} B & H \end{bmatrix} + \begin{bmatrix} I & 0 \\ 0 & I \end{bmatrix},$$

with $\boldsymbol{M}$ and $\tilde{\boldsymbol{M}}$ inner [3]. The next result shows that the maximal robust stability radius is equal to the smallest Hankel singular value of $\boldsymbol{G}(-s)$.

THEOREM 3.3 (see [6], [24]). *Let $\boldsymbol{G}^{\sim} \in \mathcal{RH}_\infty, \boldsymbol{G}(\infty) = 0$. Then the maximum stability radius r_1 for which $(\boldsymbol{G}, \boldsymbol{K})$ is $(r_1, 1)$ robustly stable for some $\boldsymbol{K} \in \mathcal{K}$ is given by $r_1 = \underline{\sigma}(\Gamma_{\boldsymbol{G}(-s)})$.*

Proof. From Theorem 3.1 $(\boldsymbol{G}, \boldsymbol{K})$ is $(r, 1)$ robustly stable if (i) $\boldsymbol{K}$ stabilizes $\boldsymbol{G}$ internally, and (ii) $\|\boldsymbol{K}(I - \boldsymbol{G}\boldsymbol{K})^{-1}\|_\infty \leq r^{-1}$. Hence,

$$r_1^{-1} = \inf \ \{\|\boldsymbol{K}(I - \boldsymbol{G}\boldsymbol{K})^{-1}\|_\infty : \boldsymbol{K} \in \mathcal{K}\},$$

where $\mathcal{K}$ is the set of all internally stabilizing controllers of $\boldsymbol{G}$. From (4),

$$r_1^{-1} = \inf \ \{\|(\boldsymbol{U} + \boldsymbol{M}\boldsymbol{Q})\tilde{\boldsymbol{M}}\|_\infty : \boldsymbol{Q} \in \mathcal{H}_\infty\} = \inf \ \{\|\boldsymbol{M}^{\sim}\boldsymbol{U} + \boldsymbol{Q}\|_\infty : \mathbf{Q} \in \mathcal{H}_\infty\} \tag{5}$$

since $\boldsymbol{M}, \tilde{\boldsymbol{M}}$ are inner. A straightforward state-space calculation shows that

$$\boldsymbol{M}^{\sim}\boldsymbol{U}(s) = -B'\Sigma^{-1}(sI - A)^{-1}\Sigma^{-1}C' \in \mathcal{RH}_\infty^- \tag{6}$$

in previously defined notation. From Nehari's theorem the infimum in (5) is attained and is given by the Hankel norm of $\boldsymbol{M}^{\sim}\boldsymbol{U}(-s)$. It is also straightforward to verify that the realization in (6) is balanced with grammians equal to $-\Sigma^{-1}$. Thus, the realization in (6) is minimal, and

$$r_1^{-2} = \|\Gamma_{\boldsymbol{M}^{\sim}\boldsymbol{U}(-s)}\|^2 = \lambda_{\max}(\Sigma^{-2}) = \underline{\sigma}^{-2}(\Gamma_{\boldsymbol{G}(-s)})$$

from which it follows that $r_1 = \underline{\sigma}(\Gamma_{\boldsymbol{G}(-s)})$, as required. ∎

Remark 3.2. Let $\boldsymbol{G}$ satisfy the assumptions of Theorem 3.3 and assume that the MacMillan degree of $\boldsymbol{G}$ is n. Let the Hankel singular values of $\boldsymbol{G}(-s)$ and $\boldsymbol{M}^{\sim}\boldsymbol{U}(-s)$ be $\{\sigma_i(\Gamma_{\boldsymbol{G}(-s)})\}$ and $\{\sigma_i(\Gamma_{\boldsymbol{M}^{\sim}\boldsymbol{U}(-s)})\}$, respectively, ordered in nonincreasing order of magnitude. Then $\sigma_n(\Gamma_{\boldsymbol{G}(-s)}) > 0$. Further, a slight adaptation of Theorem 3.3 shows that $\sigma_i(\Gamma_{\boldsymbol{G}(-s)}) = \sigma_{n-i+1}^{-1}(\Gamma_{\boldsymbol{M}^{\sim}\boldsymbol{U}(-s)})$ for each $i = 1, 2, \ldots, n$.

4. Main results. The set of all maximally robust controllers may be characterized in terms of the set of all optimal Nehari extensions of $\boldsymbol{M}^{\sim}\boldsymbol{U}$, i.e., the set of all $\boldsymbol{Q} \in \mathcal{H}_\infty$ which achieve

$$\|\boldsymbol{M}^{\sim}\boldsymbol{U} + \boldsymbol{Q}\|_\infty = r_1^{-1}. \tag{7}$$

This set can be parametrized as a linear fractional map of the set of all r_1 stable contractions [4], [5]. This parametrization is outlined next.

Remark 4.1. To avoid a messy indexing system we assume hereafter that the largest Hankel singular values of $\boldsymbol{R}(-s)$ and $\hat{\boldsymbol{R}}(-s)$ defined below in Theorems 4.1 and 4.2, respectively, are nonrepeated. These conditions are equivalent to the assumption that the first two superoptimal levels $s_1(\mathcal{T})$ and $s_2(\mathcal{T})$ are nonrepeated.

THEOREM 4.1 (see [14], [5], [7]). *Let $\boldsymbol{R} = \boldsymbol{M}^{\sim}\boldsymbol{U} \in \mathcal{RH}_\infty^{-p\times m}$ and define $r_1 = \bar{\sigma}^{-1}(\Gamma_{\boldsymbol{R}(-s)})$ (see Theorem 3.3). Then there exists an embedding of $\boldsymbol{R}$ of the form*

$$\boldsymbol{H} = \begin{bmatrix} \boldsymbol{H}_{11} & \boldsymbol{H}_{12} \\ \boldsymbol{H}_{21} & \boldsymbol{H}_{22} \end{bmatrix} = \begin{bmatrix} \boldsymbol{R} + \boldsymbol{Q}_{11} & \boldsymbol{Q}_{12} \\ \boldsymbol{Q}_{21} & \boldsymbol{Q}_{22} \end{bmatrix}$$

$$:= \begin{bmatrix} \boldsymbol{R} & 0 \\ 0 & 0 \end{bmatrix} + \boldsymbol{Q}_a \in \mathcal{RL}_\infty^{(p+m-1)\times(m+p-1)}$$

with $\boldsymbol{Q}_a \in \mathcal{RH}_\infty$*, such that* $\boldsymbol{H}\boldsymbol{H}^\sim = \boldsymbol{H}^\sim\boldsymbol{H} = r_1^{-2} I_{p+m-1}$ *and* $\|\boldsymbol{H}_{22}\|_\infty = \|\boldsymbol{Q}_{22}\|_\infty < r_1^{-1}$*. Further, the set of all (optimal)* $\boldsymbol{Q} \in \mathcal{H}_\infty^{p\times m}$ *such that* $\|\boldsymbol{R}+\boldsymbol{Q}\|_\infty = r_1^{-1}$ *is given by*

$$\mathcal{S}_1 = \mathcal{F}_l(\boldsymbol{Q}_a, r_1\mathcal{BH}_\infty^{(p-1)\times(m-1)}). \tag{8}$$

Let $\mathcal{K}_1$ denote the set of all maximally robust (r_1-robust) controllers of $\boldsymbol{G}$, and let $\mathcal{T}_1 = \{\boldsymbol{K}(I-\boldsymbol{G}\boldsymbol{K})^{-1} : \boldsymbol{K} \in \mathcal{K}_1\} \subseteq \mathcal{T}$ denote the set of all optimal interpolating functions. In view of (3) and (4), together with Theorems 3.3 and 4.1, these sets may be parametrized as $\mathcal{K}_1 = \{(\boldsymbol{U}+\boldsymbol{M}\boldsymbol{Q})(\boldsymbol{V}+\boldsymbol{N}\boldsymbol{Q})^{-1} : \boldsymbol{Q} \in \mathcal{F}_l(\boldsymbol{Q}_a, r_1\mathcal{BH}_\infty^{(p-1)\times(m-1)})\}$ and

$$\mathcal{T}_1 = \{(\boldsymbol{U}+\boldsymbol{M}\boldsymbol{Q})\tilde{\boldsymbol{M}} : \boldsymbol{Q} \in \mathcal{F}_l(\boldsymbol{Q}_a, r_1\mathcal{BH}_\infty^{(p-1)\times(m-1)})\}, \tag{9}$$

respectively. The next theorem gives an alternative parametrization of the set of all optimal solutions of (7), and therefore of $\mathcal{T}_1$ in (9). The result shows that $\mathcal{T}_1$ can be diagonalized by rational allpass transformations and is used extensively in this work.

THEOREM 4.2 (see [10]). *Let all variables be as defined in Theorem* 4.1*. Then, the following hold:*

1. *There exists an* r_1^{-1}*-allpass completion of* $\boldsymbol{H}_{22} = \boldsymbol{Q}_{22}$ *of the form*

$$\bar{\boldsymbol{H}} = \begin{bmatrix} \bar{\boldsymbol{H}}_{11} & \bar{\boldsymbol{H}}_{12} \\ \bar{\boldsymbol{H}}_{21} & \boldsymbol{H}_{22} \end{bmatrix} = \begin{bmatrix} \hat{\boldsymbol{R}}+\bar{\boldsymbol{Q}}_{11} & \bar{\boldsymbol{Q}}_{12} \\ \bar{\boldsymbol{Q}}_{21} & \boldsymbol{Q}_{22} \end{bmatrix} := \begin{bmatrix} \hat{\boldsymbol{R}} & 0 \\ 0 & 0 \end{bmatrix} + \bar{\boldsymbol{Q}}_a$$

   with $\bar{\boldsymbol{Q}}_a \in \mathcal{RH}_\infty$ *such that* $\bar{\boldsymbol{H}}\bar{\boldsymbol{H}}^\sim = \bar{\boldsymbol{H}}^\sim\bar{\boldsymbol{H}} = r_1^{-2} I_{p+m-2}$, $\hat{\boldsymbol{R}} \in \mathcal{RH}_\infty^{-(p-1)\times(m-1)}$ *and* $\bar{\boldsymbol{Q}}_{12}^{-1}, \bar{\boldsymbol{Q}}_{21}^{-1} \in \mathcal{RH}_\infty$.
2. *The set of all* $\bar{\boldsymbol{Q}} \in \mathcal{H}_\infty^{(p-1)\times(m-1)}$ *such that* $\|\hat{\boldsymbol{R}}+\bar{\boldsymbol{Q}}\|_\infty \le r_1^{-1}$ *is given by*

$$\bar{\mathcal{S}}_1 = \mathcal{F}_l(\bar{\boldsymbol{Q}}_a, r_1\mathcal{BH}_\infty^{(p-1)\times(m-1)}).$$

3. *There exist inner matrices* $\boldsymbol{X}$ *and* $\boldsymbol{Y}$ *and a scalar allpass function* $\boldsymbol{a}$ *such that*

$$\mathcal{T}_1 = \boldsymbol{Y}\ \mathrm{diag}(r_1^{-1}\boldsymbol{a}, \mathcal{F}_l(\bar{\boldsymbol{H}}, r_1\mathcal{BH}_\infty^{(p-1)\times(m-1)}))\boldsymbol{X}. \tag{10}$$

   Further,

$$\mathcal{F}_l(\bar{\boldsymbol{H}}, r_1\mathcal{BH}_\infty^{(p-1)\times(m-1)}) = \{\hat{\boldsymbol{R}}+\bar{\boldsymbol{Q}} : \bar{\boldsymbol{Q}} \in \mathcal{H}_\infty^{(p-1)\times(m-1)}, \|\hat{\boldsymbol{R}}+\bar{\boldsymbol{Q}}\|_\infty \le r_1^{-1}\}. \tag{11}$$

Proof.

1. The construction of the r_1^{-1}-allpass completion $\bar{\boldsymbol{H}}$ is an exercise in standard factorization theory which can be performed using either transfer function or state-space techniques [3]. The details can be found in [14], [11], [12], [10].
2. The fact that r_1^{-1} is a suboptimal level of $\hat{\boldsymbol{R}}$ follows from part 1 since $\bar{\boldsymbol{Q}}_{11} \in \mathcal{RH}_\infty$ and $\|\hat{\boldsymbol{R}}+\bar{\boldsymbol{Q}}_{11}\|_\infty \le \|\bar{\boldsymbol{H}}\|_\infty = r_1^{-1}$. Part 2 now follows from part 1 and [5, Theorem 3.2] since $\bar{\boldsymbol{Q}}_a, \bar{\boldsymbol{Q}}_{12}^{-1}, \bar{\boldsymbol{Q}}_{21}^{-1}, \hat{\boldsymbol{R}}^\sim \in \mathcal{RH}_\infty$ by construction.
3. Define $\boldsymbol{V}_\perp = \boldsymbol{H}_{12}\bar{\boldsymbol{H}}_{12}^{-1} \in \mathcal{RH}_\infty$ and $\boldsymbol{W}_\perp = \boldsymbol{H}_{21}^\sim\bar{\boldsymbol{H}}_{21}^{-\sim} \in \mathcal{RH}_\infty^-$. Since $r_1\boldsymbol{H}$ and $r_1\bar{\boldsymbol{H}}$ are allpass, a manipulation will verify that $\boldsymbol{V}_\perp^\sim\boldsymbol{H}_{11}\boldsymbol{W}_\perp = \bar{\boldsymbol{H}}_{11}$ and

$$\bar{\boldsymbol{H}}_{21}\bar{\boldsymbol{H}}_{21}^\sim = r_1^{-2} I_{m-1} - \boldsymbol{H}_{22}\boldsymbol{H}_{22}^\sim = \boldsymbol{H}_{21}\boldsymbol{H}_{21}^\sim \implies \boldsymbol{W}_\perp^\sim\boldsymbol{W}_\perp = I_{m-1}.$$

Similarly,

$$\bar{\boldsymbol{H}}_{12}^{\sim}\bar{\boldsymbol{H}}_{12} = r_1^{-2} I_{p-1} - \boldsymbol{H}_{22}^{\sim}\boldsymbol{H}_{22} = \boldsymbol{H}_{12}^{\sim}\boldsymbol{H}_{12} \quad \Longrightarrow \quad \boldsymbol{V}_{\perp}^{\sim}\boldsymbol{V}_{\perp} = I_{p-1}.$$

Hence there exist $\boldsymbol{v} \in \mathcal{RH}_{\infty}^{p\times 1}$, $\boldsymbol{w}^{\sim} \in \mathcal{RH}_{\infty}^{1\times m}$ such that

$$\boldsymbol{V} = \begin{bmatrix} \boldsymbol{v} & \boldsymbol{V}_{\perp} \end{bmatrix}, \qquad \boldsymbol{W} = \begin{bmatrix} \boldsymbol{w} & \boldsymbol{W}_{\perp} \end{bmatrix} \tag{12}$$

are square inner. Now consider the product

$$\begin{aligned} \begin{bmatrix} \boldsymbol{V}^{\sim} & 0 \\ 0 & I \end{bmatrix} \boldsymbol{H} \begin{bmatrix} \boldsymbol{W} & 0 \\ 0 & I \end{bmatrix} &= \begin{bmatrix} \boldsymbol{v}^{\sim}\boldsymbol{H}_{11}\boldsymbol{w} & \boldsymbol{v}^{\sim}\boldsymbol{H}_{11}\boldsymbol{W}_{\perp} & \boldsymbol{v}^{\sim}\boldsymbol{H}_{12} \\ \boldsymbol{V}_{\perp}^{\sim}\boldsymbol{H}_{11}\boldsymbol{w} & \boldsymbol{V}_{\perp}^{\sim}\boldsymbol{H}_{11}\boldsymbol{W}_{\perp} & \boldsymbol{V}_{\perp}^{\sim}\boldsymbol{H}_{12} \\ \boldsymbol{H}_{21}\boldsymbol{w} & \boldsymbol{H}_{21}\boldsymbol{W}_{\perp} & \boldsymbol{H}_{22} \end{bmatrix} \\ &= \begin{bmatrix} \boldsymbol{v}^{\sim}\boldsymbol{H}_{11}\boldsymbol{w} & \boldsymbol{Y}_{12} & \boldsymbol{Y}_{13} \\ \boldsymbol{Y}_{21} & \bar{\boldsymbol{H}}_{11} & \bar{\boldsymbol{H}}_{12} \\ \boldsymbol{Y}_{31} & \bar{\boldsymbol{H}}_{21} & \boldsymbol{H}_{22} \end{bmatrix}. \end{aligned} \tag{13}$$

Since $\boldsymbol{V}, \boldsymbol{W}, r_1\boldsymbol{H}$, and $r_1\bar{\boldsymbol{H}}$ are allpass, all $\boldsymbol{Y}_{ij}$ terms in (13) are equal to zero, and $\boldsymbol{v}^{\sim}\boldsymbol{H}_{11}\boldsymbol{w}$ is r_1^{-1}-allpass, i.e., $\boldsymbol{v}^{\sim}\boldsymbol{H}_{11}\boldsymbol{w} = r_1^{-1}\boldsymbol{a}$ for some scalar allpass rational function $\boldsymbol{a}$. A simple manipulation using (13) verifies that

$$\mathcal{F}_l(\boldsymbol{H}, r_1\mathcal{BH}_{\infty}^{(p-1)\times(m-1)}) = \boldsymbol{V}\ \mathrm{diag}(r_1^{-1}\boldsymbol{a}, \mathcal{F}_l(\bar{\boldsymbol{H}}, r_1\mathcal{BH}_{\infty}^{(p-1)\times(m-1)}))\boldsymbol{W}^{\sim}. \tag{14}$$

Since

$$\mathcal{F}_l(\boldsymbol{H}, r_1\mathcal{BH}_{\infty}^{(p-1)\times(m-1)}) = \{\boldsymbol{R} + \boldsymbol{Q} : \boldsymbol{Q} \in \mathcal{H}_{\infty},\ \|\boldsymbol{R} + \boldsymbol{Q}\|_{\infty} = r_1^{-1}\}$$

we may write from (8), (9), and (14)

$$\begin{aligned} \mathcal{T}_1 &= (\boldsymbol{U} + \boldsymbol{M}\mathcal{F}_l(\boldsymbol{Q}_a, r_1\mathcal{BH}_{\infty}^{(p-1)\times(m-1)}))\tilde{\boldsymbol{M}} \\ &= \boldsymbol{M}(\boldsymbol{M}^{\sim}\boldsymbol{U} + \mathcal{F}_l(\boldsymbol{Q}_a, r_1\mathcal{BH}_{\infty}^{(p-1)\times(m-1)}))\tilde{\boldsymbol{M}} \\ &= \boldsymbol{M}\mathcal{F}_l(\boldsymbol{H}, r_1\mathcal{BH}_{\infty}^{(p-1)\times(m-1)})\tilde{\boldsymbol{M}} \\ &= \boldsymbol{Y}\ \mathrm{diag}(r_1^{-1}\boldsymbol{a}, \mathcal{F}_l(\bar{\boldsymbol{H}}, r_1\mathcal{BH}_{\infty}^{(p-1)\times(m-1)}))\boldsymbol{X} \end{aligned}$$

as required, with $\boldsymbol{Y} := \boldsymbol{MV}$ and $\boldsymbol{X} := \boldsymbol{W}^{\sim}\tilde{\boldsymbol{M}}$ square inner. This proves the theorem. □

Remark 4.2. The theorem shows that every optimal interpolating function $\boldsymbol{T} \in \mathcal{T}_1$ has a partial pseudosingular value decomposition with largest "singular value" r_1^{-1} and corresponding left and right "singular vectors" $\boldsymbol{Mv}$ and $\boldsymbol{w}^{\sim}\tilde{\boldsymbol{M}}$, respectively. The two "singular vectors" corresponding to the largest "singular value" are real-rational.

In what follows we develop improved robust stability properties for the set of controllers which minimize the pair $\{s_1(\mathcal{T}), s_2(\mathcal{T})\}$ with respect to lexicographic ordering. We denote the set of interpolating functions with this property by $\mathcal{T}_2$ and the corresponding set of controllers by $\mathcal{K}_2$. Clearly, $\mathcal{T}_2 \subseteq \mathcal{T}_1 \subseteq \mathcal{T}$ and $\mathcal{K}_2 \subseteq \mathcal{K}_1 \subseteq \mathcal{K}$. We refer to $\mathcal{T}_2$ ($\mathcal{K}_2$) as the superoptimal set of interpolating functions (controllers) with respect to the first two levels. The following lemma gives a parametrization of the set $\mathcal{T}_2$.

LEMMA 4.3. *$\mathcal{T}_2$ may be parametrized as*

$$\mathcal{T}_2 = \boldsymbol{Y}_1\ \mathrm{diag}(s_1\boldsymbol{a}, s_2\boldsymbol{b}, \check{\boldsymbol{R}} + \mathcal{S}_2)\boldsymbol{X}_1 \tag{15}$$

in which s_1 and s_2 denote the first two superoptimal levels of $\mathcal{T}$ with $s_1 = r_1^{-1}$, $\boldsymbol{Y}_1$ and $\boldsymbol{X}_1$ are square inner matrices, $\boldsymbol{a}$ and $\boldsymbol{b}$ are scalar rational allpass functions, $\check{\boldsymbol{R}} \in \mathcal{RH}_\infty^{-(p-2)\times(m-2)}$, and $\mathcal{S}_2 = \{\check{\boldsymbol{Q}} \in \mathcal{H}_\infty : \|\check{\boldsymbol{R}} + \check{\boldsymbol{Q}}\|_\infty \le s_2\}$. Further, the first column (row) of $\boldsymbol{Y}_1$ ($\boldsymbol{X}_1$) is identical to the first column (row) of $\boldsymbol{Y}$ ($\boldsymbol{X}$) defined in Theorem 4.2.

Proof. From (10), since $\boldsymbol{X}$ and $\boldsymbol{Y}$ are square inner and $\boldsymbol{a}$ is allpass, we have

$$s_2 = \inf \ \{ \ \|\mathcal{F}_l(\bar{\boldsymbol{H}}, \boldsymbol{U})\|_\infty : \boldsymbol{U} \in r_1 \mathcal{BH}_\infty^{(p-1)\times(m-1)}\}.$$

Using (11) this is equivalent to

$$s_2 = \inf \ \{\|\hat{\boldsymbol{R}} + \hat{\boldsymbol{Q}}\|_\infty : \hat{\boldsymbol{Q}} \in \mathcal{H}_\infty^{(p-1)\times(m-1)}\} = \|\Gamma_{\hat{\boldsymbol{R}}(-s)}\|,$$

where the second equality follows from Nehari's theorem. Clearly, $\mathcal{T}_2$ may be obtained by replacing $\mathcal{F}_l(\bar{\boldsymbol{H}}, r_1 \mathcal{BH}_\infty^{(p-1)\times(m-1)})$ by the set

$$\{\hat{\boldsymbol{R}} + \hat{\boldsymbol{Q}} : \hat{\boldsymbol{Q}} \in \mathcal{H}_\infty^{(p-1)\times(m-1)} \ , \ \ \|\hat{\boldsymbol{R}} + \hat{\boldsymbol{Q}}\|_\infty = s_2\}, \tag{16}$$

which reveals the recursive character of the problem. Since $\hat{\boldsymbol{R}} \in \mathcal{RH}_\infty^{-(p-1)\times(m-1)}$, a parametrization of (16) may be obtained from Theorem 4.1 with $\boldsymbol{R}$ replaced by $\hat{\boldsymbol{R}}$ and s_1 replaced by s_2. That is, Theorem 4.1 guarantees that there exists an s_2-allpass embedding of $\hat{\boldsymbol{R}}$ of the form

$$\hat{\boldsymbol{H}} = \begin{bmatrix} \hat{\boldsymbol{H}}_{11} & \hat{\boldsymbol{H}}_{12} \\ \hat{\boldsymbol{H}}_{21} & \hat{\boldsymbol{H}}_{22} \end{bmatrix} := \begin{bmatrix} \hat{\boldsymbol{R}} + \hat{\boldsymbol{Q}}_{11} & \hat{\boldsymbol{Q}}_{12} \\ \hat{\boldsymbol{Q}}_{21} & \hat{\boldsymbol{Q}}_{22} \end{bmatrix} = \begin{bmatrix} \hat{\boldsymbol{R}} & 0 \\ 0 & 0 \end{bmatrix} + \hat{\boldsymbol{Q}}_a$$

in which $\hat{\boldsymbol{Q}}_{11} \in \mathcal{RH}_\infty^{(p-1)\times(m-1)}$, $\hat{\boldsymbol{Q}}_{12} \in \mathcal{RH}_\infty^{(p-1)\times(p-2)}$, $\hat{\boldsymbol{Q}}_{21} \in \mathcal{RH}_\infty^{(m-2)\times(m-1)}$, and $\hat{\boldsymbol{Q}}_{22} \in \mathcal{RH}_\infty^{(m-2)\times(p-2)}$ and such that $\hat{\boldsymbol{H}}\hat{\boldsymbol{H}}^\sim = \hat{\boldsymbol{H}}^\sim \hat{\boldsymbol{H}} = s_2^2 I_{p+m-3}$ and $\|\hat{\boldsymbol{H}}_{22}\|_\infty = \|\hat{\boldsymbol{Q}}_{22}\|_\infty < s_2$. Moreover, the set of all $\hat{\boldsymbol{Q}}$ in (16) is generated by $\mathcal{F}_l(\hat{\boldsymbol{Q}}_a, s_2^{-1}\mathcal{BH}_\infty^{(p-2)\times(m-2)})$. Next, we apply Theorem 4.2 to obtain an s_2-allpass embedding of $\hat{\boldsymbol{H}}_{22} = \hat{\boldsymbol{Q}}_{22}$ of the form

$$\check{\boldsymbol{H}} = \begin{bmatrix} \check{\boldsymbol{H}}_{11} & \check{\boldsymbol{H}}_{12} \\ \check{\boldsymbol{H}}_{21} & \hat{\boldsymbol{H}}_{22} \end{bmatrix} = \begin{bmatrix} \check{\boldsymbol{R}} + \check{\boldsymbol{Q}}_{11} & \check{\boldsymbol{Q}}_{12} \\ \check{\boldsymbol{Q}}_{21} & \hat{\boldsymbol{Q}}_{22} \end{bmatrix} = \begin{bmatrix} \check{\boldsymbol{R}} & 0 \\ 0 & 0 \end{bmatrix} + \check{\boldsymbol{Q}}_a$$

with $\check{\boldsymbol{Q}}_{11} \in \mathcal{RH}_\infty^{(p-2)\times(m-2)}$, $\check{\boldsymbol{Q}}_{12} \in \mathcal{RH}_\infty^{(p-2)\times(p-2)}$, and $\check{\boldsymbol{Q}}_{21} \in \mathcal{RH}_\infty^{(m-2)\times(m-2)}$. Also, $\check{\boldsymbol{R}} \in \mathcal{RH}_\infty^{-(p-2)\times(m-2)}$, $\check{\boldsymbol{Q}}_{12}^{-1}, \check{\boldsymbol{Q}}_{21}^{-1} \in \mathcal{RH}_\infty$, and $\check{\boldsymbol{H}}\check{\boldsymbol{H}}^\sim = \check{\boldsymbol{H}}^\sim \check{\boldsymbol{H}} = s_2^2 I_{p+m-4}$. Further, Theorem 4.2 shows that the set of all $\check{\boldsymbol{Q}} \in \mathcal{H}_\infty$ such that $\|\check{\boldsymbol{R}} + \check{\boldsymbol{Q}}\|_\infty \le s_2$ is given by $\mathcal{S}_2 = \mathcal{F}_l(\check{\boldsymbol{Q}}_a, s_2^{-1}\mathcal{BH}_\infty^{(p-2)\times(m-2)})$. The proof of Theorem 4.2(3) may now be repeated step by step to show that there exist square inner matrices $\boldsymbol{V}_1$ and $\boldsymbol{W}_1^\sim$ and a scalar rational allpass function $\boldsymbol{b}$ such that

$$\mathcal{F}_l(\hat{\boldsymbol{H}}, s_2^{-1}\mathcal{BH}_\infty^{(p-2)\times(m-2)}) = \boldsymbol{V}_1 \ \mathrm{diag}(s_2\boldsymbol{b}, \mathcal{F}_l(\check{\boldsymbol{H}}, s_2^{-1}\mathcal{BH}_\infty^{(p-2)\times(m-2)}))\boldsymbol{W}_1^\sim.$$

Hence from (10), $\mathcal{T}_2 = \boldsymbol{Y}_1 \ \mathrm{diag}(s_1\boldsymbol{a}, s_2\boldsymbol{b}, \mathcal{F}_l(\check{\boldsymbol{H}}, s_2^{-1}\mathcal{BH}_\infty^{(p-2)\times(m-2)}))\boldsymbol{X}_1$, where we have defined the square inner matrix functions

$$\boldsymbol{Y}_1 = \boldsymbol{Y} \ \mathrm{diag}(1, \boldsymbol{V}_1), \qquad \boldsymbol{X}_1 = \ \mathrm{diag}(1, \boldsymbol{W}_1^\sim)\boldsymbol{X}.$$

Equation (10) agrees with the parametrization in (15). Note further that the first column (row) of $\boldsymbol{Y}_1$ ($\boldsymbol{X}_1$) is identical with the first column (row) of $\boldsymbol{Y}$ ($\boldsymbol{X}$) and that for every $\check{\boldsymbol{Q}} \in \mathcal{S}_2$, $\|\check{\boldsymbol{R}} + \check{\boldsymbol{Q}}\|_\infty \le s_2$, as required. ∎

In the next part of the section we identify the set of all $\boldsymbol{\Delta} \in \partial\mathcal{D}_{r_1}(\boldsymbol{G})$ which destabilize $(\boldsymbol{G}, \boldsymbol{K})$ for *every* $\boldsymbol{K} \in \mathcal{K}_1$. We refer to such $\boldsymbol{\Delta}$'s as *uniformly destabilizing.*

LEMMA 4.4. *There exists* $\boldsymbol{\Delta} \in \partial\mathcal{D}_{r_1}(\boldsymbol{G})$ *such that* $(\boldsymbol{G}+\boldsymbol{\Delta}, \boldsymbol{K}) \notin \mathcal{S}$ *for every* $\boldsymbol{K} \in \mathcal{K}_1$. *Furthermore,* $\boldsymbol{\Delta}$ *can be chosen to be a stable real-rational matrix function.*

Proof. Pick any $\omega_o \in \mathcal{R}$ and define

$$\Delta_o = \boldsymbol{X}^{\sim}(j\omega_o)\ \mathrm{diag}(r_1/\boldsymbol{a}(j\omega_o), 0)\boldsymbol{Y}^{\sim}(j\omega_o) \in \mathcal{C}^{m\times p}.$$

Then $\eta(\boldsymbol{G}) = \eta(\boldsymbol{G}+\Delta_o)$, $\|\Delta_o\|_\infty = r_1$ and so $\Delta_o \in \partial\mathcal{D}_{r_1}(\boldsymbol{G})$. Let $\boldsymbol{T} \in \mathcal{T}_1$ be any optimal interpolating function so that $\boldsymbol{T} = \boldsymbol{Y}\ \mathrm{diag}(r_1^{-1}\boldsymbol{a}, \boldsymbol{\Phi})\boldsymbol{X}$, where $\boldsymbol{\Phi} = \mathcal{F}_l(\bar{\boldsymbol{H}}, r_1\boldsymbol{\Psi})$ for some $\boldsymbol{\Psi} \in \mathcal{BH}_\infty^{(p-1)\times(m-1)}$ from Theorem 4.2. Then it is simple to show that $\det\ [I - \Delta_o\boldsymbol{T}(j\omega_o)] = 0$ for any $\boldsymbol{\Psi}$ and so $\det\ [I - \Delta_o\boldsymbol{T}(j\omega_o)] = 0$ for all $\boldsymbol{T} \in \mathcal{T}_1$. Since

$$\det\{I - (\boldsymbol{G}+\Delta_o)\boldsymbol{K}(j\omega_o)\} = \det\{I - \boldsymbol{G}\boldsymbol{K}(j\omega_o)\}\det\{I - \Delta_o\boldsymbol{T}(j\omega_o)\} = 0,$$

$(\boldsymbol{G}+\Delta_o, \boldsymbol{K}) \notin \mathcal{S}$ for every $\boldsymbol{K} \in \mathcal{K}_1$ by the generalized Nyquist theorem [25].

In the final part of the proof, we construct a stable real-rational $\boldsymbol{\Delta}$ such that $\boldsymbol{\Delta} \in \partial\mathcal{D}_{r_1}(\boldsymbol{G})$ and $\boldsymbol{\Delta}(j\omega_o) = \Delta_o$. Define the unit vectors

$$y_1 = \boldsymbol{a}(-j\omega_o)\boldsymbol{v}'(j\omega_o)\boldsymbol{M}'(j\omega_o) \in \mathcal{C}^{1\times p}, \qquad x_1 = \tilde{\boldsymbol{M}}'(j\omega_o)\boldsymbol{w}(j\omega_o) \in \mathcal{C}^{m\times 1}$$

so that x_1 is the first column of $\boldsymbol{X}^{\sim}(j\omega_o)$ and $y_1/\boldsymbol{a}(-j\omega_o)$ is the first row of $\boldsymbol{Y}^{\sim}(j\omega_o)$. Next, express each component of x_1 and y_1 as

$$(x_1)_i = \hat{x}_i e^{j\phi_i}, \ \ (y_1)_i = \hat{y}_i e^{j\theta_i},$$

where $\hat{y}_i, \hat{x}_i \in \mathcal{R}$ and $\phi_i, \theta_i \in [0, \pi)$. (This fixes the signs of $\hat{y}_i, \hat{x}_i$.) Define the inner vector functions

$$\hat{\boldsymbol{y}}(s) = \begin{bmatrix} \hat{y}_1\frac{s-\alpha_1}{s+\alpha_1} \\ \vdots \\ \hat{y}_p\frac{s-\alpha_p}{s+\alpha_p} \end{bmatrix}, \qquad \hat{\boldsymbol{x}}(s) = \begin{bmatrix} \hat{x}_1\frac{s-\beta_1}{s+\beta_1} \\ \vdots \\ \hat{x}_m\frac{s-\beta_m}{s+\beta_m} \end{bmatrix}$$

so that

$$\arg\left\{\frac{j\omega_o - \alpha_i}{j\omega_o + \alpha_i}\right\} = \theta_i\ ,\ i = 1, 2, \ldots, p, \ \ \arg\left\{\frac{j\omega_o - \beta_i}{j\omega_o + \beta_i}\right\} = \phi_i,\ i = 1, 2, \ldots, m,$$

with $\alpha_i, \beta_i \in \mathcal{R}_+$ for all i. (If $\theta_i = 0$ or $\phi_i = 0$, we simply replace the ith entry of $\hat{\boldsymbol{y}}$ or $\hat{\boldsymbol{x}}$ by $\hat{x}_i$ or $\hat{y}_i$, respectively.) Next, define the $\mathcal{RH}_\infty$ function $\boldsymbol{\Delta} = r_1\hat{\boldsymbol{x}}\hat{\boldsymbol{y}}^T \in \mathcal{RH}_\infty^{m\times p}$ and note that $\boldsymbol{\Delta}(j\omega_o) = \Delta_o = r_1 x_1 y_1^T$ by construction. Since $\boldsymbol{\Delta} \in \mathcal{RH}_\infty$, $\eta(\boldsymbol{G}) = \eta(\boldsymbol{G}+\boldsymbol{\Delta})$. Finally, note that $\|\boldsymbol{\Delta}\|_\infty = r_1$, which implies that $\boldsymbol{\Delta} \in \partial\mathcal{D}_{r_1}(\boldsymbol{G})$. ∎

Remark 4.3. The proof is an adaptation of a result in [25]. Indeed, it is not surprising that (real-rational) destabilizing perturbations exist on $\partial\mathcal{D}_{r_1}(\boldsymbol{G})$. The new information supplied by Lemma 4.4 is that (real-rational) boundary perturbations exist which are destabilizing for *every* maximally robust controller $\boldsymbol{K} \in \mathcal{K}_1$.

Denote by $\boldsymbol{x}^T$ and $\boldsymbol{y}$ the first row and column of $\boldsymbol{X}$ and $\boldsymbol{Y}$, respectively, defined in Theorem 4.2. Then, all uniformly destabilizing perturbations constructed in Lemma 4.4 have the property that $|\boldsymbol{x}^T\boldsymbol{\Delta}\boldsymbol{y}(j\omega)| = r_1$ for some $\omega \in \mathcal{R}$. Moreover, such perturbations can be constructed for *every* $\omega \in \mathcal{R}$. The next result shows that

condition $|\boldsymbol{x}^T\boldsymbol{\Delta}\boldsymbol{y}(j\omega)| = r_1$ is necessary for a $\boldsymbol{\Delta} \in \partial\mathcal{D}_{r_1}(\boldsymbol{G})$ to be destabilizing for every $\boldsymbol{K} \in \mathcal{K}_1$.

LEMMA 4.5. *Let $\boldsymbol{\Delta} \in \partial\mathcal{D}_{r_1}(\boldsymbol{G})$ be a destabilizing perturbation of $\boldsymbol{G}$ for every $\boldsymbol{K} \in \mathcal{K}_1$. Then there exists an $\omega \in \mathcal{R}$, such that*

$$|\boldsymbol{x}^T(j\omega)\boldsymbol{\Delta}(j\omega)\boldsymbol{y}(j\omega)| = r_1. \tag{17}$$

Proof. The set $\mathcal{T}_1$ is given by (10). Pick any $\boldsymbol{\Psi} \in r_1\mathcal{BH}_\infty^{(p-1)\times(m-1)}$ so that $\|\boldsymbol{\Psi}\|_\infty < r_1$. Since $\bar{\boldsymbol{H}}$ is r_1^{-1}-allpass, $\boldsymbol{\Phi} = \mathcal{F}_l(\bar{\boldsymbol{H}}, \boldsymbol{\Psi})$ satisfies $\|\boldsymbol{\Phi}\|_\infty < r_1^{-1}$. Since $\boldsymbol{\Delta}$ destabilizes $\boldsymbol{G}$ for every $\boldsymbol{K} \in \mathcal{K}_1$, it is also destabilizing for

$$\boldsymbol{K} = (\boldsymbol{U} + \boldsymbol{M}\mathcal{F}_l(\boldsymbol{Q}_a, \boldsymbol{\Psi}))(\boldsymbol{V} + \boldsymbol{N}\mathcal{F}_l(\boldsymbol{Q}_a, \boldsymbol{\Psi}))^{-1} \in \mathcal{K}_1,$$

corresponding to the interpolation function $\boldsymbol{T} = \boldsymbol{Y}\ \mathrm{diag}(r_1^{-1}\boldsymbol{a}, \boldsymbol{\Phi})\boldsymbol{X}$. Since $\alpha\boldsymbol{\Delta}$ is a stabilizing perturbation of $\boldsymbol{G}$ for every $\alpha \in [0, 1)$, there exists $\omega_o \in \mathcal{R}$ such that $\det\{I_m - \boldsymbol{\Delta}(j\omega_o)\boldsymbol{T}(j\omega_o)\} = 0$ or equivalently that

$$\det\left\{I_m - \boldsymbol{\Delta}(j\omega_o)\boldsymbol{Y}(j\omega_o)\begin{bmatrix} r_1^{-1}\boldsymbol{a}(j\omega_o) & 0 \\ 0 & \boldsymbol{\Phi}(j\omega_o)\end{bmatrix}\boldsymbol{X}(j\omega_o)\right\} = 0,$$

which implies that

$$\det\left\{I_m - \boldsymbol{X}(j\omega_o)\boldsymbol{\Delta}(j\omega_o)\boldsymbol{Y}(j\omega_o)\begin{bmatrix} r_1^{-1}\boldsymbol{a}(j\omega_o) & 0 \\ 0 & \boldsymbol{\Phi}(j\omega_o)\end{bmatrix}\right\} = 0. \tag{18}$$

Define

$$\begin{bmatrix} x_1^T \\ X_\perp^T \end{bmatrix} = \boldsymbol{X}(j\omega_o),$$

$\begin{bmatrix} y_1 & Y_\perp \end{bmatrix} = \boldsymbol{Y}(j\omega_o)$, and

$$\tilde{\Delta} = \begin{bmatrix} \tilde{\delta}_{11} & \tilde{\Delta}_{12} \\ \tilde{\Delta}_{21} & \tilde{\Delta}_{22}\end{bmatrix} = \begin{bmatrix} x_1^T \\ X_\perp^T \end{bmatrix}\boldsymbol{\Delta}(j\omega_o)\begin{bmatrix} y_1 & \boldsymbol{a}(j\omega_o) & Y_\perp \end{bmatrix}.$$

Then (18) may be written as

$$\det\begin{bmatrix} 1 - r_1^{-1}\tilde{\delta}_{11} & -\tilde{\Delta}_{12}\boldsymbol{\Phi}(j\omega_o) \\ -r_1^{-1}\tilde{\Delta}_{21} & I_{m-1} - \tilde{\Delta}_{22}\boldsymbol{\Phi}(j\omega_o)\end{bmatrix} = 0. \tag{19}$$

Next, we show that $\tilde{\delta}_{11} = r_1$. Suppose for contradiction that

$$1 - r_1^{-1}\tilde{\delta}_{11} \neq 0. \tag{20}$$

Then (19) implies that

$$\begin{aligned} & (1 - r_1^{-1}\tilde{\delta}_{11})\ \det\left\{I_{m-1} - \tilde{\Delta}_{22}\boldsymbol{\Phi} - r_1^{-1}\tilde{\Delta}_{21}\left(1 - r_1^{-1}\tilde{\delta}_{11}\right)^{-1}\tilde{\Delta}_{12}\boldsymbol{\Phi}(j\omega_o)\right\} = 0, \\ \Rightarrow\quad & \det\left\{I_{m-1} - \left(\tilde{\Delta}_{22} + r_1^{-1}\tilde{\Delta}_{21}\left(1 - r_1^{-1}\tilde{\delta}_{11}\right)^{-1}\tilde{\Delta}_{12}\right)\boldsymbol{\Phi}(j\omega_o)\right\} = 0, \\ \Rightarrow\quad & \det\left\{I_{m-1} - \mathcal{F}_u(\tilde{\Delta}, r_1^{-1})\boldsymbol{\Phi}(j\omega_o)\right\} = 0. \end{aligned}$$

Since $1 - r_1^{-1}\tilde{\delta}_{11} \neq 0$ by assumption, the upper linear fractional map is well-posed; moreover, $\bar{\sigma}(\tilde{\Delta}) = r_1$ which implies that $\bar{\sigma}(\mathcal{F}_u(\tilde{\Delta}, r_1^{-1})) \leq r_1$ [5]. Also, since $\bar{\sigma}(\boldsymbol{\Phi}(j\omega_o)) < r_1^{-1}$, we have that $\bar{\sigma}(\mathcal{F}_u(\tilde{\Delta}, r_1^{-1})\boldsymbol{\Phi}(j\omega_o)) < 1$. Hence,

$$\underline{\sigma}\left(I - \mathcal{F}_u(\tilde{\Delta}, r_1^{-1})\boldsymbol{\Phi}(j\omega_o)\right) \geq 1 - \bar{\sigma}\left(\mathcal{F}_u(\tilde{\Delta}, r_1^{-1})\boldsymbol{\Phi}(j\omega_o)\right) > 0,$$

and thus $\det\{I - \mathcal{F}_u(\tilde{\Delta}, r_1^{-1})\boldsymbol{\Phi}(j\omega_o)\} \neq 0$, contradicting (20). Hence,

$$\tilde{\delta}_{11} = r_1 \Rightarrow | \boldsymbol{x}^T(j\omega_o)\boldsymbol{\Delta}(j\omega_o)\boldsymbol{y}(j\omega_o) | = r_1$$

since $|\boldsymbol{a}(j\omega_o)| = 1$. □

Remark 4.4. Lemma 4.5 above shows that every $\boldsymbol{\Delta} \in \partial\mathcal{D}_{r_1}(\boldsymbol{G})$ which is destabilizing for all $\boldsymbol{K} \in \mathcal{K}_1$ satisfies $|\boldsymbol{x}^T(j\omega_o)\boldsymbol{\Delta}(j\omega_o)\boldsymbol{y}(j\omega_o)| = r_1$ for some $\omega_o \in \mathcal{R}$. Define the inner product of two matrices of compatible dimensions A and B as $\langle A, B\rangle = \text{trace}(A'B)$. Then, (17) says that every $\boldsymbol{\Delta} \in \partial\mathcal{D}_{r_1}(\boldsymbol{G})$ which is destabilizing for all $\boldsymbol{K} \in \mathcal{K}_1$ satisfies $|\langle \boldsymbol{y}(j\omega_o)\boldsymbol{x}^T(j\omega_o), \boldsymbol{\Delta}(j\omega_o)\rangle| = r_1$, i.e., that it has projection of magnitude r_1 in the "most critical direction" $\boldsymbol{y}(j\omega_o)\boldsymbol{x}^T(j\omega_o)$ for some $\omega_o \in \mathcal{R}$. Moreover, the proof of Lemma 4.4 shows that all frequencies $\omega \in \mathcal{R}$ are "equally critical," in the sense that the generalized Nyquist criterion can be violated at *any* $\omega \in \mathcal{R}$. This implies that it is futile to attempt to extend the uncertainty set guaranteed to be stabilized by a subset of $\mathcal{K}_1$ in the (frequency-dependent) direction $\boldsymbol{y}(j\omega)\boldsymbol{x}^T(j\omega)$, $\omega \in \mathcal{R}$. Suppose now that we impose a "structure" on the perturbation set of the form

$$| \boldsymbol{x}^T(j\omega)\boldsymbol{\Delta}(j\omega)\boldsymbol{y}(j\omega) | \leq r_1(1-\delta) \quad \text{for all } \omega \in \mathcal{R}$$

for some (fixed) $\delta \in [0, 1)$. Note in view of Lemmas 4.4 and 4.5 that this bound is assumed to be uniform in ω. In other words, we constrain the perturbation set so that $\boldsymbol{\Delta}$ cannot have a projection of magnitude larger that $r_1(1-\delta)$ in the most critical direction for all $\omega \in \mathcal{R}$. Formally, define the set

$$\mathcal{E}(\delta, \mu) = \{\boldsymbol{\Delta} \in \mathcal{D}_\mu(\boldsymbol{G}) : \|\boldsymbol{x}^T\boldsymbol{\Delta}\boldsymbol{y}\|_\infty \leq r_1(1-\delta)\}, \tag{21}$$

where $\mathcal{D}_\mu(\boldsymbol{G})$ is defined in (1). Then, for each $\delta \in (0, 1]$ we want to find the set of controllers $\mathcal{K}_\delta \subseteq \mathcal{K}_1$ which maximize $\mu = \mu(\delta)$ under the constraint that $\boldsymbol{G} + \boldsymbol{\Delta}$ is stable for all $\boldsymbol{\Delta} \in \mathcal{D}_{r_1}(\boldsymbol{G}) \cup \mathcal{E}(\delta, \mu)$. Suppose that the maximum μ is attained and is given by $\mu^*(\delta)$. It is clear that $\mu^*(\delta)$ is a nondecreasing function of $\delta \in (0, 1]$. It is shown below that the sets $\mathcal{K}_\delta$ are identical for every $\delta \in (0, 1]$ and equal to $\mathcal{K}_2$. A closed-form expression of $\mu^*(\delta)$ is also obtained which involves the first two superoptimal levels of $\mathcal{T}$.

The problem formulation in the above remark is motivated by a related problem in [15]: Suppose that $A \in \mathcal{C}^{n\times n}$ is nonsingular. We know that if $\bar{\sigma}(E) = \underline{\sigma}(A)$, then $A - E$ is singular if and only if $\langle u_n v_n', E\rangle = u_n' E v_n = \underline{\sigma}(A)$, where u_n and v_n denote the singular vectors of A corresponding to $\underline{\sigma}(A)$. Also, if $\bar{\sigma}(E) < \underline{\sigma}(A)$, then $A - E$ is nonsingular. Suppose that $\bar{\sigma}(E) = \underline{\sigma}(A)$ and E is constrained to have a projection of magnitude (strictly) less than $\underline{\sigma}(A)$ in the direction $u_n v_n'$. This means that $A - E$ cannot become singular, and therefore, $\bar{\sigma}(E)$ must increase for $A - E$ to lose rank. To find how much $\bar{\sigma}(E)$ can increase before singularity occurs, we formulate the problem

$$d(\phi) = \min \{\|E\| : \det(A - E) = 0, |\langle u_n v_n', E\rangle| \leq \phi\} \tag{22}$$

for $\phi < \underline{\sigma}(A) := \sigma_n(A)$. The solution to this problem is provided by the next theorem.

THEOREM 4.6. *Let A be a square nonsingular complex matrix which has a singular value decomposition $A = U\Sigma V'$, where $\Sigma = \operatorname{diag}(\sigma_1, \ldots, \sigma_{n-1}, \sigma_n)$ with $\sigma_1 \geq \cdots \geq \sigma_{n-2} \geq \sigma_{n-1} > \sigma_n > 0$ and denote by u_n and v_n the last columns of U and V, respectively. Then all E which minimize* (22) *are given by*

$$E = U \begin{bmatrix} P_s & 0 & 0 \\ 0 & -\phi & \nu \\ 0 & \nu' & \phi \end{bmatrix} V',$$

where P_s is arbitrary except for the constraint

$$\|P_s\| \leq \sqrt{\sigma_n \sigma_{n-1} + \phi(\sigma_n - \sigma_{n-1})} \tag{23}$$

and ν is given by

$$\nu = \sqrt{(\phi + \sigma_{n-1})(\sigma_n - \phi)} e^{j\theta},\ \theta \in [0, 2\pi).$$

The minimum value of $d(\phi)$ in (22) *is given by the right-hand side (RHS) of* (23).

Proof. See [15]. In the original statement of the theorem [15], all singular values of A are assumed to be distinct. This assumption can be relaxed to the condition $\sigma_n \geq \cdots \geq \sigma_{n-2} \geq \sigma_{n-1} > \sigma_n > 0$ used here. □

Remark 4.5. Theorem 4.6 says that, provided $|\langle u_n v_n', E\rangle| \leq \phi < \sigma_n$, $\|E\|$ can increase from σ_n to $d(\phi) = \sqrt{\sigma_n \sigma_{n-1} + \phi(\sigma_n - \sigma_{n-1})}$ before $A - E$ becomes singular. In [15] this is exploited to derive robust-stability bounds for a class of additive, multiplicative, and inverse-multiplicative perturbations. Note that these results are a posteriori, i.e., they can be applied to assess the robust stability of a design only *after* a compensator has been designed. In our case, the results in [15] can be applied a priori in the sense that they can be used to characterize directly the subset of all maximally robust controllers which maximize the "radius" $\mu(\delta)$ of the uncertainty set $\mathcal{E}(\mu, \delta)$ defined in (21). The a priori character of these results in our case is a consequence of the alternative parametrization of the set of all optimal interpolation functions given in Theorem 3.2, which shows that there exists a (frequency-dependent) worst-case direction (defined by the vectors $\boldsymbol{y} = \boldsymbol{M}\boldsymbol{v}$ and $\boldsymbol{x}^T = \boldsymbol{w}^{\sim}\tilde{\boldsymbol{M}}$ in (12)) which is *identical* for *all* maximally robust controllers $\boldsymbol{K} \in \mathcal{K}_1$. The vectors $\boldsymbol{v}$ and $\boldsymbol{w}$ are associated with the maximal Schmidt pair of the Hankel operator $\Gamma_{\boldsymbol{R}(-s)}$ (see [16] for details).

In what follows, we use Theorem 4.6 to characterize the subset of all optimal controllers $\mathcal{K}_1$ which maximize $\mu^*(\delta)$. We first need a slightly different version of Theorem 4.6 which also allows us to treat the nonsquare and the singular cases.

THEOREM 4.7. *Suppose that $T \in \mathcal{C}^{p\times m}$ has a singular value decomposition, $T = U\ \operatorname{diag}(\Sigma, 0)V'$, with $\Sigma = \operatorname{diag}(\sigma_1, \sigma_2, \ldots, \sigma_t)$, $\sigma_1 > \sigma_2 \geq \sigma_3 \geq \cdots \geq \sigma_t > 0$. Let v and u be the first columns of V and U, respectively, and let $\phi < \sigma_1^{-1}$ be given. Define*

$$\mathcal{B}_d^{m\times p} = \{E \in \mathcal{C}^{m\times p}\ :\ \|E\| < d\},$$

$$\mathcal{P}(\phi) = \{E \in \mathcal{C}^{m\times p} : |v'Eu| \leq \phi\}, \tag{24}$$

and

$$d(\phi) = \sup\ \{d\ :\ \det(I_m - ET) \neq 0 \text{ for all } E \in \mathcal{B}_d^{m\times p} \cap \mathcal{P}(\phi)\}. \tag{25}$$

Then

1. *$d(\phi)$ is given by*

$$d(\phi) = \sqrt{\frac{1}{\sigma_1\sigma_2} - \phi\left(\frac{1}{\sigma_2} - \frac{1}{\sigma_1}\right)}; \tag{26}$$

2. *all $E \in \mathcal{P}(\phi)$ such that $\det(I_m - ET) = 0$ and $\|E\| = d(\phi)$ are given by*

$$E = V\begin{bmatrix} \phi & \nu & 0 \\ \nu' & -\phi & 0 \\ 0 & 0 & P_s \end{bmatrix} U', \qquad \nu = e^{j\theta}\sqrt{\left(\frac{1}{\sigma_2} + \phi\right)\left(\frac{1}{\sigma_1} - \phi\right)}, \tag{27}$$

where $\theta \in [0, 2\pi)$ and P_s is arbitrary except from the constraint $\|P_s\| \le d(\phi)$.

Proof. Introduce the partitions $U = [U_1 \;\; U_2]$ and $V = [V_1 \;\; V_2]$, where $U_1 \in \mathcal{C}^{p\times t}$ and $V_1' \in \mathcal{C}^{t\times m}$. For $E \in \mathcal{C}^{m\times p}$,

$$\det(I_m - ET) \neq 0 \Leftrightarrow \det\left(I_m - V'EU\ \operatorname{diag}(\Sigma, 0)\right) \neq 0$$

which is also equivalent to

$$\det\begin{bmatrix} I_t - V_1'EU_1\Sigma & 0 \\ -V_2'EU_1\Sigma & I_{m-t} \end{bmatrix} \neq 0 \Leftrightarrow \det(\Sigma^{-1} - V_1'EU_1) \neq 0.$$

Let $r = \sigma_1^{-1}$. The transformation

$$\mathcal{C}^{m\times p} \to \mathcal{C}^{t\times t} \;\; : \;\; E \to \tilde{E} = V_1'EU_1 \tag{28}$$

maps $\mathcal{B}_r^{m\times p}$ onto $V_1'\mathcal{B}_r^{m\times p}U_1 = \mathcal{B}_r^{t\times t}$: Clearly, for any $E \in \mathcal{B}_r^{m\times p}$, $\|\tilde{E}\| \le \|E\|$ and hence $\tilde{E} \in \mathcal{B}_r^{t\times t}$. Conversely, since all solutions to the equation $\tilde{E} = V_1'EU_1$ are given by

$$E = \begin{bmatrix} V_1 & V_2 \end{bmatrix}\begin{bmatrix} \tilde{E} & \tilde{E}_{12} \\ \tilde{E}_{21} & \tilde{E}_{22} \end{bmatrix}\begin{bmatrix} U_1' \\ U_2' \end{bmatrix}, \tag{29}$$

where $\tilde{E}_{21} \in \mathcal{C}^{(m-t)\times t}$, $\tilde{E}_{12} \in \mathcal{C}^{t\times(p-t)}$, and $\tilde{E}_{22} \in \mathcal{C}^{(m-t)\times(p-t)}$ are arbitrary matrices of the specified dimensions, every $\tilde{E} \in \mathcal{B}_r^{t\times t}$ is the image of the set

$$\left\{E = \begin{bmatrix} V_1 & V_2 \end{bmatrix}\begin{bmatrix} \tilde{E} & \tilde{E}_{12} \\ \tilde{E}_{21} & \tilde{E}_{22} \end{bmatrix}\begin{bmatrix} U_1' \\ U_2' \end{bmatrix} \;:\; \left\|\begin{bmatrix} \tilde{E} & \tilde{E}_{12} \\ \tilde{E}_{21} & \tilde{E}_{22} \end{bmatrix}\right\| < d\right\} \subseteq \mathcal{B}_r^{m\times p}$$

under (28). Moreover, since $|v_1'Eu_1| \le \phi \Leftrightarrow |\tilde{E}_{11}| \le \phi$, (25) is equivalent to

$$d(\phi) = \sup\ \{d \;:\; \det(\Sigma^{-1} - \tilde{E}) \neq 0 \text{ for all } \tilde{E} \in \{\tilde{E} \in \mathcal{B}_r^{t\times t} \;:\; |\tilde{E}_{11}| < \phi\}\},$$

where $\tilde{E}_{11}$ denotes the (1,1) element of $\tilde{E}$. By introducing suitable permutations, $d(\phi)$ may obtained by applying Theorem 4.6 and is given by the RHS of (26) which proves part 1. The set of all $\tilde{E}$ with $\|\tilde{E}\| = d(\phi)$ such that $\det(\Sigma^{-1} - \tilde{E}) = 0$ under the constraint $|\tilde{E}_{11}| \le \phi < \sigma_1^{-1}$ is given via a slight adaptation of Theorem 4.6 as

$$\tilde{E} = \begin{bmatrix} \phi & \nu & 0 \\ \nu' & -\phi & 0 \\ 0 & 0 & \tilde{P}_s \end{bmatrix}, \tag{30}$$

in which ν is defined in (27) and $\tilde{P}_s \in \mathcal{C}^{(t-2)\times(t-2)}$ is arbitrary except from the constraint $\|\tilde{P}_s\| \leq d(\phi)$. It then follows from (29) and the properties of (28) that all $E \in \mathcal{P}(\phi, d)$ such that $\det(I_m - ET) = 0$ and $\|E\| = d(\phi)$ are of the form given in (29) with $\tilde{E}$ given by (30) subject to the constraint $\|E\| \leq d(\phi)$, Thus all such E's are of the form

$$E = V \begin{bmatrix} \phi & \nu & 0 & E_{14} \\ \nu' & -\phi & 0 & E_{24} \\ 0 & 0 & \tilde{P}_s & E_{34} \\ E_{41} & E_{42} & E_{43} & E_{44} \end{bmatrix} U' \tag{31}$$

subject to the constraint $\|E\| \leq d(\phi)$. Since

$$\begin{bmatrix} \phi & \nu \\ \nu' & -\phi \end{bmatrix} \begin{bmatrix} \phi & \nu \\ \nu' & -\phi \end{bmatrix} = d(\phi)^2 I_2$$

we have that $E_{14} = 0, E_{24} = 0, E_{41} = 0, E_{42} = 0$, and

$$\left\| \begin{bmatrix} \tilde{P}_s & E_{34} \\ E_{43} & E_{44} \end{bmatrix} \right\| \leq d(\phi).$$

Thus (31) agrees with the parametrization of part 2. ∎

Remark 4.6. Note that $d(\phi)$ depends only on the two largest singular values of T, σ_1 and σ_2, and on ϕ (and hence on u and v, the left and right singular vectors corresponding to σ_1). Note also that $d(\phi)$ is a decreasing function of σ_2. Since all optimal interpolating functions have the same largest singular value s_1 (for all frequencies), and furthermore, share the same left and right singular vectors corresponding to s_1, Theorem 4.7 suggests a link between the maximization of $\mu^*(\delta)$ and the minimization of the second largest singular value of the elements of $\mathcal{T}_1$.

The next theorem, which is our main result, shows that $\mu^*(\delta)$ is maximized uniquely by the set of all superoptimal controllers with respect to the first two levels.

THEOREM 4.8. *Let $\mathcal{T}_1 \subseteq \mathcal{H}_\infty^{p\times m}$ be as defined in* (10). *Let $\boldsymbol{x}^T$ and $\boldsymbol{y}$ be the first row and column of $\boldsymbol{X}$ and $\boldsymbol{Y}$, respectively, and define $\mathcal{D}_r(\boldsymbol{G})$ and $\mathcal{E}(\delta, \mu)$ as in* (1) *and* (21), *respectively, for some (fixed) $\delta \in [0, 1]$. Let $\mu^*(\delta)$ be the supremum of μ such that there exists a $\boldsymbol{K}$ for which $(\boldsymbol{G}+\boldsymbol{\Delta}, \boldsymbol{K}) \in \mathcal{S}$ for every $\boldsymbol{\Delta} \in \mathcal{D}_{r_1}(\boldsymbol{G}) \cup \mathcal{E}(\delta, \mu)$. Then the following hold:*

1. *For each δ,*

$$\mu^*(\delta) = \sqrt{\frac{1}{s_1}\left(\frac{\delta}{s_2} + \frac{1-\delta}{s_1}\right)} \geq r_1,$$

   where s_1 and s_2 are the two superoptimal levels of $\mathcal{T}$ with $s_1 = r_1^{-1}$.
2. *For each $0 < \delta \leq 1$ the following two statements are equivalent:*
   (a) $(\boldsymbol{G} + \boldsymbol{\Delta}, \boldsymbol{K}) \in \mathcal{S}$ *for every* $\boldsymbol{\Delta} \in \mathcal{D}_{r_1}(\boldsymbol{G}) \cup \mathcal{E}(\delta, \mu^*(\delta))$,
   (b) $\boldsymbol{K} \in \mathcal{K}_2$.
3. (a) $\mathcal{E}(0, \mu^*(0)) = \mathcal{D}_{r_1}(\boldsymbol{G})$,
   (b) *for each* $\boldsymbol{K} \in \mathcal{K}_2$, $(\boldsymbol{G} + \boldsymbol{\Delta}, \boldsymbol{K}) \in \mathcal{S}$ *for every* $\boldsymbol{\Delta} \in \bigcup_{\delta\in[0,1]} \mathcal{E}(\delta, \mu^*(\delta))$.
4. *Let σ_n and σ_{n-1} denote the two smallest Hankel singular values of $\boldsymbol{G}(-s)$ with $\sigma_{n-1} > \sigma_n$. Then,*

$$\mu^*(\delta) \geq \sqrt{\delta\sigma_n\sigma_{n-1} + (1-\delta)\sigma_n^2}.$$

Proof. Let $\boldsymbol{K} \in \mathcal{K}_2$ and define $\boldsymbol{T} = \boldsymbol{K}(I - \boldsymbol{G}\boldsymbol{K})^{-1} \in \mathcal{H}_\infty$. Fix $\delta \in (0,1]$ and define

$$\mu_1^*(\delta, \boldsymbol{K}) = \sup \{ \mu : (\boldsymbol{G} + \boldsymbol{\Delta}, \boldsymbol{K}) \in \mathcal{S} \text{ for all } \boldsymbol{\Delta} \in \mathcal{D}_{r_1}(\boldsymbol{G}) \cup \mathcal{E}(\delta, \mu)\}$$

and $\mu_1^*(\delta) = \sup \{\mu_1^*(\delta, \boldsymbol{K}) : \boldsymbol{K} \in \mathcal{K}_2\}$. Clearly $\mu_1^*(\delta, \boldsymbol{K}) \le \mu_1^*(\delta) \le \mu^*(\delta)$. We show that

$$r_1 \ge \mu_1^*(\delta, \boldsymbol{K}) \ge \sqrt{\frac{1}{s_1}\left(\frac{\delta}{s_2} + \frac{1-\delta}{s_1}\right)}. \tag{32}$$

Since the largest Hankel singular value of $\boldsymbol{R}(-s)$ is assumed to be simple, we have from [16] that $s_2 \le \sigma_2(\Gamma_{\boldsymbol{R}(-s)}) < \sigma_1(\Gamma_{\boldsymbol{R}(-s)}) = s_1$. Set μ_1 equal to the RHS of (32) and suppose for contradiction that there exists a $\boldsymbol{\Delta} \in \mathcal{D}_{r_1}(\boldsymbol{G}) \cup \mathcal{E}(\delta, \mu_1)$ such that $(\boldsymbol{G} + \boldsymbol{\Delta}, \boldsymbol{K}) \notin \mathcal{S}$. Clearly, if $\boldsymbol{\Delta} \in \mathcal{D}_{r_1}(\boldsymbol{G})$ it cannot be destabilizing and hence $\boldsymbol{\Delta} \in \mathcal{E}(\delta, \mu_1) \backslash \mathcal{D}_{r_1}(\boldsymbol{G})$. Thus $r_1 \le \|\boldsymbol{\Delta}\|_\infty < \mu_1$ and $\|\boldsymbol{x}^T \boldsymbol{\Delta} \boldsymbol{y}\|_\infty \le r_1(1-\delta)$. Since $\boldsymbol{K}$ stabilizes $\boldsymbol{G}$, it follows from the generalized Nyquist theorem that

$$\det(I - \boldsymbol{G}\boldsymbol{K}(j\omega)) \ne 0 \text{ for all } \omega \in \mathcal{R}. \tag{33}$$

Now, let ξ vary continuously in the interval $[0,1]$ and consider the resulting deformation of the Nyquist plot of $\det(I - (\boldsymbol{G} + \xi\boldsymbol{\Delta})\boldsymbol{K}(j\omega))$. Since $\boldsymbol{\Delta}$ is destabilizing, there exist an $\omega_o \in \mathcal{R}$ and a $\xi_o \in (0,1]$ such that

$$\det(I - (\boldsymbol{G}(j\omega_o) + \xi_o \boldsymbol{\Delta}(j\omega_o))\boldsymbol{K}(j\omega_o)) = 0$$

which implies that

$$\det(I - \boldsymbol{G}(j\omega_o)\boldsymbol{K}(j\omega_o)) \det(I - \xi_o \boldsymbol{\Delta}(j\omega_o)\boldsymbol{T}(j\omega_o)) = 0$$

or equivalently that

$$\det(I - \xi_o \boldsymbol{\Delta}(j\omega_o)\boldsymbol{T}(j\omega_o)) = 0 \tag{34}$$

from (33). Now $\|\boldsymbol{x}^T \boldsymbol{\Delta} \boldsymbol{y}\|_\infty \le r_1(1-\delta)$ implies that

$$\xi_o |\boldsymbol{x}^T(j\omega_o)\boldsymbol{\Delta}(j\omega_o)\boldsymbol{y}(j\omega_o)| \le \frac{1-\delta}{s_1} := \phi \tag{35}$$

since $0 < \xi_o \le 1$. Since the two largest singular values of $\boldsymbol{T}(j\omega_o)$ are s_1 and s_2, respectively, Theorem 4.7 guarantees that $\det(I - \xi_o E \boldsymbol{T}(j\omega_o)) \ne 0$ for all E such that

$$\|E\| < \frac{1}{\xi_o}\sqrt{\frac{1}{s_1 s_2} - \phi\left(\frac{1}{s_2} - \frac{1}{s_1}\right)} = \frac{1}{\xi_o}\sqrt{\frac{1}{s_1}\left(\frac{\delta}{s_2} + \frac{1-\delta}{s_2}\right)} = \frac{\mu_1}{\xi_o} \tag{36}$$

provided that

$$\xi_o |\boldsymbol{x}^T(j\omega_o) E \boldsymbol{y}(j\omega_o)| \le \frac{1-\delta}{s_1} = \phi. \tag{37}$$

Thus, Theorem 4.7 also guarantees the nonsingularity of $I - \xi_o \boldsymbol{\Delta}(j\omega_o)\boldsymbol{T}(j\omega_o)$ from (35), (36), and the fact that $\|\boldsymbol{\Delta}(j\omega_o)\| \le \|\boldsymbol{\Delta}\|_\infty < \mu_1 \le \mu_1 \xi_o^{-1}$. This contradicts (34) and hence shows that

$$\mu^*(\delta) \ge \mu_1^*(\delta) \ge \mu_1^*(\delta, \boldsymbol{K}) \ge \sqrt{\frac{1}{s_1}\left(\frac{\delta}{s_2} + \frac{1-\delta}{s_1}\right)}. \tag{38}$$

Next, it is shown that the second and third inequalities in (38) are in fact equalities. To establish this fact, it suffices to construct a $\boldsymbol{\Delta} \in \mathcal{RH}_\infty$ ($\Rightarrow \eta(\boldsymbol{G}) = \eta(\boldsymbol{G}+\boldsymbol{\Delta})$) such that (i) $\|\boldsymbol{\Delta}\|_\infty$ is equal to the RHS of (38), (ii) $\|\boldsymbol{x}^T\boldsymbol{\Delta}\boldsymbol{y}\|_\infty \leq r_1(1-\delta)$, and (iii) $(\boldsymbol{G}+\boldsymbol{\Delta}, \boldsymbol{K})$ is unstable for every $\boldsymbol{K} \in \mathcal{K}_2$.

Using Lemma 4.3 the interpolating function $\boldsymbol{T} \in \mathcal{T}_2$ corresponding to any $\boldsymbol{K} \in \mathcal{K}_2$ can be written in the form

$$\boldsymbol{T} = \boldsymbol{Y}_1 \operatorname{diag}(s_1\boldsymbol{a}, s_2\boldsymbol{b}, \check{\boldsymbol{R}} + \check{\boldsymbol{Q}})\boldsymbol{X}_1,$$

in which $\boldsymbol{Y}_1$ and $\boldsymbol{X}_1$ are square inner matrices, $\boldsymbol{a}$ and $\boldsymbol{b}$ are scalar allpass functions, and $\|\check{\boldsymbol{R}} + \check{\boldsymbol{Q}}\|_\infty \leq s_2$. In addition, also from Lemma 4.3, the first column (row) of $\boldsymbol{Y}_1$ ($\boldsymbol{X}_1$) is identical to the first column (row) of $\boldsymbol{Y}$ ($\boldsymbol{X}$); these are denoted by $\boldsymbol{y}$ and $\boldsymbol{x}^T$, respectively. Define the allpass matrix function $\boldsymbol{Y}_2 = \boldsymbol{Y}_1 \operatorname{diag}(\boldsymbol{a}, \boldsymbol{b}, I_{p-2})$. Then $\boldsymbol{T} = \boldsymbol{Y}_2 \operatorname{diag}(s_1, s_2, \check{\boldsymbol{R}}+\check{\boldsymbol{Q}})\boldsymbol{X}_1$. Factor $\boldsymbol{X}_1^\sim$ and $\boldsymbol{Y}_2^\sim$ as $\boldsymbol{X}_1^\sim = \boldsymbol{N}_1 \operatorname{diag}(\boldsymbol{d}_1, \boldsymbol{d}_2, \ldots, \boldsymbol{d}_m)$ and $\boldsymbol{Y}_2^\sim = \operatorname{diag}(\tilde{\boldsymbol{d}}_1, \tilde{\boldsymbol{d}}_2, \ldots, \tilde{\boldsymbol{d}}_p)\boldsymbol{N}_2$ where the functions $\boldsymbol{N}_1, \boldsymbol{N}_2, \boldsymbol{d}_i^{-1}, \tilde{\boldsymbol{d}}_i^{-1}$ are in $\mathcal{RH}_\infty$, $\boldsymbol{N}_1, \boldsymbol{N}_2$ are square inner, and the $\boldsymbol{d}_i$'s, $\tilde{\boldsymbol{d}}_i$'s are scalar allpass functions; these are left and right coprime factorizations of the columns of $\boldsymbol{X}_1^\sim$ and the rows of $\boldsymbol{Y}_2^\sim$ with inner denominators. Next pick any $\omega_o \in \mathcal{R}$, and write for each $i = 1, 2$, $\boldsymbol{d}_i(j\omega_o) = \exp(j\phi_i)$, $\tilde{\boldsymbol{d}}_i(j\omega_o) = \exp(j\tilde{\phi}_i)$, where $-\pi \leq \phi_i, \tilde{\phi}_i < \pi$. Define two diagonal inner matrices $\boldsymbol{A}_1 = \operatorname{diag}(\boldsymbol{\alpha}_1, \boldsymbol{\alpha}_2)$ and $\boldsymbol{A}_2 = \operatorname{diag}(\tilde{\boldsymbol{\alpha}}_1, \tilde{\boldsymbol{\alpha}}_2)$ as follows: For each $i \in \{1, 2\}$, if $0 < \phi_i < \pi$ ($0 < \tilde{\phi}_i < \pi$), set $\boldsymbol{\alpha}_i(s) = (s-\beta_i)(s+\beta_i)^{-1}$ ($\tilde{\boldsymbol{\alpha}}_i(s) = (s-\tilde{\beta}_i)(s+\tilde{\beta}_i)^{-1}$), where $\arg(j\omega_o - \beta_i)(j\omega_o + \beta_i)^{-1} = \phi_i > 0$ ($\arg(j\omega_o - \tilde{\beta}_i)(j\omega_o + \tilde{\beta}_i)^{-1} = \tilde{\phi}_i > 0$). In the case that $-\pi < \phi_i < 0$, set $\boldsymbol{\alpha}_i(s) = -(s-\beta_i)(s+\beta_i)^{-1}$ ($\tilde{\boldsymbol{\alpha}}_i(s) = -(s-\tilde{\beta}_i)(s+\tilde{\beta}_i)^{-1}$), where $\arg(j\omega_o - \beta_i)(j\omega_o + \beta_i)^{-1} = \pi + \phi_i > 0$ ($\arg(j\omega_o - \tilde{\beta}_i)(j\omega_o + \tilde{\beta}_i)^{-1} = \pi + \tilde{\phi}_i > 0$). Finally, if $\phi_i = 0$ ($\tilde{\phi}_i = 0$) or $\phi_i = -\pi$ ($\tilde{\phi}_i = -\pi$), set $\boldsymbol{\alpha}_i$ ($\tilde{\boldsymbol{\alpha}}_i$) to 1 or -1, respectively.

Next, let $\boldsymbol{N}_{11}$ ($\boldsymbol{N}_{21}$) denote the matrix consisting of the first two columns (rows) of $\boldsymbol{N}_1$ ($\boldsymbol{N}_2$), and define $\boldsymbol{\Delta} \in \mathcal{RH}_\infty$ as

$$\boldsymbol{\Delta} = \boldsymbol{N}_{11}\boldsymbol{A}_1 \begin{bmatrix} \phi & \nu_o \\ \nu_o & -\phi \end{bmatrix} \boldsymbol{A}_2\boldsymbol{N}_{21},$$

where ϕ is defined in (35) and

$$\nu_o = \sqrt{\left(\frac{1}{s_2} + \phi\right)\left(\frac{1}{s_1} - \phi\right)} = \sqrt{\frac{\delta}{s_1}\left(\frac{1}{s_2} + \frac{1-\delta}{s_1}\right)}, \tag{39}$$

where the second equality in (39) follows by using the definition of ϕ in (35). Since $\boldsymbol{N}_{11}$, $\boldsymbol{N}_{21}^T$, $\boldsymbol{A}_1$, and $\boldsymbol{A}_2$ are inner matrices,

$$\|\boldsymbol{\Delta}\|_\infty = \left\| \begin{bmatrix} \phi & \nu_o \\ \nu_o & -\phi \end{bmatrix} \right\| = \sqrt{\phi^2 + \nu_o^2},$$

which is equal to the RHS of (38) after some simple algebra.

Since, $\boldsymbol{X}_1\boldsymbol{X}_1^\sim = I_m$ and $\boldsymbol{X}$ and $\boldsymbol{X}_1$ have the same first row ($\boldsymbol{x}^T$), we have

$$\boldsymbol{x}^T\boldsymbol{X}_1^\sim = [1\ 0\ \ldots\ 0] \ \Rightarrow\ \boldsymbol{x}^T\boldsymbol{N}_1 \operatorname{diag}(\boldsymbol{d}_1, \boldsymbol{d}_2, \ldots, \boldsymbol{d}_m) = [1\ 0\ \ldots\ 0],$$

and hence

$$\boldsymbol{x}^T\boldsymbol{N}_{11} = [\ \boldsymbol{d}_1^{-1}\ \ 0\ ]. \tag{40}$$

Similarly, since $\boldsymbol{Y}_1^{\sim}\boldsymbol{Y}_1 = I_p$ and matrices $\boldsymbol{Y}$ and $\boldsymbol{Y}_1$ share their first column ($\boldsymbol{y}$),

$$\boldsymbol{Y}_1^{\sim}\boldsymbol{y} = \begin{bmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix} \Rightarrow \begin{bmatrix} \boldsymbol{a} & 0 & 0 \\ 0 & \boldsymbol{b} & 0 \\ 0 & 0 & I_{p-2} \end{bmatrix} \boldsymbol{Y}_2^{\sim}\boldsymbol{y} = \begin{bmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix}$$

or, equivalently, that

$$\boldsymbol{N}_2\boldsymbol{y} = \begin{bmatrix} \tilde{\boldsymbol{d}}_1^{-1}\boldsymbol{a}^{-1} \\ 0 \\ \vdots \\ 0 \end{bmatrix} \Rightarrow \boldsymbol{N}_{21}\boldsymbol{y} = \begin{bmatrix} \tilde{\boldsymbol{d}}_1^{-1}\boldsymbol{a}^{-1} \\ 0 \end{bmatrix}. \tag{41}$$

Using (40) and (41) we conclude that

$$\|\boldsymbol{x}^T\boldsymbol{\Delta}\boldsymbol{y}\|_\infty = \left\| \boldsymbol{x}^T\boldsymbol{N}_{11}\boldsymbol{A}_1 \begin{bmatrix} \phi & \nu_o \\ \nu_o & -\phi \end{bmatrix} \boldsymbol{A}_2\boldsymbol{N}_{21}\boldsymbol{y} \right\|_\infty = \|\phi\boldsymbol{a}^{-1}\boldsymbol{d}_1^{-1}\tilde{\boldsymbol{d}}_1^{-1}\boldsymbol{\alpha}_1\tilde{\boldsymbol{\alpha}}_1\|_\infty = \phi,$$

using (37) and the fact that $\boldsymbol{a}, \boldsymbol{d}_1, \boldsymbol{d}_3, \boldsymbol{\alpha}_1$, and $\tilde{\boldsymbol{\alpha}}_1$ are all scalar allpass.

Using the identity $\det(I - AB) = \det(I - BA)$, $\det(I_m - \boldsymbol{\Delta}(j\omega_o)\boldsymbol{T}(j\omega_o))$ can be written as

$$\det\left\{ I - \boldsymbol{X}_1(j\omega_o)\boldsymbol{N}_{11}(j\omega_o)\boldsymbol{A}_1(j\omega_o) \begin{bmatrix} \phi & \nu_o \\ \nu_o & -\phi \end{bmatrix} \boldsymbol{A}_2(j\omega_o)\boldsymbol{N}_{21}(j\omega_o)\boldsymbol{Y}_2(j\omega_o)\Pi \right\},$$

where we have defined $\Pi = \operatorname{diag}(s_1, s_2, \check{\boldsymbol{R}}(j\omega_o) + \check{\boldsymbol{Q}}(j\omega_o))$. It is now easy to verify from the construction of $\boldsymbol{A}_1$ and $\boldsymbol{A}_2$ above that

$$\boldsymbol{X}_1(j\omega_o)\boldsymbol{N}_{11}(j\omega_o)\boldsymbol{A}_1(j\omega_o) = \begin{bmatrix} I_2 \\ 0_{m-2,2} \end{bmatrix}, \ \boldsymbol{A}_2(j\omega_o)\boldsymbol{N}_{21}(j\omega_o)\boldsymbol{Y}_2(j\omega_o) = [I_2 \ 0_{2,p-2}],$$

and hence

$$\det(I_m - \boldsymbol{\Delta}(j\omega_o)\boldsymbol{T}(j\omega_o)) = \det \begin{bmatrix} 1 - \phi s_1 & -\nu_o s_2 & 0 \\ -\nu_o s_1 & 1 + \phi s_2 & 0 \\ 0 & 0 & I_{m-2} \end{bmatrix}$$

or

$$\det(I_m - \boldsymbol{\Delta}(j\omega_o)\boldsymbol{T}(j\omega_o)) = (1 - \phi s_1)(1 + \phi s_2) - \nu_o^2 s_1 s_2 = 0,$$

after some simple algebra using (39). This implies that

$$\det(I_m - (\boldsymbol{G}(j\omega_o) + \boldsymbol{\Delta}(j\omega_o))\boldsymbol{K}(j\omega_o)) = 0,$$

and hence $\boldsymbol{\Delta}$ is destabilizing from the generalized Nyquist theorem [25]. This shows that the third inequality in (38) is indeed an equality as claimed. Since $\boldsymbol{\Delta}$ is destabilizing for every $\boldsymbol{K} \in \mathcal{K}_2$, the second equality in (38) also follows.

To establish that the first inequality in (38) is an equality it suffices to construct for each $\boldsymbol{K} \in \mathcal{K}_1 \backslash \mathcal{K}_2$ a $\boldsymbol{\Delta} \in \mathcal{RH}_\infty^{m\times p}$ such that (i) $\|\boldsymbol{\Delta}\|_\infty$ is (strictly) less than the RHS of (38), (ii) $\|\boldsymbol{x}^T\boldsymbol{\Delta}\boldsymbol{y}\|_\infty \le r_1(1-\delta)$, and (iii) $(\boldsymbol{G} + \boldsymbol{\Delta}, \boldsymbol{K}) \notin \mathcal{S}$. Take any $\boldsymbol{K} \in \mathcal{K}_1$, $\boldsymbol{K} \notin \mathcal{K}_2$ and let $\boldsymbol{T} = \boldsymbol{K}(I - \boldsymbol{G}\boldsymbol{K})^{-1}$. From Theorem 4.2, $\boldsymbol{T} \in \mathcal{T}_1$ has the form

$\boldsymbol{T} = \boldsymbol{Y}\ \mathrm{diag}(s_1\boldsymbol{a}, \hat{\boldsymbol{R}} + \bar{\boldsymbol{Q}})\boldsymbol{X}$, where $\|\hat{\boldsymbol{R}} + \bar{\boldsymbol{Q}}\|_\infty \leq s_1$. Since $\boldsymbol{T} \notin \mathcal{T}_2$, there exists an $\omega_o \in \mathcal{R}_+$ such that

$$s_2 < \overline{\sigma}\left(\hat{\boldsymbol{R}}(j\omega_o) + \bar{\boldsymbol{Q}}(j\omega_o)\right) \leq s_1.$$

Let $\hat{\boldsymbol{R}}(j\omega_o) + \bar{\boldsymbol{Q}}(j\omega_o)$ have a singular value decomposition

$$\hat{\boldsymbol{R}}(j\omega_o) + \bar{\boldsymbol{Q}}(j\omega_o) = U\ \mathrm{diag}(\Sigma, 0_{p-t-1,m-t-1})V',$$

where $\Sigma = \mathrm{diag}(\sigma_1, \sigma_2, \ldots, \sigma_t)$ with $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_t > 0$. Then, $s_2 < \sigma_1 \leq s_1$. Denote by $u = [u_1\ u_2\ \ldots\ u_{p-1}]^T$ and $v = [v_1\ v_2\ \ldots\ v_{m-1}]^T$ the first column of U and V, respectively. Define an inner vector $\boldsymbol{\psi} = [\boldsymbol{\psi}_1\ \boldsymbol{\psi}_2\ \ldots\ \boldsymbol{\psi}_{p-1}]^T \in \mathcal{RH}_\infty^{(p-1)\times 1}$ as follows: Write $u_i = \rho_i \exp(j\theta_i)$ for each $i = 1, 2, \ldots, p$, where the ρ_i's are real and $0 \leq \theta_i < \pi$; if $\theta_i \neq 0$, set $\boldsymbol{\psi}_i(s) = \rho_i(s - \beta_i)(s + \beta_i)^{-1}$, where $\beta_i > 0$ and is such that $\arg((j\omega_o - \beta_i)(j\omega_o + \beta_i)^{-1}) = \theta_i$; if $\theta_i = 0$, set $\boldsymbol{\psi}_i = \rho_i$. Clearly, $\boldsymbol{\psi} \in \mathcal{RH}_\infty^{(p-1)\times 1}$, $\boldsymbol{\psi}^\sim\boldsymbol{\psi} = 1$, and $\boldsymbol{\psi}(j\omega_o) = u$. In a similar way, construct an $\mathcal{RH}_\infty^{(m-1)\times 1}$ inner vector $\boldsymbol{\xi}$ which "interpolates" v at $s = j\omega_o$, i.e., $\boldsymbol{\xi}(j\omega_o) = v$.

Define $\hat{\boldsymbol{X}}_1 = \mathrm{diag}(1, \boldsymbol{\xi}^\sim) \in \mathcal{RL}_\infty^{2\times m}$ and $\hat{\boldsymbol{Y}}_1 = \boldsymbol{Y}\ \mathrm{diag}(\boldsymbol{a}, \boldsymbol{\psi}) \in \mathcal{RL}_\infty^{p\times 2}$. Clearly, $\hat{\boldsymbol{X}}_1\hat{\boldsymbol{X}}_1^\sim = \hat{\boldsymbol{Y}}_1^\sim\hat{\boldsymbol{Y}}_1 = I_2$. Define factorizations of the columns (rows) of $\hat{\boldsymbol{X}}_1^\sim$ ($\hat{\boldsymbol{Y}}_1^\sim$) of the form $\hat{\boldsymbol{X}}_1^\sim = \hat{\boldsymbol{N}}_1\mathrm{diag}(\boldsymbol{d}_1, \boldsymbol{d}_2)$ and $\hat{\boldsymbol{Y}}_1^\sim = \mathrm{diag}(\tilde{\boldsymbol{d}}_1, \tilde{\boldsymbol{d}}_2)\hat{\boldsymbol{N}}_2$ such that $\hat{\boldsymbol{N}}_1$ and $\hat{\boldsymbol{N}}_2^T$ are inner and $\boldsymbol{d}_1, \boldsymbol{d}_2, \tilde{\boldsymbol{d}}_1, \tilde{\boldsymbol{d}}_2$ are scalar allpass. Similar to a previous part of the proof, define 2×2 inner matrices $\boldsymbol{A}_1 = \mathrm{diag}(\boldsymbol{\alpha}_1, \boldsymbol{\alpha}_2)$ and $\boldsymbol{A}_2 = \mathrm{diag}(\tilde{\boldsymbol{\alpha}}_1, \tilde{\boldsymbol{\alpha}}_2)$ such that $\boldsymbol{d}_1(j\omega_o)\boldsymbol{\alpha}_1(j\omega_o) = 1$, $\boldsymbol{d}_2(j\omega_o)\boldsymbol{\alpha}_2(j\omega_o) = 1$, $\tilde{\boldsymbol{d}}_1(j\omega_o)\tilde{\boldsymbol{\alpha}}_1(j\omega_o) = 1$, and $\tilde{\boldsymbol{d}}_2(j\omega_o)\tilde{\boldsymbol{\alpha}}_2(j\omega_o) = 1$.

Define $\boldsymbol{\Delta} \in \mathcal{RH}_\infty^{m\times p}$ as

$$\boldsymbol{\Delta} = \hat{\boldsymbol{N}}_1\boldsymbol{A}_1\begin{bmatrix} \phi & \nu_1 \\ \nu_1 & -\phi \end{bmatrix}\boldsymbol{A}_2\hat{\boldsymbol{N}}_2, \qquad \nu_1 = \sqrt{\left(\frac{1}{\sigma_1} + \phi\right)\left(\frac{1}{s_1} - \phi\right)}.$$

The inner character of $\hat{\boldsymbol{N}}_1, \hat{\boldsymbol{N}}_2^T, \boldsymbol{A}_1$, and $\boldsymbol{A}_2$ implies that $\|\boldsymbol{\Delta}\|_\infty = \sqrt{\phi^2 + \nu_1^2} < \sqrt{\phi^2 + \nu_o^2}$ since $s_2 < \sigma_1$. Thus, $\|\boldsymbol{\Delta}\|_\infty$ is strictly less than the RHS of (38). Moreover, it can be easily verified that $\|\boldsymbol{x}^T\boldsymbol{\Delta}\boldsymbol{y}\|_\infty = \phi$. Finally, using the identity $\det(I - AB) = \det(I - BA)$ we can write

$$\det(I_m - \boldsymbol{\Delta}(j\omega_o)\boldsymbol{T}(j\omega_o)) = \det\left(I_m - Z_1\begin{bmatrix} \phi & \nu_1 \\ \nu_1 & -\phi \end{bmatrix}Z_2\begin{bmatrix} s_1 & 0 & 0 \\ 0 & \sigma_1 & 0 \\ 0 & 0 & * \end{bmatrix}\right),$$

where

$$\begin{aligned} Z_1 &= \mathrm{diag}(1, V')\boldsymbol{X}(j\omega_o)\boldsymbol{N}_1(j\omega_o)\boldsymbol{A}_1(j\omega_o), \\ Z_2 &= \boldsymbol{A}_2(j\omega_o)\boldsymbol{N}_2(j\omega_o)\boldsymbol{Y}(j\omega_o)\ \mathrm{diag}(\boldsymbol{a}(j\omega_o), U), \end{aligned}$$

and * denotes a matrix not relevant for our present purposes. It can be easily verified from the above construction that $Z_1^T = [I_2\ \ 0_{2,m-1}]$ and $Z_2 = [I_2\ \ 0_{2,p-1}]$, and hence

$$\det(I_m - \boldsymbol{\Delta}(j\omega_o)\boldsymbol{T}(j\omega_o)) = \det\begin{bmatrix} 1 - \phi s_1 & -\nu_1\sigma_1 & 0 \\ -\nu_1 s_1 & 1 + \phi\sigma_1 & 0 \\ 0 & 0 & I_{m-2} \end{bmatrix} = 0,$$

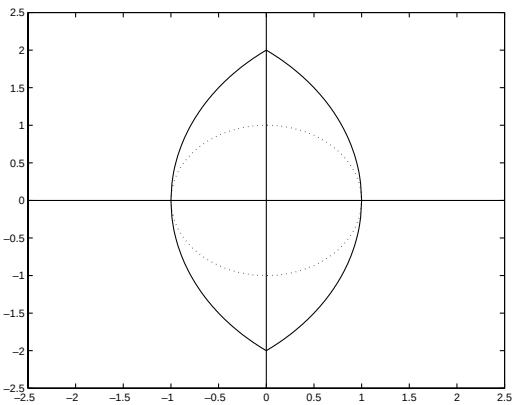


FIG. 2. *Extended permissible uncertainty set.*

which implies that $(\boldsymbol{G}+\boldsymbol{\Delta},\boldsymbol{K})$ is unstable from the generalized Nyquist theorem. Since such destabilizing $\boldsymbol{\Delta}$'s with $\|\boldsymbol{\Delta}\|_\infty < \mu_1^*(\delta)$ and $\|\boldsymbol{x}^T\boldsymbol{\Delta}\boldsymbol{y}\|_\infty = r_1(1-\delta)$ can be constructed for any $\boldsymbol{K}\in\mathcal{K}_1\backslash\mathcal{K}_2$, we conclude that $\mu^*(\delta)=\mu_1^*(\delta)$ and part 1 is proved. It is also clear that $\mathcal{K}_2$ is the set of all controllers $\boldsymbol{K}$ such that $(\boldsymbol{G}+\boldsymbol{\Delta},\boldsymbol{K})\in\mathcal{S}$ for every $\boldsymbol{\Delta}\in\mathcal{D}_{r_1}(\boldsymbol{G})\cup\mathcal{E}(\delta,\mu^*(\delta))$ and part 2 follows.

To prove part 3, note that setting $\delta=0$ in (21) gives

$$\mathcal{E}(0,\mu)=\{\boldsymbol{\Delta}\in\mathcal{L}_\infty : \|\boldsymbol{\Delta}\|_\infty<\mu,\ \|\boldsymbol{x}^T\boldsymbol{\Delta}\boldsymbol{y}\|_\infty\le r_1, \eta(\boldsymbol{G})=\eta(\boldsymbol{G}+\boldsymbol{\Delta})\}.$$

Since from Lemma 4.4 there exist uniformly destabilizing perturbations of $\boldsymbol{G}$ in $\partial\mathcal{D}(\boldsymbol{G})$, we have that $\mu^*(0)\le r_1$. Now,

$$\mathcal{E}(0,r_1)=\{\boldsymbol{\Delta}\in\mathcal{L}_\infty\ :\ \|\boldsymbol{\Delta}\|_\infty<r_1,\ \|\boldsymbol{x}^T\boldsymbol{\Delta}\boldsymbol{y}\|_\infty\le r_1, \eta(\boldsymbol{G})=\eta(\boldsymbol{G}+\boldsymbol{\Delta})\}. \tag{42}$$

In view of the condition $\|\boldsymbol{\Delta}\|_\infty < r_1$ in (42) and the fact that $\|\boldsymbol{x}\|_\infty=\|\boldsymbol{y}\|_\infty=1$, condition $\|\boldsymbol{x}^T\boldsymbol{\Delta}\boldsymbol{y}\|_\infty\le r_1$ in the characterization of $\mathcal{E}(0,r_1)$ in (42) is superfluous and thus $\mathcal{E}(0,r_1)=\mathcal{D}_{r_1}(\boldsymbol{G})$. Hence, $\mu^*(0)=r_1$ and the set of all $\boldsymbol{K}$ such that $(\boldsymbol{G}+\boldsymbol{\Delta},\boldsymbol{K})\in\mathcal{S}$ for every $\boldsymbol{\Delta}\in\mathcal{D}_{r_1}(\boldsymbol{G})\cup\mathcal{E}(0,\mu^*(0))=\mathcal{D}_{r_1}(\boldsymbol{G})$ is clearly $\mathcal{K}_1$. Since $\mathcal{K}_2\subseteq\mathcal{K}_1$, part 3 follows immediately from part 2. Finally, part 4 follows from the relations $s_1=\sigma_1(\Gamma_{\boldsymbol{R}(-s)})=\sigma_n^{-1}$, $\sigma_2(\Gamma_{\boldsymbol{R}(-s)})=\sigma_{n-1}^{-1}$ (see Theorem 3.3 and the subsequent remark), and the inequality $s_2\le\sigma_2(\Gamma_{\boldsymbol{R}(-s)})$ (see [16]). □

Remark 4.7. Figure 2 is an illustration of the set $\bigcup_{\delta\in[0,1]}\mathcal{E}(\delta,\mu^*(\delta))$ in the two-dimensional case. Here, $s_1=1$ and $s_2=0.25$. The "worst direction" is assumed to be the horizontal axis. The (open) disc of radius one represents the set of uncertainties guaranteed to be stabilized by optimal controllers. The area bounded by the solid curve represents the set of uncertainties guaranteed to be stabilized by (second-level) superoptimal controllers. Note the increase in the stability radius in all directions other than the worst direction.

5. An upper bound on the structured-singular value. So far, our results have been restricted to the unstructured uncertainty case. Our overall aim has been to use the degrees of freedom in the set of all optimal ("maximally robust") controllers $\mathcal{K}_1$ in order to extend as far as possible the region of the uncertainty space guaranteed to be stabilized by a subset of $\mathcal{K}_1$. The optimal subset of $\mathcal{K}_1$ has been identified as the set of superoptimal controllers with respect to the first two levels, $\mathcal{K}_2$.

A different interpretation of our method allows us to apply our results to structured uncertainty models as well. The crucial point is that the region of the uncertainty space which is nondestabilizing can be extended beyond $\mathcal{D}_{r_1}$ only *by imposing a structure* on the admissible set of uncertainties; this structure, in our case, is of the form of a projection (uniform in frequency) in the "worst-case direction" $\langle \boldsymbol{y}\boldsymbol{x}^T, . \rangle$, along which all uniformly destabilizing perturbations in $\partial\mathcal{D}_{r_1}(\boldsymbol{G})$ have been shown to lie. Suppose that the uncertainty is known to have a (block) diagonal structure, $\boldsymbol{\Delta}$, say. The following general procedure can be used in principle to obtain a lower bound on $r_{\boldsymbol{\Delta}}$, the robust stability radius with respect to structure $\boldsymbol{\Delta}$.

- Maximize the robust stability radius for a class of unstructured perturbations; let the maximum (unstructured) robust stability radius be r_1 and the corresponding worst-case direction be $\langle \boldsymbol{y}\boldsymbol{x}^T, . \rangle$.
- Given a specific uncertainty structure $\boldsymbol{\Delta}$, find the largest $\delta^* \in (0, 1]$ compatible with $\boldsymbol{\Delta}$, i.e., the maximum $\delta^* \in (0, 1]$ such that $|\langle \boldsymbol{y}\boldsymbol{x}^T, \Delta \rangle| \leq r_1(1 - \delta^*)$, uniformly in ω *for every* $\Delta \in \boldsymbol{\Delta}$.
- Then $\mu^*(\delta^*)$ is a guaranteed lower bound of the robust stability radius of the system with respect to uncertainty structure $\boldsymbol{\Delta}$.

This general method for calculating bounds for the structured robust-stability radius, $r_{\boldsymbol{\Delta}}$ (equivalently the structured singular value $\mu_{\boldsymbol{\Delta}}$), will be developed in future work. In this section we present preliminary results for the constant μ problem and a simple example illustrating our method. A more complete development is given in [9].

We use the definitions and notation of [19]. Let $T \in \mathcal{C}^{n\times n}$ have a singular value decomposition

$$T = U\Sigma V', \quad \Sigma = \operatorname{diag}(\sigma_1, \sigma_2, \ldots, \sigma_n), \quad U, V \in \mathcal{C}^{n\times n}, \quad U'U = V'V = I_n \tag{43}$$

and assume that

$$\sigma_1 > \sigma_2 \geq \sigma_3 \geq \cdots \geq \sigma_n > 0. \tag{44}$$

Define the structured uncertainty set

$$\boldsymbol{\Delta} = \{\operatorname{diag}(\delta_1 I_{r_1}, \ldots, \delta_S I_{r_S}, \Delta_1, \ldots, \Delta_F) : \delta_1, \ldots, \delta_S \in \mathcal{C}, \Delta_j \in \mathcal{C}^{m_j \times m_j}, j = 1, \ldots, F\}$$

with $\sum_{i=1}^{S} r_i + \sum_{j=1}^{F} m_j = n$ and let $\mathbf{B}_{\boldsymbol{\Delta}} = \{\Delta \in \boldsymbol{\Delta} : \|\Delta\| \leq 1\}$. Then the structured singular value of T is defined as

$$\mu_{\boldsymbol{\Delta}}(T)^{-1} = \min_{\substack{\Delta\in\boldsymbol{\Delta} \\ \det(I-\Delta T)=0}} \|\Delta\| = \min_{\substack{\Delta\in\boldsymbol{\Delta} \\ \det(\Sigma^{-1}-V'\Delta U)=0}} \|\Delta\|$$

(if there exists no $\Delta \in \boldsymbol{\Delta}$ such that $\det(I - \Delta T) = 0$, we define $\mu_{\boldsymbol{\Delta}}(T) = 0$). Let $u, v \in \mathcal{C}^{n\times 1}$ be the first columns of U and V, respectively. Partition u and v compatibly with $\boldsymbol{\Delta}$ as follows:

$$u = \begin{bmatrix} u_1 \\ \vdots \\ u_S \\ u_{S+1} \\ \vdots \\ u_{S+F} \end{bmatrix}, \quad v = \begin{bmatrix} v_1 \\ \vdots \\ v_S \\ v_{S+1} \\ \vdots \\ v_{S+F} \end{bmatrix}, \quad u_i, v_i \in \mathcal{C}^{r_i}, \quad u_{S+j}, v_{S+j} \in \mathcal{C}^{m_j} \tag{45}$$

for $i = 1, \ldots, S$ and $j = 1, \ldots, F$. Then it is straightforward to verify that

$$\alpha_0 := \max_{\Delta \in \mathbf{B}_{\mathbf{\Delta}}} |v'\Delta u| = \sum_{i=1}^{S} |v_i' u_i| + \sum_{j=1}^{F} \|v_{S+j}\| \|u_{S+j}\| \leq 1. \tag{46}$$

Define the set $\mathbf{\Delta}_{\alpha_0} := \{\Delta \in \mathcal{C}^{n\times n} : |v'\Delta u| \leq \alpha_0 \|\Delta\|\}$. Clearly $\mathbf{\Delta} \subseteq \mathbf{\Delta}_{\alpha_0}$. It follows that

$$\mu_{\mathbf{\Delta}}(T)^{-1} = \min_{\substack{\Delta \in \mathbf{\Delta} \\ \det(\Sigma^{-1} - V'\Delta U) = 0}} \|\Delta\| \geq \min_{\substack{\Delta \in \mathbf{\Delta}_{\alpha_0} \\ \det(\Sigma^{-1} - V'\Delta U) = 0}} \|\Delta\| =: \bar{\mu}_{\mathbf{\Delta}}(T)^{-1}.$$

Thus $\bar{\mu}_{\mathbf{\Delta}}(T)$ is an upper bound on $\mu_{\mathbf{\Delta}}(T)$. The evaluation of $\bar{\mu}_{\mathbf{\Delta}}(T)$ is related to the results of [15] (see Theorem 4.7). The next result uses Theorem 4.7 to give an expression for $\bar{\mu}_{\mathbf{\Delta}}(T)$ that involves only σ_1, σ_2, u, v, and the uncertainty set $\mathbf{\Delta}$ and shows that $\bar{\mu}_{\mathbf{\Delta}}(T)$ is increasing in σ_2.

THEOREM 5.1. *Let $T \in \mathcal{C}^{n\times n}$ have a singular value decomposition as in* (43) *and assume that* (44) *is satisfied. Let u and v and the first columns of U and V, respectively, be partitioned as in* (45) *and define α_0 as in* (46).

1. *Let*

$$d = (\sigma_1 - \sigma_2)\alpha_0/2 + \sqrt{[(\sigma_1 - \sigma_2)\alpha_0/2]^2 + \sigma_1\sigma_2}. \tag{47}$$

Then

$$\bar{\mu}_{\mathbf{\Delta}}(T) := \left(\min_{\substack{\det(\Sigma^{-1} - V'\Delta U) = 0 \\ |v'\Delta u| \leq \alpha_0 \|\Delta\|}} \|\Delta\| \right)^{-1} = d. \tag{48}$$

2. *For all $\alpha_0 \in [0, 1]$ we have*

$$\mu_{\mathbf{\Delta}}(T) \leq \bar{\mu}_{\mathbf{\Delta}}(T) \leq \sigma_1. \tag{49}$$

3. *If $\alpha_0 = 1$, then*

$$\mu_{\mathbf{\Delta}}(T) = \bar{\mu}_{\mathbf{\Delta}}(T) = \sigma_1. \tag{50}$$

4. *If $\alpha_0 < 1$, then*

$$\mu_{\mathbf{\Delta}}(T) \leq \bar{\mu}_{\mathbf{\Delta}}(T) < \sigma_1 \tag{51}$$

with $\mu_{\mathbf{\Delta}}(T) = \bar{\mu}_{\mathbf{\Delta}}(T)$ if and only if there exists $\Delta \in \mathbf{\Delta}$ such that

$$V'\Delta U = d^{-1} \begin{bmatrix} \alpha_0 & e^{j\theta}\sqrt{1-\alpha_0^2} & 0 \\ e^{-j\theta}\sqrt{1-\alpha_0^2} & -\alpha_0 & 0 \\ 0 & 0 & \Delta_{22} \end{bmatrix} \tag{52}$$

for arbitrary θ and any $\Delta_{22} \in \mathcal{C}^{(n-2)\times(n-2)}$ satisfying $\|\Delta_{22}\| \leq 1$.

Proof.

1. We first show that

$$\bar{\mu}_{\mathbf{\Delta}}(T)^{-1} \leq d^{-1}. \tag{53}$$

Let

$$\Delta = d^{-1}V \begin{bmatrix} \alpha_0 & \sqrt{1-\alpha_0^2} & 0 \\ \sqrt{1-\alpha_0^2} & -\alpha_0 & 0 \\ 0 & 0 & 0_{(n-2)\times(n-2)} \end{bmatrix} U'.$$

Then it is easy to verify that $\|\Delta\| = d^{-1},\ |v'\Delta u| = \alpha_0 d^{-1} \le \alpha_0\|\Delta\|$, and

$$\det(\Sigma^{-1} - V'\Delta U) = \sigma_1^{-1}\sigma_2^{-1} - \alpha_0 d^{-1}(\sigma_2^{-1} - \sigma_1^{-1}) - \alpha_0^2 d^{-2} = 0$$

after some manipulation, and this proves (53). Thus we can restrict our search in (48) to the set

$$\{\Delta \in \mathcal{C}^{n\times n} : \|\Delta\| \le d^{-1},\ |v'\Delta u| \le \alpha_0 d^{-1}\}$$

and so

$$\begin{aligned} \bar{\mu}_{\mathbf{\Delta}}(T)^{-1} &= \min_{\substack{\det(\Sigma^{-1}-V'\Delta U)=0 \\ |v'\Delta u|\le\alpha_0 d^{-1} \\ \|\Delta\|\le d^{-1}}} \|\Delta\| \\ &= \sqrt{\sigma_1^{-1}\sigma_2^{-1} - \alpha_0 d^{-1}(\sigma_2^{-1} - \sigma_1^{-1})} = d^{-1}, \end{aligned} \tag{54}$$

where the first equality in (54) follows from Theorem 4.7 and the second equality follows from a straightforward calculation using the definition of d (in fact, d is defined so that d^{-1} is the positive solution of (54)).

2. It is straightforward to verify that $d \le \sigma_1$. The first inequality follows from the definitions of $\bar{\mu}_{\mathbf{\Delta}}(T)$ and $\mu_{\mathbf{\Delta}}(T)$.
3. Suppose that $\alpha_0 = 1$. Then a simple calculation verifies that $d = \sigma_1$ which proves the second equality in (50). To prove the first equality, define

$$\Delta = \sigma_1^{-1}\mathrm{diag}(\delta_1 I_{r_1}, \ldots, \delta_S I_{r_S}, \Delta_1, \ldots, \Delta_F) \in \mathbf{\Delta},$$

where

$$\begin{aligned} \delta_i &= \frac{|v_i'u_i|}{v_i'u_i} \in \mathcal{C} \quad (\delta_i = 0 \text{ if } v_i'u_i = 0), \quad i = 1, \ldots, S, \\ \Delta_j &= \frac{v_j u_j'}{\|v_j\|\|u_j\|} \in \mathcal{C}^{m_j\times m_j} \quad (\Delta_j = 0 \text{ if } \|v_j\|\|u_j\| = 0), \quad j = 1, \ldots, F, \end{aligned}$$

where u_i, v_i are defined in (45). Then $\|\Delta\| = \sigma_1^{-1}$ and $v'\Delta u = \sigma_1^{-1}$. This implies that $V'\Delta U = \mathrm{diag}(\sigma_1^{-1}, \Delta_{22})$ for some Δ_{22} with $\|\Delta_{22}\| \le \sigma_1^{-1}$. It is easy to verify that $\det(\Sigma^{-1} - V'\Delta U) = 0$ and so $\mu_{\mathbf{\Delta}}(T) = \sigma_1$ and the first equality in (50) is proved.
4. Suppose that $\alpha_0 < 1$. Then a simple verification shows that $d < \sigma_1$ and establishes the second inequality in (51). Part 2 of Theorem 4.7 gives all Δ such that $\|\Delta\| = d^{-1},\ \det(\Sigma^{-1} - V'\Delta U) = 0$, and $|v'\Delta u| \le \alpha_0 d^{-1}$ as in (52). Hence $\mu_{\mathbf{\Delta}}(T) = \bar{\mu}_{\mathbf{\Delta}}(T)$ if and only if there exists such a $\Delta \in \mathbf{\Delta}$.

This completes the proof. □

Remark 5.1. Note that $\bar{\mu}_{\mathbf{\Delta}}(T)$ depends only on σ_1, σ_2, and α_0 (see (47) and (48)), and α_0 in turn depends only on u, v, and the structured uncertainty set $\mathbf{\Delta}$ (see (46)). In the context of the robust stabilization of systems with unstructured additive

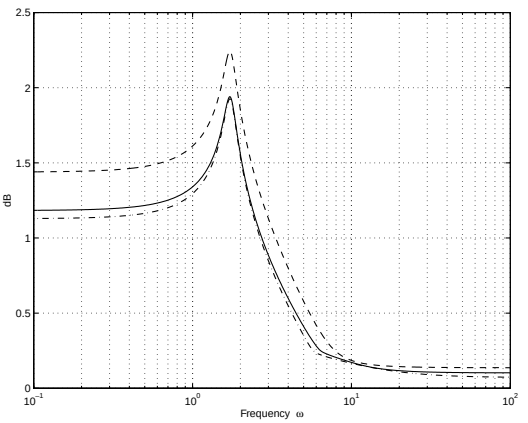


FIG. 3. $\sigma_1(\boldsymbol{T})$ *(dashed), D-iteration upper bound (dash-dot), and upper bound* $\bar{\mu}_{\boldsymbol{\Delta}}(\boldsymbol{T})$ *(solid).*

perturbations, all optimal interpolating functions share the same largest singular value s_1 at all frequencies. They also share the same (frequency-dependent) singular vectors corresponding to s_1 (the inner vectors $\boldsymbol{x}$ and $\boldsymbol{y}$). Thus the only free parameter that can be used to minimize $\bar{\mu}_{\boldsymbol{\Delta}}(\cdot)$ within the set $\mathcal{T}_1$ is the second largest singular value. Noting that $\bar{\mu}_{\boldsymbol{\Delta}}(\cdot)$ is a nonincreasing function of σ_2 (see (47)) suggests that within $\mathcal{T}_1$, $\bar{\mu}_{\boldsymbol{\Delta}}(\cdot)$ is minimized by $\mathcal{T}_2$. This will be elaborated in a future work.

Remark 5.2. The bound $\bar{\mu}_{\boldsymbol{\Delta}}$, although in general tighter than σ_1, is less tight than the upper bound of the D-iteration [19]. In fact, it is shown in [19] that at the end of the D-iteration, either

1. $\sigma_1(T) = \sigma_2(T)$, in which case our results are not applicable (see (44)), or
2. $\sigma_1(T) > \sigma_2(T)$, in which case $\mu_{\boldsymbol{\Delta}}(T) = \sigma_1(T)$. It can be shown that this corresponds to $\alpha_0 = 1$.

The main purpose in this work is to illustrate the improved robustness properties of superoptimal controllers, rather than attempting to improve the D-iteration bound.

Example 5.1. This example illustrates the upper bound $\bar{\mu}_{\boldsymbol{\Delta}}(\boldsymbol{T})$, where

$$\boldsymbol{T} \stackrel{s}{=} \left[\begin{array}{cc} A & B \\ C & D \end{array}\right] = \left[\begin{array}{ccc|cccc} -5.91 & -11.49 & 6.03 & -0.59 & -1.90 & 0.19 & 1.35 \\ -1.85 & -5.62 & 1.63 & -0.31 & 1.11 & 0.63 & 0.12 \\ -7.71 & -17.40 & 7.07 & 0.97 & 0.72 & -0.35 & -0.58 \\ \hline -0.37 & 0.49 & 0.52 & 0.01 & 0.22 & 0.71 & 0.97 \\ 1.43 & -0.09 & 1.36 & 0.60 & 0.70 & 0.23 & 0.36 \\ 0.07 & 0.37 & -0.41 & 0.82 & 0.52 & 0.45 & 0.05 \\ -0.23 & -0.15 & 0.66 & 0.98 & 0.93 & 0.17 & 0.76 \end{array}\right]$$

is chosen as random with A stable. The computation is carried out pointwise across the frequency grid, i.e., for each ω, $\sigma_1(\boldsymbol{T}(j\omega))$ and $\bar{\mu}_{\boldsymbol{\Delta}}(\boldsymbol{T}(j\omega))$ are computed and compared with the D-iteration upper bound. The uncertainty structure $\boldsymbol{\Delta}$ is taken to be diagonal, i.e., $S = 0, F = 4, m_j = 1$ for $j = 1, \ldots, 4$. The plots are shown in Figure 3.

6. Conclusions. By way of conclusion, we summarize our contribution.

- We have analyzed in detail the maximum robust stabilization problem subject to unstructured additive perturbations. We have shown that a critical direction exists in the uncertainty space, along which all maximum-norm boundary

perturbations are destabilizing for every optimal controller.

- We have shown that by imposing a parametric constraint in the most critical direction, the set of uncertainties guaranteed to be stabilized by a subset of all optimal controllers can be further extended. We have shown that the optimal solution to this problem is associated with the set of superoptimal controllers with respect to the first two levels, and we have obtained a closed-form expression for the improved robust stability radius which involves the first two superoptimal levels.
- By adapting out results to the structured uncertainty case, we have obtained an easily computable upper bound on the structured-singular value (which is tighter than the largest singular value), without the need to carry out a D-iteration. We have further shown that the minimization of this bound is equivalent to the minimization of the second largest singular value, which again motivates superoptimization.

There are a number of related research directions which we intend to pursue.

- For purposes of clarity, our technique has been restricted to unstructured additive uncertainty models. There is no conceptual difficulty, however, in extending our method to other types of unstructured uncertainty (multiplicative, inverse-multiplicative, coprime) or to include frequency weightings. Rather than analyzing each case individually, we intend to address the general problem involving linear fractional transformation uncertainty models [17]. This is likely to involve a general-distance superoptimal approximation problem, the solution of which is already in place [26], [8], [16], [13], [14], [12].
- Our method relies on Theorem 4.7 which generalizes a result in [15]. Section 5 suggests that generalizing this theorem should be useful in robust stability analysis of systems subject to structured uncertainty. We have derived some results in this direction which will be reported in a future publication.
- We intend to investigate the possibility of applying our method as an alternative to current μ-synthesis techniques. The main potential advantage of our approach is the possibility of avoiding the calculation of the optimal scaling "D-matrix" in the $D - K$ iteration [2] (currently carried out pointwise over a discretized frequency grid) by using instead the directionality information provided by the two Schmidt vectors, which define the worst-case direction in our setting. The success of such an approach will ultimately rest on how tightly we can overbound the structured singular value. Although our computational experience so far is promising, this remains an open question.

REFERENCES

[1] R. F. Curtain and H. J. Zwart, *An Introduction to Infinite-Dimensional Linear Systems Theory*, Springer-Verlag, New York, 1991.

[2] J. C. Doyle, *Lecture Notes in Advances in Multivariable Control*, ONR/Honeywell Workshop, Minneapolis, MN, 1984.

[3] B. A. Francis, *A Course in $\mathcal{H}_\infty$ Control Theory*, Springer-Verlag, Berlin, 1987.

[4] K. Glover, *All optimal Hankel-norm approximations of linear multivariable systems and their $\mathcal{L}_\infty$ error bounds*, Internat. J. Control, 39 (1984), pp. 1115–1193.

[5] K. Glover, *A tutorial on Hankel-norm approximation*, in From Data to Model, J. C. Willems, ed., Springer-Verlag, Berlin, New York, 1989.

[6] K. Glover, *Robust stabilization of linear multivariable systems: Relations to approximation*, Internat. J. Control, 43 (1986), pp. 741–746.

[7] K. Glover, D. J. N. Limebeer, J. C. Doyle, E. M. Kasenally, and M. G. Safonov, *A characterization of all solutions to the four block general distance problem*, SIAM J.

Control Optim., 29 (1991), pp. 283–324.

[8] D. W. Gu, M. C. Tsai, and I. Postlethwaite, *An algorithm for superoptimal $\mathcal{H}_\infty$ design: The 2-block case*, Automatica J. IFAC, 25 (1989), pp. 437–440.

[9] S. K. Gungah, *Maximally Robust Controllers for a Class of Unstructured Uncertainty*, Ph.D. thesis, London University, 1998.

[10] G. D. Halikias, *An affine parametrization of all one-block $\mathcal{H}_\infty$-optimal matrix interpolating functions*, Internat. J. Control, 57 (1993), pp. 1421–1441.

[11] G. D. Halikias and I. M. Jaimoukha, *Hierarchical optimization in $\mathcal{H}_\infty$*, IEEE Trans. Automat. Control, 43 (1998), pp. 1123–1128.

[12] G. D. Halikias and I. M. Jaimoukha, *The two-block superoptimal AAK problem*, Math. Control Signals Systems, 11 (1998), pp. 244–264.

[13] G. D. Halikias, D. J. N. Limebeer, and K. Glover, *A state-space analysis for the superoptimal Hankel-norm approximation problem*, SIAM J. Control Optim., 31 (1993), pp. 960–982.

[14] I. M. Jaimoukha and D. J. N. Limebeer, *A state-space algorithm for the solution of the 2-block superoptimal distance problem*, SIAM J. Control Optim., 31 (1993), pp. 1115–1134.

[15] N. A. Lehtomaki, D. A. Castanon, B. D. Levy, G. Stein, N. R. Sandel, and M. Athans, *Robustness and modeling error characterization*, IEEE Trans. Automat. Control, 29 (1984), pp. 212–220.

[16] D. J. N. Limebeer, G. D. Halikias, and K. Glover, *State-space algorithm for the computation of superoptimal matrix interpolating functions*, Internat. J. Control, 50 (1989), pp. 2431–2466.

[17] D. C. McFarlane and K. Glover, *Robust Controller Design Using Normalized Coprime Factor Plant Descriptions*, Lecture Notes in Control and Inform. Sci. 138, Springer-Verlag, Berlin, New York, 1990.

[18] P.-O. Nyman, *Improving robustness by superoptimization*, in Proceedings of the 3rd European Control Conference, Rome, Italy, 1995, pp. 1039–1044.

[19] A. Packard and J. C. Doyle, *The complex structured singular value*, Automatica J. IFAC, 21 (1993), pp. 79–109.

[20] S. Treil, *On superoptimal approximation by analytic and meromorphic matrix valued functions*, J. Funct. Anal., 131 (1995), pp. 386–414.

[21] M. C. Tsai, D. W. Gu, and I. Postlethwaite, *A state space approach to super-optimal $\mathcal{H}^\infty$ control problems*, IEEE Trans. Automat. Control, 33 (1988), pp. 833–843.

[22] M. C. Tsai, D. W. Gu, I. Postlethwaite, and B. D. O. Anderson, *A Pseudo-Singular Value Decomposition and Inner Functions in Superoptimal $\mathcal{H}_\infty$ Control*, Report OUEL 1738/88, University of Oxford, 1988.

[23] M. S. Verma, *Synthesis of Infinity-Norm Optimal Linear Feedback Systems*, Ph.D. thesis, University of Southern California, Los Angeles, CA, 1985.

[24] M. S. Verma, J. W. Helton, and E. A. Jonckheere, *Robust stabilization of a family of plants with varying number of right half plane poles*, in Proceedings of the American Control Conference, Seattle, WA, IEEE, Piscataway, NJ, 1986, pp. 1827–1832.

[25] M. Vidyasagar, *Control System Synthesis: A Factorization Approach*, MIT Press, Cambridge, MA, 1985.

[26] N. J. Young, *The Nevanlinna-Pick problem for matrix-valued functions*, J. Operator Theory, 15 (1986), pp. 239–265.

SIAM J. CONTROL OPTIM.
Vol. 38, No. 6, pp. 1830–1858

DISTURBANCE DECOUPLING FOR DESCRIPTOR SYSTEMS BY STATE FEEDBACK*

DELIN CHU† AND VOLKER MEHRMANN‡

Abstract. We study the disturbance decoupling problem for linear time invariant descriptor systems. We give necessary and sufficient conditions for the existence of a solution to the disturbance decoupling problem with or without stability via a proportional and/or derivative feedback that also makes the resulting closed-loop system regular and/or of index at most one. All results are proved constructively based on condensed forms that can be computed using orthogonal matrix transformations, i.e., transformations that can be implemented in a numerically stable way.

Key words. descriptor system, state feedback, disturbance decoupling, stability, orthogonal matrix transformation

AMS subject classifications. 93B05, 93B40, 93B52, 65F35

PII. S0363012900331891

1. Introduction.

We consider linear and time-invariant continuous descriptor systems of the form

$$\begin{aligned} E\dot{x}(t) &= Ax(t) + Bu(t) + Gq(t), \qquad x(t_0) = x_0,\ t \geq t_0, \\ y(t) &= Cx(t), \end{aligned} \tag{1}$$

where E, $A \in \mathbf{R}^{n\times n}$, $B \in \mathbf{R}^{n\times m}$, $G \in \mathbf{R}^{n\times p}$, $C \in \mathbf{R}^{q\times n}$, and $\dot{x} = dx/dt$. The term $q(t), t \geq 0$, represents a disturbance, which may represent modeling or measuring errors, noise, or higher order terms in linearization. We study the problem of constructing feedbacks that suppress this disturbance in the sense that $q(t)$ does not affect the input-output behavior of the system. In this paper, we study only square systems (E, A are square). This seems to be a restriction, since in general models that arise from automatic modeling or from heterogeneous systems are often rectangular, e.g., [16]. In [7], however, it was shown that for every rectangular system there exists an underlying square system that can be obtained via a numerically backward stable procedure by removing redundancies and higher index uncontrollable and unobservable parts in the system. It is possible to formulate all the results in this paper also for rectangular systems, by incorporating the transformations performed in [7] into the methods and the theorems. But this would make the paper, which is quite technical already, even longer and more technical. For this reason we restrict ourselves to the square case. Similarly, we also assume without loss of generality (w.l.o.g.) that B, G are full column rank and C is full row rank, i.e., $\operatorname{rank}(B) = m$, $\operatorname{rank}(G) = p$, $\operatorname{rank}(C) = q$. If this is not the case, then this can be easily achieved, via a numerically stable procedure, by removing the nullspaces and appropriate renaming of variables.

*Received by the editors December 15, 1999; accepted for publication (in revised form) January 4, 2000; published electronically July 5, 2000.
http://www.siam.org/journals/sicon/38-6/33189.html

†Department of Mathematics, National University of Singapore, Lower Kent Ridge Road, Singapore 119260 (matchudl@math.nus.edu.sg). The research of this author was supported by the Alexander von Humboldt Foundation.

‡Fakultät für Mathematik, TU Chemnitz, D-09107 Chemnitz, Germany (mehrmann@mathematik.tu-chemnitz.de). The research of this author was supported by Deutsche Forschungsgemeinschaft research grant Me 790/7-2.

The theory for (1) is well established from the analytical, geometric, and numerical points of view; see, e.g., [11, 17, 22, 30]. Existence and uniqueness of (classical) solutions to (1) for sufficiently smooth input functions and consistent initial values are guaranteed if (E, A) is *regular*, i.e., if $\det(\alpha E - \beta A) \neq 0$ for some $(\alpha, \beta) \in \mathbf{C}^2$. This means in particular that the system has to be square. The system (1) is said to have *index at most one* if the dimension of the largest nilpotent block in the Kronecker canonical form of (E, A) is at most one [14, 4]. It is well known that systems that are regular and of index at most one can be separated into purely dynamical and purely algebraic parts (fast and slow modes), and in theory the algebraic part can be eliminated to give a reduced-order standard system. The reduction process, however, may be ill-conditioned with respect to numerical computation. For this reason it is preferable to use descriptor system models rather than turning the problem into a standard system. Nonetheless most numerical simulation methods work well for systems of index at most one (see [4]) and the usual class of piecewise continuous input functions can be used. Also classical techniques for important control applications like stabilization, pole assignment, or linear quadratic control can be applied; see, e.g., [22, 5, 6].

If the index is larger than one, however, then impulses arise in the response of the system if the control is not sufficiently smooth [5, 11]. This restricts the set of admissable input functions and impulses can also arise due to the presence of modeling, measurement, linearization, and roundoff errors in the real system. Furthermore, the use of numerical integration methods is restricted; see [4].

There are essentially two possibilities to deal with higher index systems in the context of control systems. Either an index reduction is performed (see, e.g., [7, 18]) to obtain an equivalent system of index at most one, or an appropriate feedback control is chosen to ensure that the closed-loop system is regular and of index at most one. Techniques for the construction of such feedbacks were developed in [5, 6, 8] based on transformations to condensed forms via orthogonal matrix transformations, which can be implemented as numerically stable algorithms. This paper, which is strongly inspired by the work of [5, 6, 8], extends these techniques to the solution of the disturbance decoupling problem.

The *disturbance decoupling problem* for descriptor systems (1) can be stated as follows: *Find necessary and sufficient conditions under which there exists a proportional and derivative feedback of the form* $u(t) = Fx(t) - K\dot{x}(t)$, *such that matrix pencil* $(E + BK, A + BF)$ *is regular and of index at most one and* $C(s(E + BK) - (A + BF))^{-1}G = 0$, *where* $C(s(E + BK) - (A + BF))^{-1}G$ *is the transfer-function matrix of the closed-loop system*

$$(E + BK)\dot{x}(t) = (A + BF)x(t) + Gq(t), \quad y(t) = Cx(t). \tag{2}$$

For standard systems ($E = I$) this problem is well studied (see [23, 24, 25, 26, 27, 29, 30]). Our attention, however, will focus on the case that E is singular. Let us briefly summarize some previous results. The disturbance decoupling problem for continuous-time descriptor systems was first formulated in [21] and the problem was solved under the assumption, among other conditions, that the output is independent of the input disturbance in the sense that there is a set of admissible initial conditions such that the response of the system is zero. But, since the disturbance input is usually unknown, it is not clear how, and if at all, a given initial state x_0 can be qualified as an admissible initial condition. In [3] the problem was solved from the geometric point of view, using the concepts of sliding and coasting subspaces by means

of a set of necessary and sufficient conditions for obtaining disturbance decoupling in implicit discrete systems. These results are not constructive and numerically stable methods cannot be based on this approach. Furthermore the index of the system is not considered. In [19, 20] again the discrete time disturbance decoupling problem is discussed and structurally equivalent characterizations are presented for the solvability of the disturbance decoupling problems for implicit discrete-time systems.

In [1] the standard disturbance decoupling problem for continuous-time descriptor systems was considered as formulated in the standard state-space system theory [30], i.e., given the system (1), find (if possible) a proportional state feedback such that, regardless of the initial value of x_0, the disturbance input has no influence on the output of the systems for $t \geq 0$, and yet the uniqueness of solutions for the closed-loop system is ensured. Also in [1] necessary and sufficient conditions were given for solvability of the disturbance decoupling problem under the assumptions $\operatorname{rank}\begin{bmatrix} E & G \end{bmatrix} = n$ and $\operatorname{rank}\begin{bmatrix} E & B & G \end{bmatrix} = n$. But the obtained conditions are rather cumbersome and are only partly given in terms of the original data (E, A, B, C, G). Moreover, combined derivative and proportional state feedback, the index and stability of the system, and also numerical aspects of the algorithms have not been considered in the literature so far.

To demonstrate the great flexibility of our matrix pencil approach, we also discuss the extra requirement that the closed-loop system be stable, i.e., that all the finite generalized eigenvalues of $s(E + BK) - (A + BF)$ are in the open left half-plane. Furthermore a similar approach yields also the solution for partly measurable disturbances, i.e., we also study the use of a proportional and derivative feedback of the form $u(t) = Fx(t) - K\dot{x}(t) + Hq(t)$, such that the matrix pencil $(E + BK, A + BF)$ is regular, of index at most one, and

$$C(s(E + BK) - (A + BF))^{-1}(G + BH) = 0.$$

Again, we also include the stability of the closed-loop system as an extra requirement.

All our results are proved constructively, based on condensed forms under orthogonal matrix transformations which can be directly implemented as numerically stable algorithms.

The paper is organized as follows. In section 2 we introduce some notation and give some preliminary results. In sections 3 and 4 we solve the disturbance decoupling problems without and with stability, respectively. We discuss separately the case that the system is only regularized and that it also has index at most one. It should be noted that several more results on this topic could have been included but are omitted for lack of space. See the technical reports [9, 10].

2. Preliminaries. In this section we introduce the notation and give some preliminary results. We denote by $\deg(p)$ the degree of a polynomial p and by $\operatorname{rank}_g[M(s)]$ the generic rank of a rational matrix valued function $M(s)$ and by $\bar{\mathbf{C}}^+$ the closed right half-plane. Let the orthogonal complement of the space spanned by the columns of a matrix M be denoted by $M^\perp$. A matrix with orthogonal columns spanning the right nullspace of a matrix M is denoted by $S_\infty(M)$ and a matrix with orthogonal columns spanning the right nullspace of M^T by $T_\infty(M)$. For convenience of notation we identify a subspace and a matrix whose columns form an orthonormal basis of this subspace. These orthonormal bases will be available from the condensed forms that we determine.

In principle the complete analysis of descriptor systems could be based on the Kronecker canonical forms of the associated matrix pencils [14, 13], but it is in general

impossible to compute the Kronecker canonical form with a finite precision numerical algorithm, since small changes in the data can drastically change the canonical form. Instead one can obtain a condensed form under orthogonal equivalence transformations. This form, the generalized upper triangular (GUPTRI) form, is well analyzed [12, 13], and numerically stable algorithms are available and have been implemented in LAPACK [2]. The GUPTRI form displays all the invariants, in particular, the left and right Kronecker indices, but it is not the complete canonical form.

LEMMA 2.1 (see [12]). *Given a matrix pencil (E, A), $E, A \in \mathbf{R}^{n\times l}$, there exist orthogonal matrices $P \in \mathbf{R}^{n\times n}$, $Q \in \mathbf{R}^{l\times l}$ such that (PEQ, PAQ) are in the following GUPTRI form:*

$$(3)\quad P(sE-A)Q = \begin{array}{c} \\ n_1 \\ n_2 \\ n_3 \\ n_4 \end{array} \begin{array}{c} \begin{array}{cccc} l_1 & l_2 & l_3 & l_4 \end{array} \\ \begin{bmatrix} sE_{11}-A_{11} & sE_{12}-A_{12} & sE_{13}-A_{13} & sE_{14}-A_{14} \\ 0 & sE_{22}-A_{22} & sE_{23}-A_{23} & sE_{24}-A_{24} \\ 0 & 0 & sE_{33}-A_{33} & sE_{34}-A_{34} \\ 0 & 0 & 0 & sE_{44}-A_{44} \end{bmatrix} \end{array}.$$

Here $n_2 = l_2, n_3 = l_3$, $sE_{11} - A_{11}$, and $sE_{44} - A_{44}$ contain all left and right singular Kronecker blocks of $sE - A$, respectively. Furthermore, $sE_{22} - A_{22}$ and $sE_{33} - A_{33}$ are regular and contain the regular finite and infinite structure of $sE - A$, respectively.

Based on the form (3), we introduce the following spaces which we will use to describe a geometric, coordinate-free, characterization of the solution to the disturbance decoupling problem.

DEFINITION 2.2 (see [12]). *Given a matrix pencil (E, A), $E, A \in \mathbf{R}^{n\times l}$, and orthogonal matrices P, Q such that $P(sE - A)Q$ is of the form* (3). *Then*

1. *The minimal left reducing subspace $V_{m-l}[E, A]$ of (E, A) is the space spanned by the leading n_1 columns of P^T.*
2. *The minimal right reducing subspace $V_{m-r}[E, A]$ of (E, A) is the space spanned by the leading l_1 columns of Q.*
3. *The left reducing subspace corresponding to the finite spectrum of (E, A), denoted by $V_{f-l}[E, A]$, is the space spanned by the leading $n_1 + n_2$ columns of P^T.*
4. *The right reducing subspace corresponding to the finite spectrum of (E, A), denoted by $V_{f-r}[E, A]$, is the space spanned by the leading $l_1 + l_2$ columns of Q.*

The problem of constructing feedbacks such that the closed-loop system is regular, of index at most one, and stable has already been studied in detail in the literature. We summarize some of the relevant results in the following lemmas.

LEMMA 2.3 (see [5, 11, 28]). *Given $E, A \in \mathbf{R}^{n\times n}$, $B \in \mathbf{R}^{n\times m}$.*

(i) *There exists $F \in \mathbf{R}^{m\times n}$ such that $(E, A + BF)$ is regular if and only if*

$$(4)\qquad \operatorname{rank}_g \begin{bmatrix} sE - A & B \end{bmatrix} = n.$$

(ii) *There exists $F \in \mathbf{R}^{m\times n}$ such that $(E, A + BF)$ is regular and of index at most one if and only if*

$$(5)\qquad \operatorname{rank} \begin{bmatrix} E & AS_\infty(E) & B \end{bmatrix} = n.$$

(iii) *There exist $F, K \in \mathbf{R}^{m\times n}$ such that $(E+BK, A+BF)$ is regular and of index at most one if and only if*

$$(6)\qquad \operatorname{rank} \begin{bmatrix} E & AS_\infty(T_\infty^T(B)E) & B \end{bmatrix} = n.$$

(iv) *There exists a matrix* $F \in \mathbf{R}^{m\times n}$ *such that* $(E, A+BF)$ *is regular and stable if and only if*

$$\operatorname{rank}\begin{bmatrix} sE-A & B \end{bmatrix} = n \quad \forall s \in \mathbf{C}^{+}. \tag{7}$$

(v) *There exists a matrix* $F \in \mathbf{R}^{m\times n}$ *such that* $(E, A+BF)$ *is regular, stable, and of index at most one if and only if conditions* (5) *and* (7) *hold.*

(vi) *There exist matrices* $F, G \in \mathbf{R}^{m\times n}$ *such that* $(E+BG, A+BF)$ *is regular and stable if and only if condition* (7) *holds.*

(vii) *There exist matrices* $F, G \in \mathbf{R}^{m\times n}$ *such that* $(E+BG, A+BF)$ *is regular, stable, and of index at most one if and only if conditions* (6) *and* (7) *hold.*

Remark 2.1. Condition (5) is often called *controllability at infinity*, since if it holds, then the Jordan structure of the spectrum at infinity can be modified arbitrarily. This condition is not invariant under derivative feedback (see [5]), hence condition (6) is needed when combined state and derivative is used. Condition (4) is sometimes called *regularizability* [7]. If a system satisfies both (5) and (7), then it is called *strongly stabilizable*; see [22, p. 14].

The spaces occurring in Lemma 2.3 can be easily computed via numerically stable procedures, like singular value decomposition or rank revealing QR-decompositions [15, 2]. Thus, they can be checked numerically within the limitations of numerical rank decisions and nullspace computations in finite arithmetic.

For the disturbance decoupling problem, we need the following lemma.

LEMMA 2.4. *Consider a system of the form* (1). *If* (E, A) *is regular, then* $C(sE-A)^{-1}G = 0$ *if and only if*

$$\operatorname{rank}_g \begin{bmatrix} sE-A & G \\ C & 0 \end{bmatrix} = n.$$

Proof. The proof follows directly from the fact that for any $s \in \mathbf{C}$ with $\det(sE-A) \neq 0$ we have

$$\operatorname{rank}(C(sE-A)^{-1}G) = \operatorname{rank}\begin{bmatrix} sE-A & G \\ C & 0 \end{bmatrix} - \operatorname{rank}(sE-A). \qquad \square$$

We close this section with a technical lemma that we will use frequently in subsequent sections.

LEMMA 2.5. *Consider matrices* E, A, B *such that*

$$sE-A := \begin{matrix} \\ l_1 \\ l_2 \end{matrix}\begin{matrix} \begin{matrix} t \end{matrix} \\ \begin{bmatrix} sE_1-A_1 \\ -A_2 \end{bmatrix} \end{matrix}, \quad B := \begin{matrix} \\ l_1 \\ l_2 \end{matrix}\begin{matrix} \begin{matrix} r \end{matrix} \\ \begin{bmatrix} B_1 \\ B_2 \end{bmatrix} \end{matrix}$$

with $l_1 \leq t$ *and* B_2 *of full row rank.*

(i) *If*

$$\operatorname{rank}_g \begin{bmatrix} sE_1-A_1 & B_1 \\ -A_2 & B_2 \end{bmatrix} = l_1 + l_2, \tag{8}$$

then there exist a matrix $F \in \mathbf{R}^{r\times t}$ *and a nonsingular matrix* $Z \in \mathbf{R}^{t\times t}$ *such that*

$$(sE-A-BF)Z = \begin{matrix} \\ l_1 \\ l_2 \end{matrix}\begin{matrix} \begin{matrix} l_1 & t-l_1 \end{matrix} \\ \begin{bmatrix} s\Theta_1-\Phi_1 & -\Phi_2 \\ 0 & 0 \end{bmatrix} \end{matrix} \tag{9}$$

with (Θ_1, Φ_1) *regular.*

(ii) *If* (8) *holds and furthermore*

$$\operatorname{rank}\begin{bmatrix} E_1 & A_1S_\infty(E_1) & B_1 \\ 0 & A_2S_\infty(E_1) & B_2 \end{bmatrix} = l_1 + l_2, \tag{10}$$

then there exist a matrix $F \in \mathbf{R}^{r\times t}$ *and a nonsingular matrix* $Z \in \mathbf{R}^{t\times t}$ *such that* $(sE - A - BF)Z$ *has partitioning* (9) *with* (Θ_1, Φ_1) *regular and of index at most one.*

Proof. Let $W \in \mathbf{R}^{t\times t}$ and $Q \in \mathbf{R}^{r\times r}$ be orthogonal matrices such that

$$(sE-A)W = \begin{matrix} \\ l_1 \\ l_2 \end{matrix}\begin{matrix} \begin{matrix} l_1 & t-l_1 \end{matrix} \\ \begin{bmatrix} sE_{11}-A_{11} & -A_{12} \\ -A_{21} & -A_{22} \end{bmatrix} \end{matrix}, \quad \begin{bmatrix} B_1 \\ B_2 \end{bmatrix} Q = \begin{matrix} l_1 \\ l_2 \end{matrix}\begin{matrix} \begin{matrix} l_2 & r-l_2 \end{matrix} \\ \begin{bmatrix} B_{11} & B_{12} \\ B_{21} & 0 \end{bmatrix} \end{matrix}$$

with B_{21} nonsingular (note that B_2 has full row rank).

(i) Condition (8) is equivalent to

$$\operatorname{rank}_g \begin{bmatrix} sE_{11}-A_{11} & A_{12} & B_{11} & B_{12} \\ -A_{21} & A_{22} & B_{21} & 0 \end{bmatrix} = l_1 + l_2.$$

Since B_{21} is nonsingular, using Schur complements, this is equivalent to

$$\operatorname{rank}_g \begin{bmatrix} sE_{11} - (A_{11} - B_{11}B_{21}^{-1}A_{21}) & A_{12} - B_{11}B_{21}^{-1}A_{22} & B_{12} \end{bmatrix} = l_1.$$

Then, applying Lemma 2.3(iii) immediately gives the existence of a regularizing feedback.

(ii) Analogously, (10) is equivalent to

$$\operatorname{rank}\begin{bmatrix} E_{11} & A_{11}S_\infty(E_{11}) & A_{12} & B_{11} & B_{12} \\ 0 & A_{21}S_\infty(E_{11}) & A_{22} & B_{21} & 0 \end{bmatrix} = l_1 + l_2,$$

which is equivalent to

$$\operatorname{rank}\begin{bmatrix} E_{11} & (A_{11} - B_{11}B_{21}^{-1}A_{21})S_\infty(E_{11}) & (A_{12} - B_{11}B_{21}^{-1}A_{22}) & B_{12} \end{bmatrix} = l_1.$$

Lemma 2.3(i) then gives the existence of a feedback that makes the system regular and of index at most one.

Actually the feedbacks in both cases can be constructed explicitly as follows. Let

$$Z := W\begin{bmatrix} I & 0 \\ \tilde{Z} & I \end{bmatrix}, \quad F := Q\begin{bmatrix} F_{11} & F_{12} \\ F_{21} & 0 \end{bmatrix} W^T,$$

where F_{11} and F_{12} are constructed from

$$B_{21}\begin{bmatrix} F_{11} & F_{12} \end{bmatrix} = -\begin{bmatrix} A_{21} & A_{22} \end{bmatrix} \tag{11}$$

and $\tilde{Z}$ and F_{21} are constructed via the algorithm given in the appendix of [5]. ∎

Remark 2.2. The construction of the feedback in the proof of Lemma 2.5 needs the solution of a linear system in (11). This system may be very ill-conditioned and hence a numerical solution of this system may create large errors. A different approach to constructing the desired feedback is the following.

Using a rank revealing QR-decomposition, construct an orthogonal matrix P such that

$$P \begin{bmatrix} B_{11} \\ B_{21} \end{bmatrix} =: \begin{bmatrix} 0 \\ \tilde{B}_{21} \end{bmatrix}$$

with $\tilde{B}_{21}$ nonsingular. Set, with compatible partitioning,

$$P \begin{bmatrix} sE_{11} - A_{11} & A_{12} \\ -A_{21} & A_{22} \end{bmatrix} =: \begin{bmatrix} s\tilde{E}_{11} - \tilde{A}_{11} & \tilde{A}_{12} \\ s\tilde{E}_{21} - \tilde{A}_{21} & \tilde{A}_{22} \end{bmatrix}, \quad P \begin{bmatrix} B_{12} \\ 0 \end{bmatrix} =: \begin{bmatrix} \tilde{B}_{12} \\ \tilde{B}_{22} \end{bmatrix}.$$

Then apply the algorithm in the appendix of [5] to compute $\tilde{Z}$ and F_{21} such that $(\tilde{E}_{11}, \tilde{A}_{11} + \tilde{A}_{12}X + \tilde{B}_{12}F_{21})$ is regular and of index at most one.

After having introduced the preliminaries, in the next sections we discuss the disturbance decoupling problem.

3. The disturbance decoupling problem. In this section, we first establish a condensed form for matrix quintuples (E, A, B, C, G) under orthogonal equivalence transformations and then solve the disturbance decoupling problem without the stability requirement.

The general philosophy is to generate an equivalent representation of the original system that displays the system properties and can be used to construct the desired feedbacks in the solution of the disturbance decoupling problem. The main feature of the new condensed form is that it is (in contrast to canonical forms that are used in previous work on this subject, like, e.g., [1]) based on orthogonal matrix transformations which can be implemented as numerically stable algorithms, thus guaranteeing robust computation of the desired quantities, if this is possible. The spaces that we will need for the solution are the following (with the notation introduced in section 2).

$$\Pi := T_\infty \left(\begin{bmatrix} B & G \\ 0 & 0 \end{bmatrix} \right), \quad \Psi := T_\infty(G), \quad \Lambda_r := V_{f-r} \left[\Pi^T \begin{bmatrix} E \\ 0 \end{bmatrix}, \Pi^T \begin{bmatrix} A \\ C \end{bmatrix} \right],$$

$$\Lambda_l := V_{f-l} \left[\Pi^T \begin{bmatrix} E \\ 0 \end{bmatrix}, \Pi^T \begin{bmatrix} A \\ C \end{bmatrix} \right], \quad \Lambda_t := \begin{bmatrix} \Pi^\perp & \Pi \end{bmatrix} \begin{bmatrix} I & 0 \\ 0 & \Lambda_l \end{bmatrix}. \tag{12}$$

With the abbreviations $\Gamma_1 := \left[\begin{smallmatrix} 0 & \Psi^T E \\ 0 & 0 \end{smallmatrix}\right]$ and $\Gamma_2 := \left[\begin{smallmatrix} \Psi^T B & \Psi^T A \\ 0 & 0 \end{smallmatrix}\right]$ we further introduce the spaces

$$\begin{aligned} &\Lambda_1 := \Lambda_t^T \begin{bmatrix} E \\ 0 \end{bmatrix} \Lambda_r, \quad \Lambda_2 := \Lambda_t^T \begin{bmatrix} A \\ C \end{bmatrix} \Lambda_r, \\ &\Lambda_3 := \Lambda_t^T \begin{bmatrix} B \\ 0 \end{bmatrix}, \quad \Lambda_4 := (V_{f-l}^\perp[\Gamma_1, \Gamma_2])^T \, \Gamma_1 V_{f-r}^\perp[\Gamma_1, \Gamma_2]. \end{aligned} \tag{13}$$

These spaces can be easily obtained from the following condensed form under orthogonal transformations.

THEOREM 3.1. *Consider a system of the form* (1) *with* $E, A \in \mathbf{R}^{n\times n}$, $B \in \mathbf{R}^{n\times m}$, $G \in \mathbf{R}^{n\times p}$, $C \in \mathbf{R}^{q\times n}$. *Then there exist orthogonal matrices* $U, V \in \mathbf{R}^{n\times n}$ *such that*

$$UEV = \begin{array}{c} \\ p \\ \tilde{n}_2 \\ \tilde{n}_3 \\ \tilde{n}_4 \\ \tilde{n}_5 \end{array} \begin{array}{c} \begin{array}{ccc} n_1 & n_2 & n_3 \end{array} \\ \begin{bmatrix} E_{11} & E_{12} & E_{13} \\ E_{21} & E_{22} & E_{23} \\ 0 & E_{32} & E_{33} \\ 0 & E_{42} & E_{43} \\ 0 & 0 & E_{53} \end{bmatrix} \end{array}, \quad UAV = \begin{array}{c} \\ p \\ \tilde{n}_2 \\ \tilde{n}_3 \\ \tilde{n}_4 \\ \tilde{n}_5 \end{array} \begin{array}{c} \begin{array}{ccc} n_1 & n_2 & n_3 \end{array} \\ \begin{bmatrix} A_{11} & A_{12} & A_{13} \\ A_{21} & A_{22} & A_{23} \\ A_{31} & A_{32} & A_{33} \\ 0 & A_{42} & A_{43} \\ 0 & 0 & A_{53} \end{bmatrix} \end{array},$$

$$UB = \begin{array}{c} \\ p \\ \tilde{n}_2 \\ \tilde{n}_3 \\ \tilde{n}_4 \\ \tilde{n}_5 \end{array} \begin{bmatrix} B_1 \\ B_2 \\ B_3 \\ 0 \\ 0 \end{bmatrix}, \quad UG = \begin{array}{c} \\ p \\ \tilde{n}_2 \\ \tilde{n}_3 \\ \tilde{n}_4 \\ \tilde{n}_5 \end{array} \begin{bmatrix} G_1 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}, \quad CV = \begin{array}{c} \begin{array}{ccc} n_1 & n_2 & n_3 \end{array} \\ \begin{bmatrix} 0 & C_2 & C_3 \end{bmatrix} \end{array}, \tag{14}$$

where G_1, E_{21}, B_3, *and* E_{42} *are of full row rank and furthermore*

$$\operatorname{rank}(sE_{53} - A_{53}) = n_3, \quad \operatorname{rank} \begin{bmatrix} sE_{42} - A_{42} & sE_{43} - A_{43} \\ 0 & sE_{53} - A_{53} \\ C_2 & C_3 \end{bmatrix} = n_2 + n_3 \qquad \forall s \in \mathbf{C}.$$

Proof. The proof is given constructively via Algorithm 1 in Appendix A. ∎

Using condensed form (14) we can determine directly the following important spaces and their dimensions.

LEMMA 3.2. *Let* E, A, B, C, G *be in the condensed form* (14).

(i) *We have*

$$\begin{aligned} \tau &:= \dim\left(V_{f-l}\left[\begin{bmatrix} 0 & \Psi^T E \\ 0 & 0 \end{bmatrix}, \begin{bmatrix} \Psi^T B & \Psi^T A \\ 0 & C \end{bmatrix}\right]\right) = \tilde{n}_2, \\ \mu &:= \dim\left(V_{f-r}^{\perp}\left[\begin{bmatrix} \Psi^T B & \Psi^T E \\ 0 & 0 \end{bmatrix}, \begin{bmatrix} 0 & \Psi^T A \\ 0 & C \end{bmatrix}\right]\right) = n_2 + n_3, \\ \eta &:= \dim\left(V_{f-l}\left[\begin{bmatrix} \Psi^T B & \Psi^T E \\ 0 & 0 \end{bmatrix}, \begin{bmatrix} 0 & \Psi^T A \\ 0 & C \end{bmatrix}\right]\right) - \tau = \tilde{n}_3. \end{aligned} \tag{15}$$

(ii) *Let* $S := S_\infty\left(\begin{bmatrix} E_{11} \\ E_{21} \end{bmatrix}\right)$; *then*

$$\operatorname{rank} \begin{bmatrix} E_{11} & A_{11}S & B_1 \\ E_{21} & A_{21}S & B_2 \\ 0 & A_{31}S & B_3 \end{bmatrix} = \operatorname{rank} \begin{bmatrix} \Lambda_1 & \Lambda_2 S_\infty(\Lambda_1) & \Lambda_3 \end{bmatrix}, \text{ with}$$

$$\operatorname{rank} \begin{bmatrix} E_{11} \\ E_{21} \\ 0 \end{bmatrix} = \operatorname{rank}(\Lambda_1), \quad \operatorname{rank} \begin{bmatrix} E_{32} & E_{33} \\ E_{42} & E_{43} \\ 0 & E_{53} \end{bmatrix} = \operatorname{rank}(\Lambda_4). \tag{16}$$

(iii) *The matrix*

$$T_\infty^T\left(\begin{bmatrix} E_{11} & B_1 \\ E_{21} & B_2 \\ 0 & B_3 \end{bmatrix}\right) \begin{bmatrix} A_{11} \\ A_{21} \\ A_{31} \end{bmatrix} S_\infty\left(T_\infty^T\left(\begin{bmatrix} B_1 \\ B_2 \\ B_3 \end{bmatrix}\right) \begin{bmatrix} E_{11} \\ E_{21} \\ 0 \end{bmatrix}\right)$$

has full row rank if and only if $T_\infty^T(\begin{bmatrix} \Lambda_1 & \Lambda_3 \end{bmatrix})\Lambda_2 S_\infty(T_\infty^T(\Lambda_3)\Lambda_1)$ *is of full row rank.*

Proof. The proof is given in Appendix B. ∎

Based on the condensed form (14) and Lemma 2.3 we obtain the following necessary conditions for the existence of feedbacks that make the system regular and of index at most one.

LEMMA 3.3. *Given system* (1) *in condensed form* (14)*, we have the following.*

(i) *If there exists a feedback* $F \in \mathbf{R}^{m\times n}$ *such that* $(E, A+BF)$ *is regular and of index at most one, i.e., if condition* (5) *holds, then* $\tilde{n}_5 = n_3$, $E_{53} = 0$, A_{53} *is nonsingular, and furthermore,*

$$\operatorname{rank}\begin{bmatrix} E_{32} \\ E_{42} \end{bmatrix} = \operatorname{rank}\begin{bmatrix} E_{32} & E_{33} \\ E_{42} & E_{43} \end{bmatrix}. \tag{17}$$

(ii) *If there exist* $F, K \in \mathbf{R}^{m\times n}$ *such that* $(E+BK, A+BF)$ *is regular and of index at most one, i.e., if condition* (6) *holds, then* $\tilde{n}_5 = n_3$, $E_{53} = 0$, A_{53} *is nonsingular and*

$$\operatorname{rank}\begin{bmatrix} E_{11} & E_{12} & E_{13} & B_1 \\ E_{21} & E_{22} & E_{23} & B_2 \\ 0 & E_{32} & E_{33} & B_3 \\ 0 & E_{42} & E_{43} & 0 \end{bmatrix} = \operatorname{rank}\begin{bmatrix} E_{11} & E_{12} & B_1 \\ E_{21} & E_{22} & B_2 \\ 0 & E_{32} & B_3 \\ 0 & E_{42} & 0 \end{bmatrix}. \tag{18}$$

Proof. Let the system be given in the condensed form (14).

(i) If there exists $F := \begin{bmatrix} F_1 & F_2 & F_3 \end{bmatrix}$, partitioned conformly to (14), such that $(E, A+BF)$ is regular and of index at most one, then the last block row in (14), which cannot be modified by proportional feedback, must satisfy $\operatorname{rank}_g(sE_{53} - A_{53}) = \tilde{n}_5$. But $sE_{53} - A_{53}$ is of full column rank for any $s \in C$ and thus $n_3 = \tilde{n}_5$ and $\det(sE_{53} - A_{53}) = \det(-A_{53})$. Therefore, the nonsingularity of A_{53} follows directly from the regularity of $(E, A+BF)$. Moreover,

$$\begin{aligned}
\operatorname{rank}(E) &= \deg(\det(sE - A - BF)) \\
&= \deg\left(\det\left(\begin{bmatrix} sE_{11} - A_{11} - B_1F_1 & sE_{12} - A_{12} - B_1F_2 \\ sE_{21} - A_{21} - B_2F_1 & sE_{22} - A_{22} - B_2F_2 \\ -A_{31} - B_3F_1 & sE_{32} - A_{32} - B_3F_2 \\ 0 & sE_{42} - A_{42} \end{bmatrix}\right)\right) \\
&\quad + \deg(\det(sE_{53} - A_{53})) \\
&= \deg\left(\det\left(\begin{bmatrix} sE_{11} - A_{11} - B_1F_1 & sE_{12} - A_{12} - B_1F_2 \\ sE_{21} - A_{21} - B_2F_1 & sE_{22} - A_{22} - B_2F_2 \\ -A_{31} - B_3F_1 & sE_{32} - A_{32} - B_3F_2 \\ 0 & sE_{42} - A_{42} \end{bmatrix}\right)\right) \\
&\leq \operatorname{rank}\left(\begin{bmatrix} E_{11} & E_{12} \\ E_{21} & E_{22} \\ 0 & E_{32} \\ 0 & E_{42} \end{bmatrix}\right).
\end{aligned}$$

But, we have

$$\operatorname{rank}(E) \geq \operatorname{rank}\left(\begin{bmatrix} E_{11} & E_{12} \\ E_{21} & E_{22} \\ 0 & E_{32} \\ 0 & E_{42} \end{bmatrix}\right) + \operatorname{rank}(E_{53}),$$

and hence $E_{53} = 0$. We also have

$$\operatorname{rank}\begin{bmatrix} E_{11} & E_{12} & E_{13} \\ E_{21} & E_{22} & E_{23} \\ 0 & E_{32} & E_{33} \\ 0 & E_{42} & E_{43} \end{bmatrix} = \operatorname{rank}(E) = \operatorname{rank}\begin{bmatrix} E_{11} & E_{12} \\ E_{21} & E_{22} \\ 0 & E_{32} \\ 0 & E_{42} \end{bmatrix},$$

which implies (17).

(ii) If there exist $F := \begin{bmatrix} F_1 & F_2 & F_3 \end{bmatrix}$ and $K := \begin{bmatrix} K_1 & K_2 & K_3 \end{bmatrix}$ such that $(E + BK, A + BF)$ is regular and of index at most one, then, since E_{53}, A_{53} are not affected by these feedbacks, it follows from part (i) that $\tilde{n}_5 = n_3$, $E_{53} = 0$, and A_{53} is nonsingular. Similarly to part (i), we get that

$$\operatorname{rank}(E + BK) = \operatorname{rank}\left(\begin{bmatrix} E_{11} + B_1K_1 & E_{12} + B_1K_2 \\ E_{21} + B_2K_1 & E_{22} + B_2K_2 \\ B_3K_1 & E_{32} + B_3K_2 \\ 0 & E_{42} \end{bmatrix}\right).$$

However, since $E_{53} = 0$, we also have

$$\operatorname{rank}(E + BK) = \operatorname{rank}\left(\begin{bmatrix} E_{11} + B_1K_1 & E_{12} + B_1K_2 & E_{13} + B_1K_3 \\ E_{21} + B_2K_1 & E_{22} + B_2K_2 & E_{23} + B_2K_3 \\ B_3K_1 & E_{32} + B_3K_2 & E_{33} + B_3K_3 \\ 0 & E_{42} & E_{43} \end{bmatrix}\right),$$

and hence

$$\begin{aligned} &\operatorname{rank}\left(\begin{bmatrix} E_{11} + B_1K_1 & E_{12} + B_1K_2 & E_{13} + B_1K_3 \\ E_{21} + B_2K_1 & E_{22} + B_2K_2 & E_{23} + B_2K_3 \\ B_3K_1 & E_{32} + B_3K_2 & E_{33} + B_3K_3 \\ 0 & E_{42} & E_{43} \end{bmatrix}\right) \\ &= \operatorname{rank}\left(\begin{bmatrix} E_{11} + B_1K_1 & E_{12} + B_1K_2 \\ E_{21} + B_2K_1 & E_{22} + B_2K_2 \\ B_3K_1 & E_{32} + B_3K_2 \\ 0 & E_{42} \end{bmatrix}\right), \end{aligned}$$

which implies (18). □

We will now apply these results to solve the disturbance decoupling problem.

We present the results in a coordinate free way, but note that all the quantities are available via numerically stable procedures from the results presented before.

THEOREM 3.4. *Consider a system of the form* (1) *and let the spaces* Λ_i *be as in* (13) *and* τ, η, *and* μ *as in* (15).

(a) *There exists a feedback matrix* $F \in \mathbf{R}^{m\times n}$ *such that the pencil* $(E, A + BF)$ *is regular, of index at most one, and* $C(sE - (A + BF))^{-1}G = 0$ *if and only if condition* (5) *and furthermore the following three conditions hold:*

$$\tau + \mu \le n - p; \tag{19}$$

$$\operatorname{rank}(\Lambda_1) + \operatorname{rank}(\Lambda_4) = \operatorname{rank}(E); \tag{20}$$

$$\operatorname{rank}\begin{bmatrix} \Lambda_1 & \Lambda_2 S_\infty(\Lambda_1) & \Lambda_3 \end{bmatrix} = p + \tau + \eta. \tag{21}$$

(b) *There exist feedback matrices* $F, K \in \mathbf{R}^{m\times n}$*, such that the pencil* $(E+BK, A+BF)$ *is regular of index at most one and* $C(s(E+BK)-(A+BF))^{-1}G = 0$ *if and only if condition* (6) *holds,*

$$\text{(22)}\quad \operatorname{rank}_g \begin{bmatrix} T_\infty^T(G)(sE-A) & T_\infty^T(G)B \\ C & 0 \end{bmatrix} - \operatorname{rank} \begin{bmatrix} T_\infty^T(G)B \\ 0 \end{bmatrix} \le n-p,$$

and $W_1 := T_\infty^T(\begin{bmatrix} \Lambda_1 & \Lambda_3 \end{bmatrix})\Lambda_2 S_\infty(T_\infty^T(\Lambda_3)\Lambda_1)$ *has full row rank.*

Proof. By Theorem 3.1 there exist orthogonal matrices that transform the system to the form (14). Thus, for the proof we may assume, w.l.o.g., that the system is already in form (14).

(a) *Necessity.* Let $F \in \mathbf{R}^{m\times n}$ be such that $(E, A+BF)$ is regular, of index at most one, and $C(sE-(A+BF))^{-1}G = 0$. Partition $F =: \begin{bmatrix} F_1 & F_2 & F_3 \end{bmatrix}$ compatibly with (E, A, B). Then (5) follows directly from Lemma 2.3(ii). Furthermore, by Lemma 2.4 we have that $\operatorname{rank}_g \begin{bmatrix} sE_{21}-A_{21}-B_2F_1 \\ -A_{32}-B_3F_1 \end{bmatrix} = n-p-n_2-n_3$. Hence, we obtain

$$n-p-n_2-n_3 \ge \operatorname{rank}_g \begin{bmatrix} sE_{21}-A_{21}-B_2F_1 \end{bmatrix} \ge \operatorname{rank}(E_{21}) = \tilde{n}_2,$$

i.e., by (15), condition (19) holds.

To prove conditions (20)–(21), observe from Lemma 3.3, and since $(E, A+BF)$ is regular and of index at most one, we have

$$\begin{aligned} \operatorname{rank}(E) &= \deg(\det(sE-A-BF)) = \deg(\det(sE_{53}-A_{53})) \\ &\quad + \deg\left(\det\begin{bmatrix} sE_{11}-A_{11}-B_1F_1 & sE_{12}-A_{12}-B_1F_2 \\ sE_{21}-A_{21}-B_2F_1 & sE_{22}-A_{22}-B_2F_2 \\ -A_{31}-B_3F_1 & sE_{32}-A_{32}-B_3F_3 \\ 0 & sE_{42}-A_{42} \end{bmatrix}\right) \\ &= \deg\left(\det\begin{bmatrix} sE_{11}-A_{11}-B_1F_1 & sE_{12}-A_{12}-B_1F_2 \\ sE_{21}-A_{21}-B_2F_1 & sE_{22}-A_{22}-B_2F_2 \\ -A_{31}-B_3F_1 & sE_{32}-A_{32}-B_3F_3 \\ 0 & sE_{42}-A_{42} \end{bmatrix}\right). \end{aligned} \tag{23}$$

Hence, from (23) it follows that

$$\begin{aligned} \operatorname{rank}(E) &= \operatorname{rank}\begin{bmatrix} E_{11} & E_{12} \\ E_{21} & E_{22} \\ 0 & E_{32} \\ 0 & E_{42} \end{bmatrix} \\ &= \deg\left(\det\begin{bmatrix} sE_{11}-A_{11}-B_1F_1 & sE_{12}-A_{12}-B_1F_2 \\ sE_{21}-A_{21}-B_2F_1 & sE_{22}-A_{22}-B_2F_2 \\ -A_{31}-B_3F_1 & sE_{32}-A_{32}-B_3F_2 \\ 0 & sE_{42}-A_{42} \end{bmatrix}\right). \end{aligned} \tag{24}$$

Using that E_{21}, E_{42} are of full row rank, we may assume, w.l.o.g. (by performing appropriate equivalence transformations), that

$$E_{11} = \begin{bmatrix} 0 & 0 & 0 \\ 0 & \Sigma_{11} & 0 \end{bmatrix}, \quad E_{21} = \begin{bmatrix} \Sigma_{21} & 0 & 0 \end{bmatrix}, \quad E_{32} = \begin{bmatrix} 0 & 0 & 0 \\ 0 & \Theta_{32} & 0 \end{bmatrix},$$
$$E_{42} = \begin{bmatrix} 0 & 0 & \Sigma_{42} \end{bmatrix}, \quad E_{12} = \begin{bmatrix} \tilde{E}_{12} & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}, \quad E_{22} = 0,$$

where $\Sigma_{11} \in \mathbf{R}^{p_1\times p_1}, \Sigma_{21} \in \mathbf{R}^{\tilde{n}_2\times\tilde{n}_2}, \Sigma_{42} \in \mathbf{R}^{\tilde{n}_4\times\tilde{n}_4}$ are nonsingular and $\Theta_{32} \in \mathbf{R}^{(n_2-\tilde{n}_4)\times t}$ is of full column rank. Partition accordingly, $A_{42} = \begin{bmatrix} \Phi_{64} & \Phi_{65} & \Phi_{66} \end{bmatrix}$, and

$$A_{11}+B_1F_1 = \begin{bmatrix} \Phi_{11} & \Phi_{12} & \Phi_{13} \\ \Phi_{21} & \Phi_{22} & \Phi_{23} \end{bmatrix}, \quad A_{12}+B_1F_2 = \begin{bmatrix} \Phi_{14} & \Phi_{15} & \Phi_{16} \\ \Phi_{24} & \Phi_{25} & \Phi_{26} \end{bmatrix},$$
$$A_{21}+B_2F_1 = \begin{bmatrix} \Phi_{31} & \Phi_{32} & \Phi_{33} \end{bmatrix}, \quad A_{22}+B_2F_2 = \begin{bmatrix} \Phi_{34} & \Phi_{35} & \Phi_{36} \end{bmatrix},$$
$$A_{31}+B_3F_1 = \begin{bmatrix} \Phi_{41} & \Phi_{42} & \Phi_{43} \\ \Phi_{51} & \Phi_{52} & \Phi_{53} \end{bmatrix}, \quad A_{32}+B_3F_2 = \begin{bmatrix} \Phi_{44} & \Phi_{45} & \Phi_{46} \\ \Phi_{54} & \Phi_{55} & \Phi_{56} \end{bmatrix}.$$

Then (24) yields that

$$\begin{bmatrix} T_\infty^T(\tilde{E}_{12})\Phi_{13} & T_\infty^T(\tilde{E}_{12})\Phi_{14}S_\infty(\tilde{E}_{12}) \\ \Phi_{43} & \Phi_{44}S_\infty(\tilde{E}_{12}) \\ T_\infty^T(\Theta_{32})\Phi_{53} & T_\infty^T(\Theta_{32})\Phi_{54}S_\infty(\tilde{E}_{12}) \end{bmatrix}$$

is nonsingular.
Hence, we obtain that

$$\operatorname{rank}\begin{bmatrix} T_\infty^T(\tilde{E}_{12})\Phi_{13} \\ \Phi_{43} \\ T_\infty^T(\Theta_{32})\Phi_{53} \end{bmatrix} = n_1 - p_1 - \tilde{n}_2. \tag{25}$$

But, from Lemma 2.4 we have

$$\operatorname{rank}_g \begin{bmatrix} s\Sigma_{21}-\Phi_{31} & -\Phi_{32} & -\Phi_{33} \\ -\Phi_{41} & -\Phi_{42} & -\Phi_{43} \\ -\Phi_{51} & -\Phi_{52} & -\Phi_{53} \end{bmatrix} = n-p-n_2-n_3 = n_1-p,$$

and hence

$$\operatorname{rank}\begin{bmatrix} \Phi_{43} \\ T_\infty^T(\Theta_{32})\Phi_{53} \end{bmatrix} \le \operatorname{rank}\begin{bmatrix} \Phi_{43} \\ \Phi_{53} \end{bmatrix} \le n_1-p-\tilde{n}_2. \tag{26}$$

Thus, by (25), we have

$$\operatorname{rank}(T_\infty^T(\tilde{E}_{12})\Phi_{13}) \ge p-p_1, \tag{27}$$

where $p-p_1$ is the number of rows of Φ_{13}. This implies that $\tilde{E}_{12}=0$, and hence (24) implies that

$$\operatorname{rank}(E) = \operatorname{rank}\begin{bmatrix} E_{11} \\ E_{21} \\ 0 \end{bmatrix} + \operatorname{rank}\begin{bmatrix} E_{32} \\ E_{42} \end{bmatrix},$$

and thus, Lemma 3.3(i) yields that

$$\operatorname{rank}(E) = \operatorname{rank}\begin{bmatrix} E_{11} \\ E_{21} \\ 0 \end{bmatrix} + \operatorname{rank}\begin{bmatrix} E_{32} & E_{33} \\ E_{42} & E_{43} \\ 0 & E_{53} \end{bmatrix}.$$

Then, Lemma 3.2(ii) gives (20). Furthermore, using (26) and (27) we have that

$$\operatorname{rank}(\Phi_{13}) = p-p_1, \qquad \operatorname{rank}\begin{bmatrix} \Phi_{43} \\ \Phi_{53} \end{bmatrix} = \operatorname{rank}\begin{bmatrix} \Phi_{43} \\ T_\infty^T(\Theta_{32})\Phi_{53} \end{bmatrix} = n_1-p-\tilde{n}_2,$$

and thus with $S := S_\infty([\begin{smallmatrix} E_{11} \\ E_{21} \end{smallmatrix}])$, we obtain

$$\operatorname{rank} \begin{bmatrix} E_{11} & A_{11}S & B_1 \\ E_{21} & A_{21}S & B_2 \\ 0 & A_{31}S & B_3 \end{bmatrix} = p + \tilde{n}_2 + \tilde{n}_3. \tag{28}$$

Then (21) follows directly from Lemma 3.2(ii).

Sufficiency. Using conditions (19) and (21) and Lemmas 2.3(ii) and (2.5)(ii), there exist $\tilde{F}_1 \in \mathbf{R}^{m\times n_1}$ and a nonsingular matrix $Z \in \mathbf{R}^{n_1\times n_1}$ such that

$$\begin{bmatrix} sE_{11} - A_{11} - B_1\tilde{F}_1 \\ sE_{21} - A_{21} - B_2\tilde{F}_1 \\ -A_{31} - B_3\tilde{F}_1 \end{bmatrix} Z$$

$$= \begin{array}{c} \\ p \\ \tilde{n}_2 \\ \tilde{n}_3 \end{array} \begin{array}{c} \begin{array}{cc} p+\tilde{n}_2 & n_1-p-\tilde{n}_2 \end{array} \\ \begin{bmatrix} s\Theta_{11} - \Phi_{11} & -\Phi_{12} \\ s\Theta_{21} - \Phi_{21} & -\Phi_{22} \\ 0 & 0 \end{bmatrix} \end{array}$$

with $([\begin{smallmatrix} \Theta_{11} \\ \Theta_{21} \end{smallmatrix}], [\begin{smallmatrix} \Phi_{11} \\ \Phi_{21} \end{smallmatrix}])$ regular and of index at most one.

By (5) and (20) and Lemma 3.3 it follows that $\tilde{n}_5 = n_3$, $E_{53} = 0$, A_{53} is nonsingular, and furthermore,

$$\operatorname{rank}(E) = \operatorname{rank} \begin{bmatrix} E_{11} \\ E_{21} \\ 0 \end{bmatrix} + \operatorname{rank} \begin{bmatrix} E_{32} \\ E_{42} \end{bmatrix}. \tag{29}$$

Note that B_3 and E_{42} are of full row rank, so by Lemma 2.3(ii) there exist matrices $\hat{F}_1 \in \mathbf{R}^{m\times(n_1-p-\tilde{n}_2)}$ and $F_2 \in \mathbf{R}^{m\times n_2}$ such that

$$\left(\begin{bmatrix} 0 & E_{32} \\ 0 & E_{42} \end{bmatrix}, \begin{bmatrix} B_3\hat{F}_1 & A_{32} + B_3F_2 \\ 0 & A_{42} \end{bmatrix} \right)$$

is regular and of index at most one. Taking $F_1 := \tilde{F}_1 + \begin{bmatrix} 0 & \hat{F}_1 \end{bmatrix} Z^{-1}$, $F := \begin{bmatrix} F_1 & F_2 & F_3 \end{bmatrix}$ with F_3 arbitrary, it is easy to see that $(E, A + BF)$ is regular, of index at most one, and $C(sE - A - BF)^{-1}G = 0$.

(b) *Necessity.* Let $F, K \in \mathbf{R}^{m\times n}$ be such that $(E + BK, A + BF)$ is regular, of index at most one, and $C(s(E+BK)-(A+BF))^{-1}G = 0$. Then condition (6) follows directly from Lemma 2.3(iii). As in (a), by Lemma 2.4 we have that

$$\operatorname{rank}_g \begin{bmatrix} s(E_{21} + B_2K_1) - (A_{21} + B_2F_1) \\ sB_3K_1 - (A_{31} + B_3F_1) \end{bmatrix} = n - p - n_2 - n_3,$$

which implies that

$$\begin{aligned} &\operatorname{rank}_g \begin{bmatrix} sE_{21} - A_{21} & B_2 \\ -A_{31} & B_3 \end{bmatrix} - \operatorname{rank} \begin{bmatrix} B_2 \\ B_3 \end{bmatrix} \\ &= \operatorname{rank}_g \begin{bmatrix} s(E_{21} + B_2K_1) - (A_{21} + B_2F_1) & B_2 \\ sB_3K_1 - (A_{31} + B_3F_1) & B_3 \end{bmatrix} - \operatorname{rank} \begin{bmatrix} B_2 \\ B_3 \end{bmatrix} \\ &\le \operatorname{rank}_g \begin{bmatrix} s(E_{21} + B_2K_1) - (A_{21} + B_2F_1) \\ sB_3K_1 - (A_{31} + B_3F_1) \end{bmatrix} = n - p - n_2 - n_3. \end{aligned}$$

Thus,

$$n-p \geq n_2+n_3+\operatorname{rank}_g \begin{bmatrix} sE_{21}-A_{21} & B_2 \\ -A_{31} & B_3 \end{bmatrix} - \operatorname{rank} \begin{bmatrix} B_2 \\ B_3 \end{bmatrix}$$
$$= \operatorname{rank}_g \begin{bmatrix} T_\infty^T(G)(sE-A) & T_\infty^T(G)B \\ C & 0 \end{bmatrix} - \operatorname{rank} \begin{bmatrix} T_\infty^T(G)B \\ 0 \end{bmatrix},$$

i.e., (22) holds.

Since $\left[\begin{smallmatrix} E_{21}+B_2K_1 & B_2 \\ B_3K_1 & B_3 \end{smallmatrix}\right]$ is of full row rank, from (28) we have that

$$\begin{bmatrix} E_{11}+B_1K_1 & (A_{11}+B_1F_1)\tilde{S} & B_1 \\ E_{21}+B_2K_1 & (A_{21}+B_2F_1)\tilde{S} & B_2 \\ B_3K_1 & (A_{31}+B_3F_1)\tilde{S} & B_3 \end{bmatrix}$$

is of full row rank, where

$$\tilde{S} := S_\infty \left(\begin{bmatrix} E_{11}+B_1K_1 \\ E_{21}+B_2K_1 \\ B_3K_1 \end{bmatrix} \right).$$

Equivalently, we obtain that

$$\begin{bmatrix} E_{11} & A_{11}\tilde{S}_\infty & B_1 \\ E_{21} & A_{21}\tilde{S}_\infty & B_2 \\ 0 & A_{31}\tilde{S}_\infty & B_3 \end{bmatrix}$$

is of full row rank. Thus, using the relation between

$$S_\infty \left(T_\infty^T \left(\begin{bmatrix} B_1 \\ B_2 \\ B_3 \end{bmatrix} \right) \begin{bmatrix} E_{11} \\ E_{21} \\ 0 \end{bmatrix} \right)$$

and $\tilde{S}_\infty$, we have that

$$T_\infty^T \left(\begin{bmatrix} E_{11} & B_1 \\ E_{21} & B_2 \\ 0 & B_3 \end{bmatrix} \right) \begin{bmatrix} A_{11} \\ A_{21} \\ A_{31} \end{bmatrix} S_\infty \left(T_\infty^T \left(\begin{bmatrix} B_1 \\ B_2 \\ B_3 \end{bmatrix} \right) \begin{bmatrix} E_{11} \\ E_{21} \\ 0 \end{bmatrix} \right)$$

is of full row rank and using Lemma 3.2(iii) so is W_1.

Sufficiency. Since E_{21}, B_3 are of full row rank, using Lemma 2.1, we can determine orthogonal matrices $P_1 \in \mathbf{R}^{p\times p}$, $P_2 \in \mathbf{R}^{(\tilde{n}_2+\tilde{n}_3)\times(\tilde{n}_2+\tilde{n}_3)}$, and $W \in \mathbf{R}^{m\times m}$ such that

$$\begin{bmatrix} P_1 & \\ & P_2 \end{bmatrix} \begin{bmatrix} sE_{11}-A_{11} & sE_{12}-A_{12} & sE_{13}-A_{13} \\ sE_{21}-A_{21} & sE_{22}-A_{22} & sE_{23}-A_{23} \\ -A_{31} & sE_{32}-A_{32} & sE_{33}-A_{33} \end{bmatrix}$$
$$= \begin{array}{c} \\ l_1 \\ p-l_1 \\ l_2 \\ \tilde{n}_2+\tilde{n}_3-l_2 \end{array} \begin{array}{c} \begin{array}{ccc} n_1 & n_2 & n_3 \end{array} \\ \begin{bmatrix} s\Theta_{11}-\Phi_{11} & s\Theta_{12}-\Phi_{12} & s\Theta_{13}-\Phi_{13} \\ s\Theta_{21}-\Phi_{21} & s\Theta_{22}-\Phi_{12} & s\Theta_{23}-\Phi_{13} \\ s\Theta_{31}-\Phi_{21} & s\Theta_{32}-\Phi_{32} & s\Theta_{33}-\Phi_{33} \\ s\Theta_{41}-\Phi_{41} & s\Theta_{42}-\Theta_{42} & s\Theta_{43}-\Phi_{43} \end{bmatrix} \end{array},$$

$$\begin{bmatrix} P_1 & \\ & P_2 \end{bmatrix} \begin{bmatrix} B_1 \\ B_2 \\ B_3 \end{bmatrix} W = \begin{array}{c} \\ l_1 \\ p-l_1 \\ l_2 \\ \tilde{n}_2+\tilde{n}_3-l_2 \end{array} \begin{bmatrix} \tilde{B}_{11} & 0 \\ \tilde{B}_{21} & \tilde{B}_{22} \\ 0 & 0 \\ \tilde{B}_{41} & 0 \end{bmatrix}$$

with $\tilde{B}_{22}, \Theta_{31}$ of full row rank and $\tilde{B}_{41}$ nonsingular. Now determine K_{11}, K_{21} from

$$\begin{bmatrix} \tilde{B}_{21} & \tilde{B}_{22} \\ \tilde{B}_{41} & 0 \end{bmatrix} W \begin{bmatrix} K_{11} \\ K_{21} \end{bmatrix} = - \begin{bmatrix} \Theta_{21} \\ \Theta_{41} \end{bmatrix}.$$

The last two conditions imply that $p + l_2 \le n_1$ and

$$\operatorname{rank} \begin{bmatrix} \Theta_{11} + \tilde{B}_{11}K_{11} & \Phi_{11}\tilde{S} & \tilde{B}_{11} & 0 \\ 0 & \Phi_{21}\tilde{S} & \tilde{B}_{21} & \tilde{B}_{22} \\ \Theta_{31} & \Phi_{31}\tilde{S} & 0 & 0 \\ 0 & \Phi_{41}\tilde{S} & \tilde{B}_{41} & 0 \end{bmatrix} = p + \tilde{n}_2 + \tilde{n}_3,$$

where $\tilde{S} = S_\infty([\begin{smallmatrix} \Theta_{11}+\tilde{B}_{11}K_{11} \\ \Theta_{31} \end{smallmatrix}])$. By Lemma 2.5 there exist F_{11}, F_{21}, and a nonsingular matrix Z satisfying

$$\begin{bmatrix} s(\Theta_{11} + \tilde{B}_{11}K_{11}) - \Phi_{11} - \tilde{B}_{11}F_{11} \\ -\Phi_{21} - \tilde{B}_{21}F_{11} - \tilde{B}_{22}F_{21} \\ s\Theta_{31} \\ -\Phi_{41} - \tilde{B}_{41}F_{11} \end{bmatrix} Z$$

$$= \begin{array}{c} \\ l_1 \\ p - l_1 \\ l_2 \\ \tilde{n}_2 + \tilde{n}_3 - l_2 \end{array} \begin{array}{c} \begin{array}{cc} p + l_2 & n_1 - p - l_2 \end{array} \\ \begin{bmatrix} s\tilde{\Theta}_{11} - \tilde{\Phi}_{11} & -\tilde{\Phi}_{12} \\ -\tilde{\Phi}_{21} & -\tilde{\Phi}_{22} \\ s\tilde{\Theta}_{31} - \tilde{\Phi}_{31} & -\tilde{\Phi}_{32} \\ 0 & 0 \end{bmatrix} \end{array}$$

with

$$\left(\begin{bmatrix} \tilde{\Theta}_{11} \\ 0 \\ \tilde{\Theta}_{31} \end{bmatrix}, \begin{bmatrix} \tilde{\Phi}_{11} \\ \tilde{\Phi}_{21} \\ \tilde{\Phi}_{31} \end{bmatrix} \right)$$

regular and of index at most one. Using condition (6) and Lemma 3.3(ii), we obtain that $\tilde{n}_5 = n_3$. Hence, we can determine $\hat{K}_{11}, \hat{K}_{21}, K_{12}, K_{22}$ such that

$$\begin{bmatrix} \tilde{B}_{21} & \tilde{B}_{22} \end{bmatrix} \begin{bmatrix} \hat{K}_{11} & K_{12} \\ \hat{K}_{21} & K_{22} \end{bmatrix} = - \begin{bmatrix} 0 & \Theta_{22} \end{bmatrix},$$

and $\begin{bmatrix} \tilde{B}_{41}\hat{K}_{11} & \Theta_{42} + \tilde{B}_{41}K_{12} \\ 0 & E_{42} \end{bmatrix}$ is nonsingular. Let

$$K_1 := \begin{bmatrix} K_{11} \\ K_{21} \end{bmatrix} + \begin{bmatrix} 0 & \hat{K}_{11} \\ 0 & \hat{K}_{21} \end{bmatrix} Z^{-1}, \quad K_2 := \begin{bmatrix} K_{12} \\ K_{22} \end{bmatrix}, \quad K_3 := \begin{bmatrix} K_{13} \\ K_{23} \end{bmatrix},$$

and

$$F_1 := \begin{bmatrix} F_{11} \\ F_{21} \end{bmatrix},$$

where K_{13}, K_{23} will be determined later, and set $K := W \begin{bmatrix} K_1 & K_2 & K_3 \end{bmatrix}$ and $F := W \begin{bmatrix} F_1 & 0 & 0 \end{bmatrix}$. Then we have

$$\begin{bmatrix} P_1 & & \\ & P_2 & \\ & & I \end{bmatrix} (s(E+BK)-(A+BF)) \begin{bmatrix} Z & & \\ & I & \end{bmatrix}$$

$$= \begin{bmatrix} s\tilde{\Theta}_{11}-\tilde{\Phi}_{11} & s\tilde{B}_{11}\hat{K}_{11}-\tilde{\Phi}_{12} & s(\Theta_{12}+\tilde{B}_{11}K_{12})-\Phi_{12} & s(\Theta_{13}+\tilde{B}_{11}K_{13})-\Phi_{13} \\ -\tilde{\Phi}_{21} & -\tilde{\Phi}_{12} & -\Phi_{22} & s(\Theta_{23}+\tilde{B}_{21}K_{13}+\tilde{B}_{22}K_{23})-\Phi_{23} \\ s\tilde{\Theta}_{31}-\tilde{\Phi}_{31} & -\tilde{\Phi}_{32} & s\Theta_{32}-\Phi_{32} & s\Theta_{33}-\Phi_{33} \\ 0 & s(\tilde{B}_{41}\hat{K}_{11}) & s(\Theta_{42}+\tilde{B}_{41}K_{12})-\Phi_{42} & s(\Theta_{43}+\tilde{B}_{41}K_{13})-\Phi_{43} \\ 0 & 0 & sE_{42}-A_{42} & sE_{43}-A_{43} \\ 0 & 0 & 0 & sE_{53}-A_{53} \end{bmatrix}.$$

By (18), and since $\begin{bmatrix} \tilde{B}_{41}\hat{K}_{11} & \Theta_{42}+\tilde{B}_{41}K_{12} \\ 0 & E_{42} \end{bmatrix}$ is nonsingular, we have that

$$\operatorname{rank} \begin{bmatrix} \tilde{\Theta}_{11} & \Theta_{12} & \Theta_{13} & \tilde{B}_{11} \\ \tilde{\Theta}_{31} & \Theta_{32} & \Theta_{33} & 0 \\ 0 & \Theta_{42} & \Theta_{43} & \tilde{B}_{41} \\ 0 & E_{42} & E_{43} & 0 \end{bmatrix} = \operatorname{rank} \begin{bmatrix} \tilde{\Theta}_{11} & \Theta_{12} & \tilde{B}_{11} \\ \tilde{\Theta}_{31} & \Theta_{32} & 0 \\ 0 & \Theta_{42} & \tilde{B}_{41} \\ 0 & E_{42} & 0 \end{bmatrix}.$$

Therefore, there exists K_{13} such that

$$\operatorname{rank} \begin{bmatrix} \tilde{\Theta}_{11} & \Theta_{12}+\tilde{B}_{11}K_{12} & \Theta_{13}+\tilde{B}_{11}K_{13} \\ \tilde{\Theta}_{31} & \Theta_{32} & \Theta_{33} \\ 0 & \Theta_{42}+\tilde{B}_{41}K_{12} & \Theta_{43}+\tilde{B}_{41}K_{13} \\ 0 & E_{42} & E_{43} \end{bmatrix} = \operatorname{rank} \begin{bmatrix} \tilde{\Theta}_{11} & \Theta_{12}+\tilde{B}_{11}K_{12} \\ \tilde{\Theta}_{31} & \Theta_{32} \\ 0 & \Theta_{42}+\tilde{B}_{41}K_{12} \\ 0 & E_{42} \end{bmatrix}.$$

Finally, we determine K_{23} from

$$\tilde{B}_{22}K_{23} = -(\Theta_{23}+\tilde{B}_{21}K_{13}).$$

We have from Lemma 3.3 that $E_{53}=0$ and that A_{53} is nonsingular. Thus, for F, K determined by the described procedure, $(E+BK, A+BF)$ is regular and of index at most one. It is also easy to see that $C(s(E+BK)-(A+BF))^{-1}G = 0$. □

Remark 3.1. If the index one condition is not required, then in case (a) necessary and sufficient conditions are given by (4), (19), and

$$\operatorname{rank}_g \begin{bmatrix} sE-A & B \\ C & 0 \end{bmatrix} = \operatorname{rank}_g \begin{bmatrix} sE-A & B & G \\ C & 0 & 0 \end{bmatrix} \tag{30}$$

and in case (b) by (4), (22), and (30). The proof of the necessary and sufficient conditions and the construction of feedbacks that regularize the system without achieving index at most one are discussed in detail in [9].

Remark 3.2. The previous results can also be modified to the case when only derivative feedback is used. Since the results are essentially dual results to the ones for state feedback by exchanging the roles of E and A, we omit them. Details are given in the reports [9, 10].

4. The disturbance decoupling problem with stability. In this section we study the case where the extra requirement that the closed-loop system be stable is added. Similarly to section 3, first we prove a condensed form for matrix quintuples (E, A, B, C, G) under orthogonal equivalence transformations and determine different left and right reducing subspaces that are needed for the solution of the disturbance decoupling problem with stability.

THEOREM 4.1. *Given a system of the form* (1)*, there exist orthogonal matrices* $U, V \in \mathbf{R}^{n\times n}$ *such that*

$$U(sE-A)V = \begin{array}{c} \\ \tilde{n}_1 \\ \tilde{n}_2 \\ \tilde{n}_3 \\ \tilde{n}_4 \\ \tilde{n}_5 \\ \tilde{n}_6 \end{array} \begin{array}{c} \begin{array}{cccc} n_1 & n_2 & n_3 & n_4 \end{array} \\ \begin{bmatrix} sE_{11}-A_{11} & sE_{12}-A_{12} & sE_{13}-A_{13} & sE_{14}-A_{14} \\ -A_{21} & sE_{22}-A_{22} & sE_{23}-A_{23} & sE_{24}-A_{24} \\ -A_{31} & sE_{32}-A_{32} & sE_{33}-A_{33} & sE_{34}-A_{34} \\ 0 & sE_{42}-A_{42} & sE_{43}-A_{43} & sE_{44}-A_{44} \\ 0 & 0 & sE_{53}-A_{53} & sE_{54}-A_{54} \\ 0 & 0 & 0 & sE_{64}-A_{64} \end{bmatrix} \end{array},$$

$$UB = \begin{array}{c} \\ \tilde{n}_1 \\ \tilde{n}_2 \\ \tilde{n}_3 \\ \tilde{n}_4 \\ \tilde{n}_5 \\ \tilde{n}_6 \end{array} \begin{bmatrix} B_1 \\ 0 \\ B_3 \\ 0 \\ 0 \\ 0 \end{bmatrix}, \quad UG = \begin{array}{c} \\ \tilde{n}_1 \\ \tilde{n}_2 \\ \tilde{n}_3 \\ \tilde{n}_4 \\ \tilde{n}_5 \\ \tilde{n}_6 \end{array} \begin{bmatrix} G_1 \\ G_2 \\ G_3 \\ 0 \\ 0 \\ 0 \end{bmatrix}, \quad CV = \begin{array}{c} \begin{array}{cccc} n_1 & n_2 & n_3 & n_4 \end{array} \\ \begin{bmatrix} 0 & 0 & C_3 & C_4 \end{bmatrix} \end{array}, \tag{31}$$

where E_{11}*,* G_2*,* B_3*, and* E_{53} *are of full row rank,* E_{42} *is square and nonsingular (i.e.,* $n_2 = \tilde{n}_4$*,), and furthermore* $\forall s \in \mathbf{C}$

$$\operatorname{rank}(sE_{64}-A_{64}) = n_4, \quad \operatorname{rank}\begin{bmatrix} sE_{53}-A_{53} \\ C_3 \end{bmatrix} = n_3,$$
$$\operatorname{rank}\begin{bmatrix} sE_{11}-A_{11} & B_1 & G_1 \\ -A_{21} & 0 & G_2 \\ -A_{31} & B_3 & G_3 \end{bmatrix} = \tilde{n}_1 + \tilde{n}_2 + \tilde{n}_3.$$

Proof. The proof is constructive and can be obtained by an algorithm similar to Algorithm 1. A detailed description of the method can be found in [10]. ∎

Similarly to the proof of Lemma 3.2, using the condensed form (31) we can characterize the following spaces:

$$\begin{aligned} &\Delta_r := V_{m-r}\left[\Pi^T\begin{bmatrix} E \\ 0 \end{bmatrix}, \Pi^T\begin{bmatrix} A \\ C \end{bmatrix}\right], \quad \Delta_l := V_{m-l}\left[\Pi^T\begin{bmatrix} E \\ 0 \end{bmatrix}, \Pi^T\begin{bmatrix} A \\ C \end{bmatrix}\right], \\ &\Delta_t := \begin{bmatrix} \Pi^{\perp} & \Pi \end{bmatrix}\begin{bmatrix} I & 0 \\ 0 & \Delta_l \end{bmatrix}, \\ &\Delta_1 := \Delta_t^T\begin{bmatrix} E \\ 0 \end{bmatrix}\Delta_r, \qquad \Delta_2 := \Delta_t^T\begin{bmatrix} A \\ C \end{bmatrix}\Delta_r, \\ &\Delta_3 := \Delta_t^T\begin{bmatrix} B \\ 0 \end{bmatrix}, \qquad \Delta_4 := \Delta_t^T\begin{bmatrix} G \\ 0 \end{bmatrix}, \end{aligned} \tag{32}$$

where Π is as in (12). We introduce furthermore the following indices which are determined by the condensed form of Theorem 4.1:

$$\begin{aligned}\nu &:= \operatorname{rank}\begin{bmatrix} B & G \end{bmatrix} + \operatorname{rank}(\Delta_l), \quad \xi := \operatorname{rank}(\Delta_r),\\ \chi &:= \dim\left(V_{m-l}\left[\begin{bmatrix} E & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}, \begin{bmatrix} A & B & G \\ C & 0 & 0 \end{bmatrix}\right]\right)\\ &\quad + \dim\left(V_{f-l}\left[\begin{bmatrix} E & B & G \\ 0 & 0 & 0 \end{bmatrix}, \begin{bmatrix} A & 0 & 0 \\ C & 0 & 0 \end{bmatrix}\right]\right)\\ &\quad - \dim\left(V_{f-l}\left[\begin{bmatrix} E & B & 0 \\ 0 & 0 & 0 \end{bmatrix}, \begin{bmatrix} A & 0 & G \\ C & 0 & 0 \end{bmatrix}\right]\right). \end{aligned} \tag{33}$$

From the form (31) we then immediately obtain that

$$\xi = n_1, \qquad \chi = \tilde{n}_1 + \tilde{n}_2, \qquad \nu = \tilde{n}_1 + \tilde{n}_2 + \tilde{n}_3.$$

THEOREM 4.2. *Consider a system of the form* (1)*, spaces* Δ_i *as in* (32)*, and indices* χ*,* ξ*, and* ν *as in* (33)*.*

(a) *There exist feedback matrices* $F \in \mathbf{R}^{m\times n}$ *and* $H \in \mathbf{R}^{m\times d}$ *such that* $(E, A + BF)$ *is regular, of index at most one, stable, and* $C(sE - (A + BF))^{-1}(G + BH) = 0$ *if and only if conditions* (5) *and* (7) *and the following three conditions hold:*

$$\operatorname{rank}\begin{bmatrix} s\Delta_1 - \Delta_2 & \Delta_3 \end{bmatrix} = \nu \quad \forall s \in \mathbf{C}^+, \tag{34}$$

$$\operatorname{rank}\begin{bmatrix} \Delta_1 & \Delta_2 S_\infty(\Delta_1) & \Delta_3 \end{bmatrix} = \nu, \tag{35}$$

$$\chi \le \xi. \tag{36}$$

(b) *There exist feedback matrices* $F, K \in \mathbf{R}^{m\times n}$*, and* $H \in \mathbf{R}^{m\times d}$ *such that* $(E + BK, A + BF)$ *is regular, of index at most one, stable, and* $C(s(E + BK) - (A + BF))^{-1}(G + BH) = 0$ *if and only if conditions* (6)*,* (7)*, and* (34)*, as well as*

$$\operatorname{rank}_g \begin{bmatrix} T_\infty^T(B)(sE - A) & T_\infty^T(B)G \\ C & 0 \end{bmatrix} \le n, \tag{37}$$

hold and furthermore $W_2 := T_\infty^T(\begin{bmatrix} \Delta_1 & \Delta_3 \end{bmatrix})\Delta_2 S_\infty(T_\infty^T(\Delta_3)\Delta_1)$ *is of full row rank.*

Proof. Let $U, V \in \mathbf{R}^{n\times n}$ be orthogonal matrices such that $U(sE - A)V$, UB, CV, and UG are in the form (31). Then condition (34) translates to

$$\operatorname{rank}\begin{bmatrix} sE_{11} - A_{11} & B_1 \\ -A_{21} & 0 \\ -A_{31} & B_3 \end{bmatrix} = \tilde{n}_1 + \tilde{n}_2 + \tilde{n}_3 \quad \forall s \in \mathbf{C}^+$$

and (36) means that $\tilde{n}_1 + \tilde{n}_2 \le n_1$. Condition (35) translates to

$$\operatorname{rank}\begin{bmatrix} E_{11} & A_{11}S_\infty(E_{11}) & B_1 \\ 0 & A_{21}S_\infty(E_{11}) & 0 \\ 0 & A_{31}S_\infty(E_{11}) & B_3 \end{bmatrix} = \tilde{n}_1 + \tilde{n}_2 + \tilde{n}_3$$

and W_2 has full rank if and only if

$$T_\infty^T\left(\begin{bmatrix} E_{11} & B_1 \\ 0 & 0 \\ 0 & B_3 \end{bmatrix}\right) \begin{bmatrix} A_{11} \\ A_{21} \\ A_{31} \end{bmatrix} S_\infty\left(T_\infty^T\left(\begin{bmatrix} B_1 \\ 0 \\ B_3 \end{bmatrix}\right)\begin{bmatrix} E_{11} \\ 0 \\ 0 \end{bmatrix}\right)$$

has full row rank.

(a) *Necessity.* Let $F \in \mathbf{R}^{m\times n}$ and $H \in \mathbf{R}^{m\times d}$ be such that $(E, A+BF)$ is regular, of index at most one, and stable and $C(sE-(A+BF))^{-1}(G+BH)=0$. Conditions (5) and (7) follow directly by Lemma 2.3(v). If we partition

$$FV = \begin{matrix} \\ \end{matrix}\begin{matrix} n_1 & n_2 & n_3 & n_4 \end{matrix}$$
$$\begin{bmatrix} F_1 & F_2 & F_3 & F_4 \end{bmatrix},$$

then by Lemma 2.4 we have that

$$\sum_{i=1}^{4} n_i = n = \operatorname{rank}_g \begin{bmatrix} sE-(A+BF) & G+BH \\ C & 0 \end{bmatrix}$$
$$= n_2+n_3+n_4+\operatorname{rank}_g \begin{bmatrix} sE_{11}-(A_{11}+B_1F_1) & G_1+B_1H \\ -A_{21} & G_2 \\ -(A_{31}+B_3F_1) & G_3+B_3H \end{bmatrix}.$$

Hence,

$$\operatorname{rank}_g \begin{bmatrix} sE_{11}-(A_{11}+B_1F_1) & G_1+B_1K \\ -A_{21} & G_2 \\ -(A_{31}+B_3F_1) & G_3+B_3K \end{bmatrix} = n_1, \tag{38}$$

which implies condition (36), since

$$\tilde{n}_1+\tilde{n}_2 = \operatorname{rank}_g \begin{bmatrix} sE_{11}-(A_{11}+B_1F_1) & G_1+B_1H \\ -A_{21} & G_2 \end{bmatrix} \le n_1.$$

To show (34) let P_1 be an orthogonal matrix such that

$$P_1^T \begin{bmatrix} G_1+B_1H \\ G_2 \\ G_3+B_3H \end{bmatrix} = \begin{matrix} t_1 \\ \tilde{t}_2 \end{matrix} \begin{bmatrix} \tilde{G}_1 \\ 0 \end{bmatrix} \tag{39}$$

with $\tilde{G}_1$ of full row rank. Set

$$\begin{matrix} t_1 \\ \tilde{t}_2 \end{matrix} \begin{bmatrix} s\tilde{E}_{11}-\tilde{A}_{11} \\ s\tilde{E}_{21}-\tilde{A}_{21} \end{bmatrix} := P_1^T \begin{bmatrix} sE_{11}-(A_{11}+B_1F_1) \\ -A_{21} \\ -(A_{31}+B_3F_1) \end{bmatrix} \tag{40}$$

and compute the GUPTRI form of $(\tilde{E}_{21}, \tilde{A}_{21})$

$$\hat{P}_1^T(s\tilde{E}_{21}-\tilde{A}_{21})Q_1 = \begin{matrix} \\ t_2 \\ t_3 \end{matrix} \begin{matrix} r_1 & r_2 \end{matrix} \begin{bmatrix} s\Theta_{21}-\Phi_{21} & s\Theta_{22}-\Phi_{22} \\ 0 & s\Theta_{32}-\Phi_{32} \end{bmatrix} \tag{41}$$

with Θ_{21} of full row rank and $s\Theta_{32}-\Phi_{32}$ of full column rank $\forall s \in \mathbf{C}$. Set

$$\begin{matrix} r_1 & r_2 \end{matrix}$$
$$\begin{bmatrix} s\Theta_{11}-\Phi_{11} & s\Theta_{12}-\Phi_{12} \end{bmatrix}$$

$$:= (s\tilde{E}_{11}-\tilde{A}_{11})Q_1, \quad \begin{matrix} t_1 \\ t_2 \\ t_3 \end{matrix} \begin{bmatrix} \Psi_1 \\ \Psi_2 \\ \Psi_3 \end{bmatrix} := \begin{bmatrix} I & \\ & \hat{P}_1^T \end{bmatrix} P_1^T \begin{bmatrix} B_1 \\ 0 \\ B_3 \end{bmatrix}. \tag{42}$$

Since

$$\operatorname{rank}\begin{bmatrix} sE_{11}-A_{11} & B_1 & G_1 \\ -A_{21} & 0 & G_2 \\ -A_{31} & B_3 & G_3 \end{bmatrix} = \operatorname{rank}\begin{bmatrix} s\Theta_{11}-\Phi_{11} & s\Theta_{12}-\Phi_{12} & \Psi_1 & \tilde{G}_1 \\ s\Theta_{21}-\Phi_{21} & s\Theta_{22}-\Phi_{22} & \Psi_2 & 0 \\ 0 & s\Theta_{32}-\Phi_{32} & \Psi_3 & 0 \end{bmatrix}$$
$$= \tilde{n}_1+\tilde{n}_2+\tilde{n}_3 = t_1+t_2+t_3$$

$\forall s \in \mathbf{C}$, it follows that $\operatorname{rank}\begin{bmatrix} s\Theta_{32}-\Phi_{32} & \Psi_3 \end{bmatrix} = t_3 \ \forall s \in \mathbf{C}$. By (38) we also have that $t_1+t_2+r_2 = t_1+\operatorname{rank}_g(s\Theta_{21}-\Phi_{21})+r_2 = n_1 = r_1+r_2$, or equivalently we have that

$$t_1+t_2 = r_1 \text{ and } \begin{bmatrix} s\Theta_{11}-\Phi_{11} \\ s\Theta_{21}-\Phi_{21} \end{bmatrix} \text{ is square.} \tag{43}$$

We know that $(E, A+BF)$ is regular and stable, so we have that $([\begin{smallmatrix}\Theta_{11}\\ \Theta_{21}\end{smallmatrix}], [\begin{smallmatrix}\Phi_{11}\\ \Phi_{21}\end{smallmatrix}])$ is regular and stable. Therefore, we have $\forall s \in \mathbf{C}^+$ that

$$\operatorname{rank}\begin{bmatrix} sE_{11}-A_{11} & B_1 \\ -A_{21} & 0 \\ -A_{31} & B_3 \end{bmatrix} = \operatorname{rank}\begin{bmatrix} sE_{11}-(A_{11}+B_1F_1) & B_1 \\ -A_{21} & 0 \\ -(A_{31}+B_3F_1) & B_3 \end{bmatrix}$$
$$= \operatorname{rank}\begin{bmatrix} s\Theta_{11}-\Phi_{11} & s\Theta_{12}-\Phi_{12} & \Psi_1 \\ s\Theta_{21}-\Phi_{21} & s\Theta_{22}-\Phi_{22} & \Psi_2 \\ 0 & s\Theta_{32}-\Phi_{32} & \Psi_3 \end{bmatrix}$$
$$= t_1+t_2+\operatorname{rank}\begin{bmatrix} s\Theta_{32}-\Phi_{32} & \Psi_3 \end{bmatrix} = t_1+t_2+t_3 = \tilde{n}_1+\tilde{n}_2+\tilde{n}_3,$$

which gives condition (34).

We have that E_{11}, G_2, and B_3 are of full row rank, so that $\begin{bmatrix} \Theta_{32} & \Psi_3 \end{bmatrix}$ is also of full row rank. Since $(E, A+BF)$ is regular and of index at most one, we have that $([\begin{smallmatrix}\Theta_{11}\\ \Theta_{21}\end{smallmatrix}], [\begin{smallmatrix}\Phi_{11}\\ \Phi_{21}\end{smallmatrix}])$ is also regular and of index at most one. Then, using the standard characterization of pencils that are regular and of index at most one, e.g., [5], we obtain that $\operatorname{rank}[\begin{smallmatrix}\Theta_{11} & \Phi_{11}\hat{S}\\ \Theta_{21} & \Phi_{21}\hat{S}\end{smallmatrix}] = t_1+t_2$ with $\hat{S} = S_\infty([\begin{smallmatrix}\Theta_{11}\\ \Theta_{21}\end{smallmatrix}])$. Therefore, we have

$$\tilde{n}_1+\tilde{n}_2+\tilde{n}_3 = t_1+t_2+t_3 \geq \operatorname{rank}\begin{bmatrix} E_{11} & A_{11}S_\infty(E_{11}) & B_1 \\ 0 & A_{21}S_\infty(E_{11}) & 0 \\ 0 & A_{31}S_\infty(E_{11}) & B_3 \end{bmatrix}$$
$$= \operatorname{rank}\begin{bmatrix} E_{11} & (A_{11}+B_1F_1)S_\infty(E_{11}) & B_1 \\ 0 & A_{21}S_\infty(E_{11}) & 0 \\ 0 & (A_{31}+B_3F_1)S_\infty(E_{11}) & B_3 \end{bmatrix}$$
$$= \operatorname{rank}\begin{bmatrix} \Theta_{11} & \Theta_{12} & \tilde{\Phi}_1\tilde{S} & \Psi_1 \\ \Theta_{21} & \Theta_{22} & \tilde{\Phi}_2\tilde{S} & \Psi_3 \\ 0 & \Theta_{32} & \tilde{\Phi}_3\tilde{S} & \Psi_3 \end{bmatrix}$$
$$\geq \operatorname{rank}\begin{bmatrix} \Theta_{11} & \Theta_{12} & \Phi_{11}\hat{S} & \Psi_1 \\ \Theta_{21} & \Theta_{22} & \Phi_{21}\hat{S} & \Psi_2 \\ 0 & \Theta_{32} & 0 & \Psi_3 \end{bmatrix} = t_1+t_2+t_3,$$

where

$$\begin{bmatrix} \tilde{\Phi}_1 \\ \tilde{\Phi}_2 \\ \tilde{\Phi}_3 \end{bmatrix} = \begin{bmatrix} \Phi_{11} & \Phi_{12} \\ \Phi_{21} & \Phi_{22} \\ 0 & \Phi_{32} \end{bmatrix}, \quad \tilde{S} = S_\infty\left(\begin{bmatrix} \Theta_{11} & \Theta_{12} \\ \Theta_{21} & \Theta_{22} \\ 0 & \Theta_{32} \end{bmatrix}\right).$$

Hence, (35) follows.

Sufficiency. Since conditions (7) and (5) hold, similarly to Lemma 3.3, we have

(44)
$$\operatorname{rank}(E) = \operatorname{rank}(E_{11}) + \operatorname{rank}\begin{bmatrix} E_{22} & E_{23} \\ E_{32} & E_{33} \\ E_{42} & E_{43} \\ 0 & E_{53} \end{bmatrix}, \quad \tilde{n}_6 = n_4,\ E_{64} = 0,\ \det(A_{64}) \neq 0.$$

Hence, by condition (36), we have that $n_2 + n_3 \leq \tilde{n}_3 + n_2 + \tilde{n}_5$, and therefore, as in Theorem 2.4 of [8], we can compute a matrix $X \in \mathbf{R}^{\tilde{n}_3 \times n_2}$ such that

$$\operatorname{rank}\begin{bmatrix} E_{32} + XE_{22} & E_{33} + XE_{23} \\ E_{42} & E_{43} \\ 0 & E_{53} \end{bmatrix} = \operatorname{rank}\begin{bmatrix} E_{22} & E_{23} \\ E_{32} & E_{33} \\ E_{42} & E_{43} \\ 0 & E_{53} \end{bmatrix}.$$

As a consequence, we obtain

$$\operatorname{rank}(E) = \operatorname{rank}(E_{11}) + \operatorname{rank}\begin{bmatrix} E_{32} + XE_{22} & E_{33} + XE_{23} \\ E_{42} & E_{43} \\ 0 & E_{53} \end{bmatrix}. \tag{45}$$

Since conditions (34), (36), and (35) hold, by a slight modification of Lemma 2.5, there exist $\tilde{F}_1 \in \mathbf{R}^{m \times n_1}$ and a nonsingular matrix Z such that

$$\begin{bmatrix} sE_{11} - (A_{11} + B_1\tilde{F}_1) \\ -A_{21} \\ -(A_{31} + XA_{21} + B_3\tilde{F}_1) \end{bmatrix} Z$$
$$= \begin{array}{c} \\ \tilde{n}_1 \\ \tilde{n}_2 \\ \tilde{n}_3 \end{array} \begin{array}{c} \begin{array}{cc} \tilde{n}_1 + \tilde{n}_2 & n_1 - (\tilde{n}_1 + \tilde{n}_2) \end{array} \\ \begin{bmatrix} s\Theta_{11} - \Phi_{11} & -\Phi_{12} \\ s\Theta_{21} - \Phi_{21} & -\Phi_{22} \\ 0 & 0 \end{bmatrix} \end{array}$$

with $([\begin{smallmatrix}\Theta_{11}\\ \Theta_{21}\end{smallmatrix}], [\begin{smallmatrix}\Phi_{11}\\ \Phi_{21}\end{smallmatrix}])$ regular, of index at most one, and stable. By condition (7), we obtain $\forall\ s \in \mathbf{C}^+$

$$\operatorname{rank}\begin{bmatrix} s\tilde{E}_{32} - \tilde{A}_{32} & s\tilde{E}_{33} - \tilde{A}_{33} & s\tilde{E}_{34} - \tilde{A}_{34} & B_3 \\ sE_{42} - A_{42} & sE_{43} - A_{43} & sE_{44} - A_{44} & 0 \\ 0 & sE_{53} - A_{53} & sE_{54} - A_{54} & 0 \\ 0 & 0 & sE_{64} - A_{64} & 0 \end{bmatrix} = n - (\tilde{n}_1 + \tilde{n}_2),$$

where

$$\tilde{E}_{32} = E_{32} + XE_{22}, \quad \tilde{A}_{32} = A_{32} + XA_{22}, \quad \tilde{E}_{33} = E_{33} + XE_{23},$$
$$\tilde{A}_{33} = A_{33} + XA_{23}, \quad \tilde{E}_{34} = E_{34} + XE_{24}, \quad \tilde{A}_{34} = A_{34} + XA_{24}.$$

We also have that B_3, E_{53} are of full row rank, E_{42} and A_{64} are nonsingular, and $E_{64} = 0$. Moreover, condition (45) obviously says that

$$\operatorname{rank}\begin{bmatrix} E_{32} + XE_{22} & E_{33} + XE_{23} \\ E_{42} & E_{43} \\ 0 & E_{53} \end{bmatrix}$$
$$= \operatorname{rank}\begin{bmatrix} E_{32} + XE_{22} & E_{33} + XE_{23} & E_{34} + XE_{24} \\ E_{42} & E_{43} & E_{44} \\ 0 & E_{53} & E_{54} \end{bmatrix}.$$

Thus, by Lemma 2.3(iv), there exists a matrix

$$\begin{matrix} n_1-(\tilde{n}_1+\tilde{n}_2) & n_2 & n_3 & n_4 \end{matrix}$$
$$\begin{bmatrix} \hat{F}_1 & F_2 & F_3 & F_4 \end{bmatrix}$$

such that

$$\left(\begin{bmatrix} 0 & \tilde{E}_{32} & \tilde{E}_{33} & \tilde{E}_{34} \\ 0 & E_{42} & E_{43} & E_{44} \\ 0 & 0 & E_{53} & E_{54} \\ 0 & 0 & 0 & E_{64} \end{bmatrix}, \begin{bmatrix} B_3\hat{F}_1 & \tilde{A}_{32}+B_3F_2 & \tilde{A}_{33}+B_3F_3 & \tilde{A}_{34}+B_3F_4 \\ 0 & A_{42} & A_{43} & A_{44} \\ 0 & 0 & A_{53} & A_{54} \\ 0 & 0 & 0 & A_{64} \end{bmatrix} \right)$$

is regular, of index at most one, and stable.

Let

$$F_1 := \tilde{F}_1 + \begin{bmatrix} 0 & \hat{F}_1 \end{bmatrix} Z^{-1}, \quad F := \begin{bmatrix} F_1 & F_2 & F_3 & F_4 \end{bmatrix} V^T;$$

we have that $(E, A+BF)$ is regular, of index at most one, and stable. Furthermore, if we compute H from $(G_3+XG_2)+B_3H=0$, then we also have disturbance decoupling.

(b) *Necessity.* Let $F, K \in \mathbf{R}^{m\times n}$, and $H \in \mathbf{R}^{m\times d}$ be such that $(E+BK, A+BF)$ is regular, of index at most one, and stable and the disturbances are decoupled. Then conditions (6)–(7) and (37) follow directly Lemma 2.3(vii), Lemma 2.4, and the inequality

$$\operatorname{rank}_g \begin{bmatrix} T_\infty^T(B)(sE-A) & T_\infty^T(B)G \\ C & 0 \end{bmatrix} \le \operatorname{rank}_g \begin{bmatrix} sE-(A+BF) & G+BH \\ C & 0 \end{bmatrix} = n.$$

Partition

$$\begin{matrix} & n_1 & n_2 & n_3 & n_4 \end{matrix} \qquad \begin{matrix} & n_1 & n_2 & n_3 & n_4 \end{matrix}$$
$$FV =: \begin{bmatrix} F_1 & F_2 & F_3 & F_4 \end{bmatrix}, \quad KV =: \begin{bmatrix} K_1 & K_2 & K_3 & K_4 \end{bmatrix}.$$

To prove (34), note that E_{11}, G_2, and B_3 are of full row rank, so there exists an orthogonal matrix $\hat{P}_1$ such that in

$$\begin{matrix} \hat{t}_1 \\ \hat{t}_2 \\ \hat{t}_3 \end{matrix} \begin{bmatrix} s\hat{E}_{11}-\hat{A}_{11} \\ -\hat{A}_{21} \\ -\hat{A}_{31} \end{bmatrix} := \hat{P}_1^T \begin{bmatrix} s(E_{11}+B_1K_1)-A_{11} \\ -A_{21} \\ s(B_3K_1)-A_{31} \end{bmatrix},$$

$$\begin{matrix} \hat{t}_1 \\ \hat{t}_2 \\ \hat{t}_3 \end{matrix} \begin{bmatrix} \hat{B}_1 \\ 0 \\ \hat{B}_3 \end{bmatrix} := \hat{P}_1^T \begin{bmatrix} B_1 \\ 0 \\ B_3 \end{bmatrix}, \quad \begin{matrix} \hat{t}_1 \\ \hat{t}_2 \\ \hat{t}_3 \end{matrix} \begin{bmatrix} \hat{G}_1 \\ \hat{G}_2 \\ \hat{G}_3 \end{bmatrix} := \hat{P}_1^T \begin{bmatrix} G_1 \\ G_2 \\ G_3 \end{bmatrix},$$

$\hat{E}_{11}$, $\hat{G}_2$, and $\hat{B}_3$ have full row rank.

If we set $\hat{P} := \begin{bmatrix} \hat{P}_1 & \\ & I \end{bmatrix} \in \mathbf{R}^{n\times n}$, then we obtain that $\hat{P}^TU(s(E+BK)-A)V$, $\hat{P}^TUB$, $\hat{P}^TUG$, CV are in the condensed form (31). Since there exist $F \in \mathbf{R}^{m\times n}$ and $K \in \mathbf{R}^{m\times p}$ such that $(E+BK, A+BF)$ is regular and stable and the disturbances are decoupled, it follows from part (a) that

$$\operatorname{rank} \begin{bmatrix} sE_{11}-A_{11} & B_1 \\ -A_{21} & 0 \\ -A_{31} & B_3 \end{bmatrix} = \operatorname{rank} \begin{bmatrix} s\hat{E}_{11}-\hat{A}_{11} & \hat{B}_1 \\ -\hat{A}_{21} & 0 \\ -\hat{A}_{31} & \hat{B}_3 \end{bmatrix} = \hat{t}_1+\hat{t}_2+\hat{t}_3 = \tilde{n}_1+\tilde{n}_2+\tilde{n}_3$$

$\forall s \in \mathbf{C}^+$, which is condition (34). Moreover, by part (a) we also have

$$\begin{aligned}\operatorname{rank}\begin{bmatrix} E_{11}+B_1K_1 & A_{11}\hat{S} & B_1 \\ 0 & A_{21}\hat{S} & 0 \\ B_3K_1 & A_{31}\hat{S} & B_3 \end{bmatrix} &= \operatorname{rank}\begin{bmatrix} \hat{E}_{11} & \hat{A}_{11}S_\infty(\hat{E}_{11}) & \hat{B}_1 \\ 0 & \hat{A}_{21}S_\infty(\hat{E}_{11}) & 0 \\ 0 & \hat{A}_{31}S_\infty(\hat{E}_{11}) & \hat{B}_3 \end{bmatrix} \\ &= \hat{t}_1+\hat{t}_2+\hat{t}_3 \\ &= \tilde{n}_1+\tilde{n}_2+\tilde{n}_3\end{aligned}$$

with

$$\hat{S} = S_\infty\left(\begin{bmatrix} E_{11}+B_1K_1 \\ 0 \\ B_3K_1 \end{bmatrix}\right).$$

Thus we have that

$$T_\infty^T\left(\begin{bmatrix} E_{11} & B_1 \\ 0 & 0 \\ 0 & B_3 \end{bmatrix}\right)\begin{bmatrix} A_{11} \\ A_{21} \\ A_{31} \end{bmatrix}\hat{S}$$

is of full row rank and hence W_2 is also.

Sufficiency. Since (6) and (7) hold, it follows by Lemma 2.3(vii) that there exist F_0, K_0 such that $(E+BK_0, A+BF_0)$ is regular and of index at most one. Hence

$$\begin{aligned}\det(U(s(E+BK_0)-(A+BF_0))V) &\neq 0, \\ \deg(\det(U(s(E+BK_0)-(A+BF_0))V)) &= \operatorname{rank}(U(E+BK_0)).\end{aligned}$$

Similarly to Lemma 3.3, a direct calculation yields that

$$\operatorname{rank}\begin{bmatrix} E_{22} & E_{23} \\ E_{42} & E_{43} \\ 0 & E_{53} \end{bmatrix} = \operatorname{rank}\begin{bmatrix} E_{22} & E_{23} & E_{24} \\ E_{42} & E_{43} & E_{44} \\ 0 & E_{53} & E_{54} \end{bmatrix}, \quad \tilde{n}_6 = n_4, \quad E_{64} = 0,$$

$$A_{64} \text{ is nonsingular.} \tag{46}$$

Note that since E_{11}, G_2, and B_3 are of full row rank, there exists an orthogonal matrix $P_1 \in \mathbf{R}^{\nu\times\nu}$ such that

$$\begin{aligned} & P_1^T\begin{bmatrix} sE_{11}-A_{11} & sE_{12}-A_{12} & sE_{13}-A_{13} & sE_{14}-A_{14} & B_1 & G_1 \\ -A_{21} & sE_{22}-A_{22} & sE_{23}-A_{23} & sE_{24}-A_{24} & 0 & G_2 \\ -A_{31} & sE_{32}-A_{32} & sE_{33}-A_{33} & sE_{34}-A_{34} & B_3 & G_3 \end{bmatrix} \\ & \begin{matrix} \\ t_1 \\ = \tilde{n}_2 \\ t_3 \end{matrix}\begin{matrix} \begin{matrix} n_1 & n_2 & n_3 & n_4 & m & d \end{matrix} \\ \begin{bmatrix} s\tilde{E}_{11}-\tilde{A}_{11} & s\tilde{E}_{12}-\tilde{A}_{12} & s\tilde{E}_{13}-\tilde{A}_{13} & s\tilde{E}_{14}-\tilde{A}_{14} & 0 & \tilde{G}_1 \\ A_{21} & sE_{22}-A_{22} & sE_{23}-A_{23} & sE_{24}-A_{24} & 0 & G_2 \\ s\tilde{E}_{31}-\tilde{A}_{31} & s\tilde{E}_{32}-\tilde{A}_{32} & s\tilde{E}_{33}-\tilde{A}_{33} & s\tilde{E}_{34}-\tilde{A}_{34} & \tilde{B}_3 & \tilde{G}_3 \end{bmatrix} \end{matrix},\end{aligned} \tag{47}$$

with $\tilde{E}_{11}$, and $\tilde{B}_3$ full row rank. Then (37) implies that $t_1+\tilde{n}_2 \le n_1$. Furthermore, by (46) we have $n_2+n_3 \le t_3+n_2+\tilde{n}_5$. Note that E_{42} is nonsingular and E_{53} is full row rank; thus, as in the construction given Theorem 2.4 of [8], there exists a matrix $X \in \mathbf{R}^{t_3\times\tilde{n}_2}$ such that

$$\operatorname{rank}\begin{bmatrix} XE_{22} & XE_{23} \\ E_{42} & E_{43} \\ 0 & E_{53} \end{bmatrix} = \operatorname{rank}\begin{bmatrix} E_{22} & E_{23} \\ E_{42} & E_{43} \\ 0 & E_{53} \end{bmatrix}. \tag{48}$$

With this X, by (46) and (48) we have

$$\text{(49)}\quad \operatorname{rank}\begin{bmatrix} \tilde{E}_{11} & \tilde{E}_{12} & \tilde{E}_{13} & \tilde{E}_{14} \\ 0 & E_{22} & E_{23} & E_{24} \\ 0 & XE_{22} & XE_{23} & XE_{24} \\ 0 & E_{42} & E_{43} & E_{44} \\ 0 & 0 & E_{53} & E_{54} \end{bmatrix} = \operatorname{rank}(\tilde{E}_{11}) + \operatorname{rank}\begin{bmatrix} XE_{22} & XE_{23} \\ E_{42} & E_{43} \\ 0 & E_{53} \end{bmatrix}.$$

Determine

$$K := \begin{matrix} n_1 & n_2 & n_3 & n_4 \end{matrix} \\ \begin{bmatrix} K_1 & K_2 & K_3 & K_4 \end{bmatrix}$$

from

$$\tilde{B}_3 \begin{bmatrix} K_1 & K_2 & K_3 & K_4 \end{bmatrix} + \begin{bmatrix} \tilde{E}_{31} & \tilde{E}_{32} & \tilde{E}_{33} & \tilde{E}_{34} \end{bmatrix} = 0;$$

then, since W_2 has full rank if and only if $\operatorname{rank}(\tilde{A}_{11} S_\infty(\tilde{E}_{11})) = \tilde{n}_2$, and since we have already shown that (46) and (49) hold, similarly to the sufficiency in part (a), we obtain the desired feedback matrices F and H. See [10] for details. □

Remark 4.1. If the index one condition is not required, then in (a) necessary and sufficient conditions are given by (7), (34), and (36) and in (b) by (7), (34), and (37). The proof of these conditions and the construction of feedbacks that regularize the system without achieving index at most one are discussed in detail in [10].

Remark 4.2. We can extend these results to systems that include a feedthrough term, which arises frequently in H_∞ and LQG control (see [26]), i.e., systems of the form

$$\begin{aligned} E\dot{x}(t) &= Ax(t) + Bu(t) + Gq(t);\ x(t_0) = x_0, \qquad t \geq t_0, \\ y(t) &= Cx(t) + Du(t), \end{aligned} \tag{50}$$

where E, A, B, G, C are as in (1) and $D \in \mathbf{R}^{p\times m}$.

Using the singular value decomposition of D, e.g., [15],

$$P^T D W = \begin{matrix} r_d \\ q - r_d \end{matrix} \begin{matrix} \begin{matrix} r_d & m - r_d \end{matrix} \\ \begin{bmatrix} D_1 & 0 \\ 0 & 0 \end{bmatrix} \end{matrix}, \tag{51}$$

with P, W orthogonal and D_1 of size $r_d \times r_d$ and nonsingular, we can set

$$\begin{aligned} BW =: & \begin{matrix} r_d & m - r_d \end{matrix} \\ & \begin{bmatrix} B_1 & B_2 \end{bmatrix}, \quad P^T C =: \begin{matrix} r_d \\ q - r_d \end{matrix} \begin{bmatrix} C_1 \\ C_2 \end{bmatrix}, \\ W^T u =: & \begin{bmatrix} u_1 \\ u_2 \end{bmatrix}, \quad P^T y =: \begin{bmatrix} y_1 \\ y_2 \end{bmatrix} \end{aligned} \tag{52}$$

and furthermore for $F \in \mathbf{R}^{m\times n}$

$$\begin{bmatrix} F_1 \\ F_2 \end{bmatrix} := W^T F - \begin{bmatrix} D_1^{-1} C_1 \\ 0 \end{bmatrix}. \tag{53}$$

Then for the solution of the disturbance decoupling problem it suffices to perform the analysis of the augmented system

$$\begin{aligned} \hat{E}\dot{\hat{x}} &= \hat{A}\hat{x} + \hat{B}\hat{u} + \hat{G}q, \\ \hat{y} &= \hat{C}\hat{x}, \end{aligned} \tag{54}$$

where

$$\hat{E} := \begin{bmatrix} E & 0 \\ 0 & 0 \end{bmatrix}, \quad \hat{A} := \begin{bmatrix} A & B_1 \\ C_1 & D_1 \end{bmatrix} \in \mathbf{R}^{\hat{n}\times\hat{n}},$$
$$\hat{B} := \begin{bmatrix} B_2 \\ 0 \end{bmatrix} \in \mathbf{R}^{\hat{n}\times\hat{m}}, \quad \hat{C} := \begin{bmatrix} C_2 & 0 \end{bmatrix} \in \mathbf{R}^{\hat{p}\times\hat{n}}, \quad \hat{G} := \begin{bmatrix} G \\ 0 \end{bmatrix} \in \mathbf{R}^{\hat{n}\times\hat{d}},$$

and $\hat{n} := n+r_d$, $\hat{m} := m-r_d$, $\hat{p} := p-r_d$. For this system the results from the previous section apply, and hence the case of systems with feedthrough can be reduced to the previous results. See [10] for more details.

5. Conclusions. In this paper we have studied the disturbance decoupling problem with or without stability for descriptor systems. We have given necessary and sufficient conditions for solving this problem and at the same time ensuring that the resulting closed-loop system is regular and has index at most one. The proofs are constructive, based on condensed forms that can be computed via orthogonal matrix transformations and can be implemented as numerically stable procedures.

Appendix A.

Constructive proof of Theorem 3.1. In the following algorithm we need row compressions, column compressions, or simultaneous row and column compressions of matrices. Such compressions can be obtained in the usual way via QR-factorizations, rank revealing QR-factorizations, URV-decompositions, or singular value decompositions; see [15, 2].

We also need the computation of GUPTRI forms which can be obtained via the LAPACK routine DGGBAK from LAPACK [2].

ALGORITHM 1.

Input: Matrices E, $A \in \mathbf{R}^{n\times n}$, $B \in \mathbf{R}^{n\times m}$, $G \in \mathbf{R}^{n\times p}$, $C \in \mathbf{R}^{q\times n}$.

Output: Orthogonal matrices U, V, P, W, and the condensed form (14).

Step 1. Perform a row compression and a column compression such that

$$UB =: \begin{bmatrix} B_1 \\ B_2 \\ 0 \end{bmatrix}, \quad UG =: \begin{bmatrix} G_1 \\ 0 \\ 0 \end{bmatrix}, \quad CV =: \begin{bmatrix} 0 & C_3 \end{bmatrix}$$

with G_1, B_2 of full row rank and C_3 of full column rank. Set

$$U(sE-A)V =: \begin{bmatrix} sE_{11}-A_{11} & sE_{13}-A_{13} \\ sE_{21}-A_{21} & sE_{23}-A_{23} \\ sE_{31}-A_{31} & sE_{33}-A_{33} \end{bmatrix}.$$

Step 2. Compute the GUPTRI form of (E_{31}, A_{31}),

$$U_2(sE_{31}-A_{31})V_2 =: \begin{bmatrix} sE_{31}-A_{31} & sE_{32}-A_{32} \\ 0 & sE_{42}-A_{42} \end{bmatrix}$$

with E_{31} of full row rank and $sE_{42} - A_{42}$ of full column rank for any $s \in \mathbf{C}$. Set

$$\begin{bmatrix} sE_{11} - A_{11} \\ sE_{21} - A_{21} \end{bmatrix} V_2 =: \begin{bmatrix} sE_{11} - A_{11} & sE_{12} - A_{12} \\ sE_{21} - A_{21} & sE_{22} - A_{22} \end{bmatrix},$$
$$U_2(sE_{33} - A_{33}) =: \begin{bmatrix} sE_{33} - A_{33} \\ sE_{43} - A_{43} \end{bmatrix}.$$

Step 3. Compute the GUPTRI form of $(\begin{bmatrix} E_{42} & E_{43} \end{bmatrix}, \begin{bmatrix} A_{42} & A_{43} \end{bmatrix})$,

$$U_3 \begin{bmatrix} sE_{42} - A_{42} & sE_{43} - A_{43} \end{bmatrix} V_3 =: \begin{bmatrix} sE_{42} - A_{42} & sE_{43} - A_{43} \\ 0 & sE_{53} - A_{53} \end{bmatrix}$$

with E_{42} of full row rank and $sE_{53} - A_{53}$ of full column rank for any $s \in \mathbf{C}$. Set

$$\begin{bmatrix} sE_{12} - A_{12} & sE_{13} - A_{13} \\ sE_{22} - A_{22} & sE_{23} - A_{23} \\ sE_{32} - A_{32} & sE_{33} - A_{33} \end{bmatrix} := \begin{bmatrix} sE_{12} - A_{12} & sE_{13} - A_{13} \\ sE_{22} - A_{22} & sE_{23} - A_{23} \\ sE_{32} - A_{32} & sE_{33} - A_{33} \end{bmatrix} V_3$$

and $\begin{bmatrix} C_2 & C_3 \end{bmatrix} := \begin{bmatrix} 0 & C_3 \end{bmatrix} V_3$.

Step 4. Perform a row compression

$$U_4 \begin{bmatrix} sE_{21} - A_{21} \\ sE_{31} - A_{31} \end{bmatrix} =: \begin{bmatrix} sE_{21} - A_{21} \\ -A_{31} \end{bmatrix}$$

with E_{21} of full row rank. Set

$$\begin{bmatrix} sE_{22} - A_{22} & sE_{23} - A_{23} \\ sE_{32} - A_{32} & sE_{33} - A_{33} \end{bmatrix} := U_4 \begin{bmatrix} sE_{22} - A_{22} & sE_{23} - A_{23} \\ sE_{32} - A_{32} & sE_{33} - A_{33} \end{bmatrix},$$
$$\begin{bmatrix} B_2 \\ B_3 \end{bmatrix} := U_4 \begin{bmatrix} B_2 \\ 0 \end{bmatrix},$$

and

$$U := \begin{bmatrix} I & & \\ & U_4 & \\ & & I \end{bmatrix} \begin{bmatrix} I & \\ & U_3 \end{bmatrix} \begin{bmatrix} I & \\ & U_2 \end{bmatrix} U_1, \quad V := V_1 \begin{bmatrix} V_2 & \\ & I \end{bmatrix} \begin{bmatrix} I & \\ & V_3 \end{bmatrix}.$$

Appendix B.

Proof of Lemma 3.2. Let $\tilde{U}$ be an orthogonal matrix such that

$$\tilde{U} \begin{bmatrix} B_2 \\ B_3 \end{bmatrix} = \begin{matrix} l_2 \\ \tilde{n}_2 + \tilde{n}_3 - l_2 \end{matrix} \begin{bmatrix} \tilde{B}_2 \\ 0 \end{bmatrix}$$

with $\tilde{B}_2$ of full row rank. Set

$$\tilde{U} \begin{bmatrix} sE_{21} - A_{21} & sE_{22} - A_{22} & sE_{23} - A_{23} \\ -A_{31} & sE_{32} - A_{32} & sE_{33} - A_{33} \end{bmatrix}$$
$$=: \begin{matrix} l_2 \\ \tilde{n}_2 + \tilde{n}_3 - l_2 \end{matrix} \begin{matrix} \begin{matrix} n_1 & \quad n_2 & \quad n_3 \end{matrix} \\ \begin{bmatrix} s\tilde{E}_{21} - \tilde{A}_{21} & s\tilde{E}_{22} - \tilde{A}_{22} & s\tilde{E}_{23} - \tilde{A}_{23} \\ s\tilde{E}_{31} - \tilde{A}_{31} & s\tilde{E}_{32} - \tilde{A}_{32} & s\tilde{E}_{33} - \tilde{A}_{33} \end{bmatrix} \end{matrix}$$

and partition $\tilde{U} = \begin{smallmatrix} l_2 \\ \tilde{n}_2+\tilde{n}_3-l_2 \end{smallmatrix} [\begin{smallmatrix} \tilde{U}_2 \\ \tilde{U}_3 \end{smallmatrix}]$. Then since

$$\begin{bmatrix} sE_{42}-A_{42} & sE_{43}-A_{43} \\ 0 & sE_{53}-A_{53} \\ C_2 & C_3 \end{bmatrix}$$

is of full column rank and G_1, E_{21}, and B_3 are of full row rank $\forall s \in \mathbf{C}$, we have that $\tilde{E}_{31}$ is also of full row rank, and

$$\Pi = \begin{bmatrix} 0 & 0 \\ \tilde{U}_3^T & 0 \\ 0 & I_{\tilde{n}_4+\tilde{n}_5+q} \end{bmatrix} \in \mathbf{R}^{(n+q)\times(n+q-p-l_2)},$$

$$\begin{bmatrix} \Pi^\perp & \Pi \end{bmatrix} = \begin{bmatrix} W_\pi^T & \\ & I_{\tilde{n}_4+\tilde{n}_5+q} \end{bmatrix} \in \mathbf{R}^{(n+q)\times(n+q)},$$

$$\Lambda_r = \begin{bmatrix} I_{n_1} \\ 0 \end{bmatrix} W_r \in \mathbf{R}^{n\times n_1}, \quad \Lambda_l = \begin{bmatrix} I_{\tilde{n}_2+\tilde{n}_3-l_2} \\ 0 \end{bmatrix} \tilde{W}_l^T \in \mathbf{R}^{(n+q-p-l_2)\times(\tilde{n}_2+\tilde{n}_3-l_2)},$$

$$\Lambda_t = \begin{bmatrix} W_\pi^T & \\ & I_{\tilde{n}_4+\tilde{n}_5+q} \end{bmatrix} \begin{bmatrix} I_{p+\tilde{n}_2+\tilde{n}_3} \\ 0 \end{bmatrix} \begin{bmatrix} I_{p+l_2} & \\ & \tilde{W}_l^T \end{bmatrix},$$

where $W_\pi, W_r, \tilde{W}_l$ are orthogonal. If we set $W_l = [\begin{smallmatrix} I_{p+l_2} & \\ & \tilde{W}_l \end{smallmatrix}] W_\pi$, then we obtain

$$\Lambda_1 = W_l \begin{bmatrix} E_{11} \\ E_{21} \\ 0 \end{bmatrix} W_r, \qquad \Lambda_2 = W_l \begin{bmatrix} A_{11} \\ A_{21} \\ A_{31} \end{bmatrix} W_r, \qquad \Lambda_3 = W_l \begin{bmatrix} B_1 \\ B_2 \\ B_3 \end{bmatrix}.$$

Note that

$$\begin{bmatrix} A_{11} \\ A_{21} \\ A_{31} \end{bmatrix} W_r S_\infty \left(\begin{bmatrix} E_{11} \\ E_{21} \\ 0 \end{bmatrix} W_r \right) = \begin{bmatrix} A_{11} \\ A_{21} \\ A_{31} \end{bmatrix} S_\infty \left(\begin{bmatrix} E_{11} \\ E_{21} \\ 0 \end{bmatrix} \right),$$

$$T_\infty^T \left(W_l \begin{bmatrix} E_{11}W_r & B_1 \\ E_{21}W_r & B_2 \\ 0 & B_3 \end{bmatrix} \right) W_l \begin{bmatrix} A_{11} \\ A_{21} \\ A_{31} \end{bmatrix} W_r = T_\infty^T \left(\begin{bmatrix} E_{11} & B_1 \\ E_{21} & B_2 \\ 0 & B_3 \end{bmatrix} \right) \begin{bmatrix} A_{11} \\ A_{21} \\ A_{31} \end{bmatrix} W_r,$$

$$S_\infty \left(T_\infty^T \left(W_l \begin{bmatrix} B_1 \\ B_2 \\ B_3 \end{bmatrix} \right) W_l \begin{bmatrix} E_{11} \\ E_{21} \\ 0 \end{bmatrix} W_r \right) = W_r^T S_\infty \left(T_\infty^T \left(\begin{bmatrix} B_1 \\ B_2 \\ B_3 \end{bmatrix} \right) \begin{bmatrix} E_{11} \\ E_{21} \\ 0 \end{bmatrix} \right).$$

Since B_3 is of full row rank, there exists an orthogonal matrix $\tilde{V}$ such that $\begin{bmatrix} B_3 & A_{31} \end{bmatrix} \tilde{V} = \begin{bmatrix} 0 & \Phi_{32} \end{bmatrix}$ with Φ_{32} nonsingular. Set (with conformal partitioning) $\begin{bmatrix} B_2 & A_{21} \end{bmatrix} \tilde{V} = \begin{bmatrix} \Phi_{21} & \Phi_{22} \end{bmatrix}$, $\begin{bmatrix} 0 & E_{21} \end{bmatrix} \tilde{V} = \begin{bmatrix} \Theta_{21} & \Theta_{22} \end{bmatrix}$, and $\hat{V} := [\begin{smallmatrix} \tilde{V} & \\ & I \end{smallmatrix}] \in \mathbf{R}^{(n+m)\times(n+m)}$. Then, we obtain that

$$\begin{bmatrix} -\Psi B & \Psi(sE-A) \\ 0 & C \end{bmatrix} \hat{V}$$

$$= \begin{bmatrix} s\Theta_{21}-\Phi_{21} & s\Theta_{22}-\Phi_{22} & sE_{22}-A_{22} & sE_{23}-A_{23} \\ 0 & -\Phi_{32} & sE_{32}-A_{32} & sE_{33}-A_{33} \\ 0 & 0 & sE_{42}-A_{42} & sE_{43}-A_{43} \\ 0 & 0 & 0 & sE_{53}-A_{53} \\ 0 & 0 & C_2 & C_3 \end{bmatrix}.$$

Since E_{21}, B_2 are of full row rank and Φ_{32} is nonsingular, it follows that Θ_{21} is also of full row rank. Note that

$$\begin{bmatrix} -\Phi_{32} & sE_{42} - A_{42} & sE_{43} - A_{43} \\ 0 & 0 & sE_{53} - A_{53} \\ 0 & C_2 & C_3 \end{bmatrix}$$

is of full column rank $\forall\ s \in \mathbf{C}$. Hence,

$$\operatorname{rank}(\Lambda_4) = \operatorname{rank} \begin{bmatrix} 0 & E_{32} & E_{33} \\ 0 & E_{42} & E_{43} \\ 0 & 0 & E_{53} \\ 0 & 0 & 0 \end{bmatrix} = \operatorname{rank} \begin{bmatrix} E_{32} & E_{33} \\ E_{42} & E_{43} \\ 0 & E_{53} \end{bmatrix}.$$

Since

$$\begin{bmatrix} \Psi(sB) & \Psi(sE - A) \\ 0 & -C \end{bmatrix} = \begin{bmatrix} sB_2 & sE_{21} - A_{21} & sE_{22} - A_{22} & sE_{23} - A_{23} \\ sB_3 & -A_{31} & sE_{32} - A_{32} & sE_{33} - A_{33} \\ 0 & 0 & sE_{42} - A_{42} & sE_{43} - A_{43} \\ 0 & 0 & 0 & sE_{53} - A_{53} \\ 0 & 0 & -C_2 & -C_3 \end{bmatrix},$$

$\left[\begin{smallmatrix} B_2 & E_{21} \\ B_3 & 0 \end{smallmatrix}\right]$ are of full row rank and

$$\begin{bmatrix} sE_{42} - A_{42} & sE_{43} - A_{43} \\ 0 & sE_{53} - A_{53} \\ -C_2 & -C_3 \end{bmatrix}$$

is of full column rank $\forall s \in \mathbf{C}$. □

REFERENCES

[1] A. Ailon, *A solution to the disturbance decoupling problem in singular systems via analogy with state-space systems*, Automatica J. IFAC, 29 (1993), pp. 1541–1545.

[2] E. Anderson, Z. Bai, C. Bischof, J. Demmel, J. Dongarra, J. Du Croz, A. Greenbaum, S. Hammarling, A. McKenney, S. Ostrouchov, and D. Sorensen, *LAPACK Users' Guide*, 2nd ed., SIAM, Philadelphia, 1995.

[3] A. Banaszuk, M. Kociecki, and K. A. Przyluski, *The disturbance decoupling problem for implicit linear discrete-time systems*, SIAM J. Control Optim., 28 (1990), pp. 1270–1293.

[4] K. Brenan, S. Campbell, and L. Petzold, *Numerical Solution of Initial-Value Problems in Differential Algebraic Equations*, North–Holland, New York, 1989.

[5] A. Bunse-Gerstner, V. Mehrmann, and N. K. Nichols, *Regularization of descriptor systems by derivative and proportional state feedback*, SIAM J. Matrix Anal. Appl., 13 (1992), pp. 46–67.

[6] A. Bunse-Gerstner, V. Mehrmann, and N. Nichols, *Regularization of descriptor systems by output feedback*, IEEE Trans. Automat. Control, 39 (1994), pp. 1742–1747.

[7] R. Byers, T. Geerts, and V. Mehrmann, *Descriptor systems without controllability at infinity*, SIAM J. Control Optim., 35 (1997), pp. 462–479.

[8] D. L. Chu, H. C. Chan, and D. W. C. Ho, *Regularization of singular systems by derivative and proportional output feedback*, SIAM J. Matrix Anal. Appl., 19 (1998), pp. 21–38.

[9] D. Chu and V. Mehrmann, *Disturbance Decoupling for Descriptor Systems*, Tech. Rep. 97-7, Fakultät für Mathematik, TU Chemnitz, Germany, 1997.

[10] D. Chu and V. Mehrmann, *Disturbance Decoupling for Descriptor Systems* II, Tech. Rep. 97-21, Fakultät für Mathematik, TU Chemnitz, Germany, 1997.

[11] L. Dai, *Singular Control Systems*, Lecture Notes in Control and Inform. Sci. 118, Springer-Verlag, Berlin, 1989.

[12] J. Demmel and B. Kågström, *The generalized Schur decomposition of an arbitrary pencil $A-\lambda B$: Robust software with error bounds and applications. Part* I: *Theory and algorithms*, ACM Trans. Math. Software, 19 (1993), pp. 160–174.
[13] P. Van Dooren, *The generalized eigenstructure problem in linear system theory*, IEEE Trans. Automat. Control, 26 (1981), pp. 111–129.
[14] F. Gantmacher, *Theory of Matrices*, Vol. I, Chelsea, New York, 1959.
[15] G. Golub and C. V. Loan, *Matrix Computations*, 3rd ed., The Johns Hopkins University Press, Baltimore, MD, 1996.
[16] S. Hagel, *Systematische Modellbildung mechanischer Systeme auf der Basis von Teilkomponenten*, in Notes of the Workshop Identifizierung-, Analyse und Entwurfsmethoden für mechanische Mehrkörpersysteme in Deskriptorform, Paderborn, P. C. Müller, ed., 1994, pp. 67–72.
[17] H. Knobloch and H. Kwakernaak, *Lineare Kontrolltheorie*, Springer-Verlag, Berlin, 1985.
[18] P. Kunkel and V. Mehrmann, *The linear quadratic optimal control problem for linear descriptor systems with variable coefficients*, Math. Control Signals Systems, 10 (1997), pp. 247–264.
[19] G. Lebret, *Structural solution of the disturbance decoupling problem for implicit linear discrete-time systems*, in Proceedings of the Second International Symposium on Implicit and Robust Systems, Warsaw, 1991, pp. 127–130.
[20] G. Lebret, *Structural solution of the disturbance decoupling problem for implicit linear discrete-time systems*, Circuits Systems Signal Process., 13 (1994), pp. 311–327.
[21] L. R. Fletcher and A. Asaraai, *On disturbance decoupling in descriptor systems*, SIAM J. Control Optim., 27 (1989), pp. 1319–1332.
[22] V. Mehrmann, *The Autonomous Linear Quadratic Control Problem: Theory and Numerical Algorithms*, Lecture Notes in Control and Inform. Sci. 163, Springer-Verlag, Heidelberg, 1991.
[23] A. S. Morse and W. M. Wonham, *Decoupling and pole assignment by dynamic compensation*, SIAM J. Control, 8 (1970), pp. 317–337.
[24] M. Rakowski, *Transfer function approach to disturbance decoupling problem*, in Linear Algebra for Control Theory, B. W. P. Van Dooren, ed., IMA Vol. Math. Appl. 62, Springer-Verlag, 1994, pp. 159–176.
[25] J. M. Schumacher, *Compensator synthesis using (C, A, B)-pairs*, IEEE Trans. Automat. Control, 25 (1980), pp. 1133–1138.
[26] A. A. Stoorvogel and J. W. V. der Woude, *The disturbance decoupling problem with measurement feedback and stability for systems with direct feedthrough matrices*, Systems Control Lett., 17 (1991), pp. 217–226.
[27] V. L. Syrmos, *Disturbance decoupling using constrained Sylvester equations*, IEEE Trans. Automat. Control, 39 (1994), pp. 797–803.
[28] A. Varga, *On stabilization methods of descriptor systems*, Systems Control Lett., 24 (1995), pp. 133–138.
[29] J. C. Willems and C. Commault, *Disturbance decoupling by measurement feedback with stability or pole placement*, SIAM J. Control Optim., 19 (1981), pp. 490–504.
[30] W. Wonham, *Linear Multivariable Control: A Geometric Approach*, 2nd ed., Springer-Verlag, New York, 1979.

SIAM J. CONTROL OPTIM.
Vol. 38, No. 6, pp. 1859–1873

(C, A)-INVARIANCE OF MODULES OVER PRINCIPAL IDEAL DOMAINS*

NAOHARU ITO†, WILAND SCHMALE‡, AND HARALD K. WIMMER§

Abstract. For discrete-time linear systems over a principal ideal domain, three types of (C, A)-invariance can be distinguished. Connections between these notions are investigated. For pure submodules, necessary and sufficient conditions for dynamic (C, A)-injection invariance are given. Sufficient conditions are obtained in the general case.

Key words. systems over rings, (C, A)-invariance, injection-invariance, (A, B)-invariance, feedback invariance, pure submodules, closure of submodules

AMS subject classifications. 93B07, 93B99, 15A33, 13C99

PII. S0363012999355083

1. Introduction.

Consider a real n-dimensional linear system

$$x(t+1) = Ax(t) + Bu(t),\ y(t) = Cx(t), \tag{1.1}$$

with a dynamical observer of the form

$$z(t+1) = Az(t) + Bu(t) - K[y(t) - Cz(t)].$$

The error $e(t) = x(t) - z(t)$ satisfies $e(t+1) = (A+KC)e(t)$. Hence, if ϕ is a subspace of $\mathbb{R}^n$ which is invariant under $A + KC$, then the error $e(t)$ will remain in ϕ provided that the initial error $e(0)$ is in ϕ. Such (C, A)-invariant subspaces, which are basic tools for the construction of observers, have been studied in [3], [28], [25], [8], [11]. The differentiable structure of sets of (C, A)-invariant subspaces was investigated in [10]. Classification and parametrization results were obtained in [16]. Applications to disturbance decoupling by observation feedback and disturbance decoupled estimation can be found in [25].

If (1.1) is a system over a field, then the following important fact is well known. Given ϕ, there exists a K such that $(A + KC)\phi \subseteq \phi$ if and only if

$$A(\phi \cap \operatorname{Ker} C) \subseteq \phi. \tag{1.2}$$

For systems over rings, Example 1.2 below shows that the equivalence between $(A + KC)$-invariance and the geometric property (1.2) is not valid in general, which gives rise to distinct notions of (C, A)-invariance. It is indispensable for prospective applications of systems over rings to clarify the interdependence of those invariance concepts. If a real linear system depends polynomially on a parameter or if it has

*Received by the editors April 21, 1999; accepted for publication (in revised form) January 20, 2000; published electronically July 11, 2000.
http://www.siam.org/journals/sicon/38-6/35508.html

†Department of Mathematical Engineering and Information Physics, University of Tokyo, 7-3-1 Hongo, Bunkyo-ku, Tokyo, 113-8656 Japan (naoharu@nara-edu.ac.jp). The research of this author was supported by the Japanese Society for the Promotion of Sciences.

‡Fachbereich 6 Mathematik, Carl-von-Ossietzky-Universität, D-26111 Oldenburg, Germany (wiland.schmale@uni-oldenburg.de).

§Mathematisches Institut, Universität Würzburg, Am Hubland, D-97074 Würzburg, Germany (wimmer@mathematik.uni-wuerzburg.de). The research of this author was supported by the Deutsche Forschungsgemeinschaft (grant 446 JAP-113/162/0).

commensurable delays, then it can be viewed as a system over the ring $\mathbb{R}[s]$ or more generally as a system over a principal ideal domain (PID) (see, e.g., [26], [23], [12]). Accordingly in this paper the underlying ring R will be a PID. We refer to [7] for linear systems over PIDs and to [9], [17], [2] for applications. In a subsequent paper [19] we apply our present results to disturbance decoupling by dynamic measurement feedback. An example of an engineering application is the decoupling control of a ship propelled by a pitch propeller. The problem was studied by Kono et al. [21]. It was solved in [17] by applying results on diagonal decoupling to delay systems.

Let (C, A) and (A, B) be pairs of matrices with $A \in R^{n\times n}$, $C \in R^{l\times n}$, $B \in R^{n\times m}$. In this paper, ϕ will always be a submodule of R^n with $\operatorname{rank}\phi = r$ and $\operatorname{rank}(\phi \cap \operatorname{Ker} C) = k$.

DEFINITION 1.1.

(i) *A submodule ϕ is called (C, A)-invariant if*

$$A(\phi \cap \operatorname{Ker} C) \subseteq \phi.$$

(ii) *We say ϕ is (C, A)-injection-invariant if there exists a matrix $K \in R^{n\times l}$ such that*

$$(A + KC)\phi \subseteq \phi. \tag{1.3}$$

It is well known that (1.3) implies (1.2). If R is not a field, then the converse need not be true. The following example can be found in [8].

Example 1.2. Take $R = \mathbb{Q}[\tau]$, $n = 2$, and define

$$A = \begin{pmatrix} 1 & 0 \\ 1 & 1 \end{pmatrix}, \quad C = (\tau \quad 0), \quad \phi = \operatorname{Im}\begin{pmatrix} 1 \\ 0 \end{pmatrix}. \tag{1.4}$$

Clearly ϕ is (C, A)-invariant because of $\phi \cap \operatorname{Ker} C = 0$. If $K = (\alpha\ \beta)^T$ is a matrix such that (1.3) holds, then $1 + \beta\tau = 0$, which is possible for $\beta \in R[1/\tau]$ but not for $\beta \in R$. We point out that the submodule $C\phi = \tau R$ does not satisfy $(\phi^\perp)^\perp = \phi$ such that the example later will confirm Corollary 5.2.

The next example shows that injection-invariance can be achieved by extending the data A, C, ϕ in a way which corresponds to dynamic feedback.

Example 1.3. Define A, C, ϕ as in (1.4). Set

$$A^e = \left(\begin{array}{cc|c} 1 & 0 & 0 \\ 1 & 1 & 0 \\ \hline 0 & 0 & 0 \end{array}\right), \quad C^e = \left(\begin{array}{cc|c} \tau & 0 & 0 \\ \hline 0 & 0 & 1 \end{array}\right), \quad \phi^e = \operatorname{Im} M^e, \quad M^e = \left(\begin{array}{c|c} 1 & 0 \\ 0 & 1 \\ \hline 0 & \tau \end{array}\right).$$

Define

$$K^e = \begin{pmatrix} 0 & 0 \\ 0 & 0 \\ 1 & 1 \end{pmatrix}.$$

Then

$$(A^e + K^eC^e)M^e = M^e \begin{pmatrix} 1 & 0 \\ 1 & 1 \end{pmatrix},$$

i.e., $(A^e + K^eC^e)\phi^e \subseteq \phi^e$, and ϕ^e extends ϕ in the sense of $\phi^e \cap (R^2 \oplus 0) = \phi \oplus 0$.

With regard to the preceding example and in particular motivated by a dual concept in the theory of (A, B)-invariant submodules (to be discussed in section 2) we introduce another invariance notion. Let q be a nonnegative integer, and let 0^q be the zero submodule of R^q. We consider extensions of order q of A and C of the form

$$A_q^e = \begin{pmatrix} A & 0 \\ 0 & 0_{q\times q} \end{pmatrix}, \quad C_q^e = \begin{pmatrix} C & 0 \\ 0 & I_q \end{pmatrix}. \tag{1.5}$$

To a submodule ϕ we associate the sets

$$E(\phi, q) = \{\psi \mid \psi \text{ submodule of } R^{n+q},\ \psi \cap (R^n \oplus 0^q) = \phi \oplus 0^q\},$$

and

$$E(\phi) = \cup\{E(\phi, q),\ q \geq 0\}.$$

This type of module extensions appears already in [25].

DEFINITION 1.4. *We call ϕ dynamically (C, A)-injection-invariant of order $q \geq 0$ if there exist a matrix $K^e \in R^{(n+q)\times(l+q)}$ and a submodule $\phi^e \in E(\phi, q)$ such that*

$$(A_q^e + K^e C_q^e)\phi^e \subseteq \phi^e. \tag{1.6}$$

For a pair (C, A) and a submodule ϕ we now have three notions of invariance. It is the purpose of this paper to study the relations between those three concepts. In section 5 we shall prove the following result.

PROPOSITION 1.5. *If ϕ is dynamically (C, A)-injection-invariant, then ϕ is (C, A)-invariant.*

The theory of (A, B)-invariant submodules in section 2 might suggest that a converse of Proposition 1.5 should be true, at least for those submodules ϕ that are pure, i.e., which are direct summands of R^n, or equivalently, which satisfy $(\phi^\perp)^\perp = \phi$. As usual we define $\phi^\perp = \{x \in R^n \mid x^T y = 0 \text{ for all } y \in R^n\}$, and we call $\mathcal{S}\phi = (\phi^\perp)^\perp$ the closure of ϕ. Because of $\mathcal{S}\phi = \phi$, or equivalent closure properties, it has also been common (see, e.g., [9], [22]) to speak of closed instead of pure submodules. Basic properties of the closure operator $\mathcal{S}$ and of pure submodules will be reviewed in the appendix. For our purposes it is important to know that (dynamic) (C, A)-injection-invariance of ϕ is inherited by $\mathcal{S}\phi$. Our main result for pure submodules is the following.

THEOREM 1.6. *Let ϕ be a pure submodule of R^n of rank r. Then ϕ is dynamically (C, A)-injection-invariant if and only if $\phi^\perp$ contains a submodule ψ satisfying*

(i) $\operatorname{rank} \psi = n - r$,

(ii) $A^T\psi \subseteq \psi + \operatorname{Im}(C^T)$.

Note that condition (ii) means that ψ is (A^T, C^T)-invariant according to Definition 2.1 below. Hence the two conditions (i) and (ii) hold if and only if the unique maximal (A^T, C^T)-invariant submodule of $\phi^\perp$ has rank $n - r$.

Having Theorem 1.6 at our disposal we claim that (C, A)-invariance is not sufficient for dynamical (C, A)-injection-invariance.

Example 1.7. Take $R = \mathbb{Q}[\tau]$, $n = 2$, $r = 1$, and

$$A = \begin{pmatrix} 0 & 0 \\ 1 & 0 \end{pmatrix}, \quad C = (\tau \quad 1), \quad \phi = \operatorname{Im} \begin{pmatrix} 1 \\ 0 \end{pmatrix}.$$

Again we have $\phi \cap \operatorname{Ker} C = 0$. Hence ϕ is (C, A)-invariant. It requires a considerable effort to verify directly that there does not exist an extension $\phi^e \in E(\phi)$ and a matrix

K^e such that $(A^e + K^e C^e)\phi^e \subseteq \phi^e$. On the other hand, it follows immediately from Theorem 1.6 that ϕ cannot be dynamically (C, A)-injection-invariant, since in $\phi^\perp = \mathrm{Im}(0, 1)^T$ only the zero submodule is (A^T, C^T)-invariant.

We continue our paper as follows. In section 2 we review types of (A, B)-invariance that correspond to the notions in Definitions 1.1 and 1.4. It will become evident that our results with respect to (C, A) are not dual ones of (A, B)-invariance results. In section 3 we give the proof of Theorem 1.6 and obtain further criteria for pure submodules. In section 4 we work with bases of submodules and develop matrix representations which will then be exploited in section 5 to derive sufficient conditions for (dynamic) (C, A)-injection-invariance in the general nonpure case.

2. (A, B)-invariance. If R is a field, then the (C, A)-invariance condition (1.2) is equivalent to a dual (A^T, C^T)-invariance condition for $\phi^\perp$. That is, we have

$$A(\phi \cap \ker C) \subseteq \phi \iff A^T \phi^\perp \subseteq \phi^\perp + \mathrm{Im}\, C^T. \tag{2.1}$$

Over a PID, (2.1) in general is no longer true. We have only

$$A\,\mathcal{S}(\phi \cap \ker C) \subseteq \mathcal{S}\,\phi \iff A^T \phi^\perp \subseteq \mathcal{S}\left[\phi^\perp + \mathcal{S}(\mathrm{Im}\, C^T)\right]. \tag{2.2}$$

Therefore, in general, results on (A, B)-invariance cannot be carried over to dual ones on (C, A)-invariance. In the special case where ϕ, $\mathrm{Im}\, C^T$, and $\phi^\perp + \mathrm{Im}\, C^T$ are pure submodules the equivalence (2.2) implies (2.1).

For systems over rings dualization, in particular duality between controllability (or reachability) and observability has been a fundamental issue (see, e.g., [5], [6], and [15]). It is known [15] that among PIDs only fields have the property that controllability of (A, B) is always equivalent to observability of the corresponding transposed system. In general, controlled and observed invariance are intrinsically different concepts. Note that (A, B)-invariance (2.3) is preserved under homomorphisms. On the other hand consider Example 1.7 where the homomorphism induced by $\tau \mapsto 0$ destroys (C, A)-invariance.

Let us briefly review some facts and definitions on (A, B)-invariance, for which we refer to [14], [8], and [18]. In accordance with Definition 1.1 we have two concepts, a geometric and a feedback characterization.

DEFINITION 2.1.

(i) *A submodule ψ of R^n is called (A, B)-invariant if*

$$A\psi \subseteq \psi + \mathrm{Im}\, B. \tag{2.3}$$

(ii) *We say ψ is (A, B)-feedback-invariant if there exists a matrix $F \in R^{m\times n}$ such that*

$$(A + BF)\psi \subseteq \psi. \tag{2.4}$$

The notion of dynamic (A, B)-feedback-invariance was introduced in [9] and then also studied in [17] and [18]. For $q \geq 0$ let A_q^e be given as in (1.5). Define

$$B_q^e = \begin{pmatrix} B & 0 \\ 0 & I_q \end{pmatrix}$$

correspondingly. Let $\mathcal{P}$ be the projection of $R^n \oplus R^q$ onto R^n. To a submodule $\psi \subseteq R^n$ we associate the sets

$$G(\psi, q) = \{\mu \mid \mu \,\text{submodule of}\, R^n \oplus R^q \,\text{with}\, \mathcal{P}\mu = \psi\}$$

and $G(\psi) = \cup\{G(\psi, q),\, q \geq 0\}$.

DEFINITION 2.2. *A submodule ψ of R^n is called dynamically (A, B)-feedback-invariant of order q if there exists a $\psi^b \in G(\psi, q)$ and a matrix $F^b \in R^{(n+q)\times(m+q)}$ such that*

$$(A_q^e + B_q^e F^b)\psi^b \subseteq \psi^b.$$

We note that for pure submodules and pure extensions, Definition 1.4 and Definition 2.2 are related by duality. Namely, if $(\phi^\perp)^\perp = \phi$ and $\phi^e \in E(\phi, q)$, then (by Lemma 3.1 below) we have $\mathcal{P}[(\phi^e)^\perp] = \phi^\perp$. Hence $(\phi^e)^\perp \in G(\phi^\perp, q)$. Similarly, if $\psi^b \in G(\psi, q)$, then

$$(\psi^b)^\perp \cap (R^n \oplus 0^q) = \psi^\perp$$

and $(\psi^b)^\perp \in E(\psi^\perp, q)$.

The interdependence between the preceding three types of invariance with respect to (A, B) is well understood. It is clear that over a commutative ring (2.4) implies (2.3). The following example from [14] shows that the converse need not be true.

Example 2.3. Consider $R = \mathbb{Q}[\tau]$, $n = 2$, and

$$A = \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}, \; B = \begin{pmatrix} \tau & 0 \\ 0 & 1 \end{pmatrix}, \; \psi = \operatorname{Im} \begin{pmatrix} 0 \\ \tau \end{pmatrix}.$$

The submodule ψ is (A, B)-invariant but not (A, B)-feedback-invariant.

For pure submodules the two notions of Definition 2.1 coincide.

PROPOSITION 2.4 (see [14]). *A pure submodule of R^n is (A, B)-invariant if and only if it is (A, B)-feedback-invariant.*

For the next result, which is in contrast to Proposition 1.5 and Example 1.7, we refer to [17], [18], and also to [9].

THEOREM 2.5.

(i) *A submodule ψ of R^n is dynamically (A, B)-feedback-invariant if and only if it is (A, B)-invariant.*

(ii) *If ψ is (A, B)-invariant and $\psi = \operatorname{Im} N$, $N \in R^{n\times t}$, $\operatorname{rank} \psi = t$, then there exists an*

$$F^b = \begin{pmatrix} 0_{n\times n} & F_{12} \\ 0_{t\times n} & F_{22} \end{pmatrix} \in R^{(n+t)\times(n+t)}$$

such that

$$(A_t^e + B_t^e F^b) \operatorname{Im} \begin{pmatrix} N \\ I_t \end{pmatrix} \subseteq \operatorname{Im} \begin{pmatrix} N \\ I_t \end{pmatrix}. \tag{2.5}$$

Proof. We include a proof of (ii), since for Theorem 1.6 we shall need a submodule of the form $\psi^b = \operatorname{Im} \binom{N}{I_t}$. Note that (2.5) holds if

$$\begin{pmatrix} AN + BF_{12} \\ F_{22} \end{pmatrix} = \begin{pmatrix} N \\ I_t \end{pmatrix} T \tag{2.6}$$

for some T. Now (2.3) is equivalent to

$$AN = NS_1 + BS_2 \tag{2.7}$$

for some matrices S_1, S_2, over R. Hence we can take $F_{12} = -S_2$ and $F_{22} = S_1$, and (2.7) yields (2.6) with $T = S_1$. ∎

We have indicated before that (C, A)-invariance is preserved under the closure operation $\mathcal{S}$. In general, a corresponding result is not true for (A, B)-invariance, not even for reachable pairs (A, B).

Example 2.6. Take $R = \mathbb{Z}$. If

$$A = \begin{pmatrix} 0 & 0 \\ 1 & 0 \end{pmatrix}, B = \begin{pmatrix} 1 & 0 \\ 0 & 2 \end{pmatrix}, \phi = \operatorname{Im} \begin{pmatrix} 2 \\ 4 \end{pmatrix},$$

then ϕ is (A, B)-invariant but $\mathcal{S}\,\phi$ is not.

Hautus [13] has characterized (A, B)-invariance by a frequency domain condition, which remains valid over commutative rings. For a submodule μ of R^n let

$$\mu[[s^{-1}]] = \left\{ \sum_{j=0}^{\infty} v_j s^{-j}, \quad v_j \in \mu \right\}$$

be the set of formal power series in s^{-1} over μ.

PROPOSITION 2.7. *A submodule ϕ of R^n is (A, B)-invariant if and only if for each $x \in \phi$ there exists a $\xi(s) \in s^{-1}\phi[[s^{-1}]]$ and a $w(s) \in s^{-1}R^m[[s^{-1}]]$ such that*

$$x = (sI - A)\xi(s) - Bw(s). \tag{2.8}$$

Proof. "⇒" Set $\xi_1 = x$. Then we have $A\xi_1 = \xi_2 - Bw_1$ for some $\xi_2 \in \phi$ and $w_1 \in R^m$ and we can define recursively sequences (ξ_j) and (w_j) such that $\xi_j \in \phi$ and $\xi_{j+1} = A\xi_j + Bw_j$, $j \geq 1$. If we set

$$\xi(s) = \xi_1 s^{-1} + \xi_2 s^{-2} + \cdots \quad \text{and} \quad w(s) = w_1 s^{-1} + w_2 s^{-2} + \cdots, \tag{2.9}$$

we obtain (2.8).

"⇐" Now assume (2.8) and (2.9). Then $x = \xi_1$ and $0 = -A\xi_1 + \xi_2 - Bw_1$ yield $Ax \in \phi + \operatorname{Im} B$. ∎

It would be of interest to know whether (C, A)-invariance can be characterized in a similar way. Based on the matrix

$$\begin{pmatrix} sI - A \\ C \end{pmatrix}$$

such a frequency domain description could lead to a computable criterion for (C, A)-invariance.

3. Criteria for pure submodules. Our main objective in this section is the proof of Theorem 1.6. We begin with a technical result.

LEMMA 3.1. *If $\phi^e \in E(\phi)$, then*

$$\mathcal{S}\,\mathcal{P}\big[(\phi^e)^\perp\big] = \phi^\perp. \tag{3.1}$$

Proof. Because of (6.1) we have

$$\mathcal{S}\big[(\phi^e)^\perp + (R^n \oplus 0)^\perp\big] = \big(\phi^e \cap (R^n \oplus 0)\big)^\perp = \phi^\perp \oplus R^q,$$

and we obtain

$$\mathcal{S}\,\mathcal{P}\big[(\phi^e)^\perp\big] = \mathcal{S}\,\mathcal{P}\big[(\phi^e)^\perp + (R^n \oplus 0)^\perp\big] = \mathcal{S}\,\mathcal{P}\big[\mathcal{S}\big((\phi^e)^\perp + (R^n \oplus 0)^\perp\big)\big] = \phi^\perp. \qquad ∎$$

Proof of Theorem 1.6. First assume that ϕ is dynamically (C, A)-injection-invariant. Let $\phi^e \in E(\phi, q)$ and K^e be such that

$$(A^e + K^e C^e)\phi^e \subseteq \phi^e. \tag{3.2}$$

Then $(A^{eT} + C^{eT} K^{eT})\phi^{e\perp} \subseteq \phi^{e\perp}$ implies

$$A^T \mathcal{P}(\phi^{e\perp}) \subseteq \mathcal{P}(\phi^{e\perp}) + \operatorname{Im} C^T.$$

Hence the submodule $\mathcal{P}(\phi^{e\perp})$ is (A^T, C^T)-invariant and because of (3.1) it is of rank $n - r$.

To prove the converse we have a submodule $\hat{\phi}$ of $\phi^\perp$ at our disposal that is (A^T, C^T)-invariant and has rank $n - r$. Then $\hat{\phi} = \operatorname{Im} N$ with $N \in R^{n\times(n-r)}$ and $\operatorname{Ker} N = 0$. Define $\hat{\phi}^b = \operatorname{Im} \binom{N}{I_{n-r}}$ and $\phi^e = (\hat{\phi}^b)^\perp$. Then

$$\phi^e = \operatorname{Im} \begin{pmatrix} I_n \\ -N^T \end{pmatrix} \in E(\phi, n - r).$$

Take $q = n - r$ in (1.5). From Theorem 2.5(ii) we know that there exists a matrix K^b such that

$$(A^{eT} + C^{eT} K^{bT})\hat{\phi}^b \subseteq \hat{\phi}^b.$$

Hence

$$(A^e + K^b C^e)\phi^e \subseteq \phi^e, \quad \phi^e \in E(\phi),$$

which shows that ϕ is dynamically (C, A)-injection-invariant. □

Corollary 3.2. *Let ϕ be a pure submodule of R^n. If*

$$A^T \phi^\perp \subseteq \phi^\perp + \mathcal{S}(\operatorname{Im} C^T), \tag{3.3}$$

or if

$$A \operatorname{Ker} C \subseteq \phi, \tag{3.4}$$

then ϕ is dynamically (C, A)-injection-invariant.

Proof. Let $\lambda \in R$, $\lambda \neq 0$, be such that $\lambda \mathcal{S}(\operatorname{Im} C^T) \subseteq \operatorname{Im} C^T$. Then $\operatorname{rank} \lambda\phi^\perp = n - r$ and (3.3) implies that $\lambda\phi^\perp$ is (A^T, C^T)-invariant. Note that $(\operatorname{Ker} C)^\perp = \mathcal{S}(\operatorname{Im} C^T)$. Hence (3.4) yields $A^T \phi^\perp \subseteq \mathcal{S}(\operatorname{Im} C^T)$, which is a special case of (3.3). □

The following example shows that condition (3.3) is sufficient but not necessary for dynamic (C, A)-injection-invariance.

Example 3.3. Consider $R = \mathbb{Q}[\tau]$ and

$$A = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 1 & 0 \\ 1 & 0 & 0 \end{pmatrix}, C = \begin{pmatrix} \tau & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}, \phi = \operatorname{Im} \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix}.$$

Take $q = 2$ and $\phi^e = \operatorname{Im} M^e$, where

$$M^e = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \\ 0 & \tau & 0 \\ 0 & 0 & 1 \end{pmatrix}.$$

Then $\phi^e \in E(\phi, 2)$. For

$$K^e = \begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & -1 \\ 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & -1 \\ 0 & 0 & 0 & 0 \end{pmatrix}$$

we obtain

$$\operatorname{Im}(A^e + K^e C^e) M^e = \operatorname{Im} \begin{pmatrix} 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & -1 \\ 1 & 0 & 0 & 0 & 0 \\ \tau & 1 & 0 & 0 & -1 \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix} M^e = \operatorname{Im} \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 1 & 0 & 0 \\ \tau & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix} \subseteq \operatorname{Im} M^e.$$

Hence ϕ is dynamically (C, A)-injection-invariant. As far as condition (3.3) is concerned we note that

$$\phi^\perp = \operatorname{Im} \begin{pmatrix} 0 & 0 \\ 1 & 0 \\ 0 & 1 \end{pmatrix} = \operatorname{Im}(e_2,\, e_3).$$

Now $A^T e_3 = (1, 0, 0)^T \in \phi^\perp + \mathcal{S}(\operatorname{Im} C^T)$ would mean

$$\begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix} = \begin{pmatrix} 0 & 0 & \tau & 0 \\ 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 \end{pmatrix} \begin{pmatrix} a_1 \\ a_2 \\ a_3 \\ a_4 \end{pmatrix}$$

for some $a_j \in R = \mathbb{Q}[\tau]$. But $\tau a_3 = 1$ is impossible in R. Hence ϕ is a submodule which does not satisfy (3.3).

4. Bases and matrices. A nonsingular matrix $U \in R^{n\times n}$ is called unimodular if $U^{-1} \in R^{n\times n}$. Let us say that there exists a basis of $R^{n\times n}$ such that $\phi = \operatorname{Im} M$ if there exists a unimodular U such that $U \operatorname{Im} M = \phi$. We first describe transformations of A, C, ϕ that preserve the three types of (C, A)-invariance.

LEMMA 4.1. *Let $U \in R^{n\times n}$ and $V \in R^{l\times l}$ be unimodular. Set*

$$\tilde{A} = U^{-1} A U,\ \tilde{C} = VCU,\ \tilde{\phi} = U^{-1}\phi. \tag{4.1}$$

The submodule ϕ is dynamically (C, A)-injection-invariant of order q if and only if $\tilde{\phi}$ is dynamically $(\tilde{C}, \tilde{A})$-injection-invariant of the same order.

Proof. Assume $\phi = \operatorname{Im} M$, $M \in R^{n\times r}$, and set $\tilde{M} = U^{-1}M$. Then $\tilde{\phi} = \operatorname{Im} \tilde{M}$. Let $K^e \in R^{(n+q)\times(l+q)}$ and

$$M^e = \begin{pmatrix} M_1^e \\ M_2^e \end{pmatrix} \in R^{(n+q)\times r}$$

be such that $\operatorname{Im}(A^e + K^e C^e) M^e \subseteq \operatorname{Im} M^e$ and $\operatorname{Im} M^e \in E(\phi, q)$, i.e., $\operatorname{Im} M = M_1^e \operatorname{Ker}(M_2^e)$. Now define

$$\tilde{K}^e = \begin{pmatrix} U^{-1} & 0 \\ 0 & I_q \end{pmatrix} K^e \begin{pmatrix} V^{-1} & 0 \\ 0 & I_q \end{pmatrix}$$

and

$$\tilde{M}^e = \begin{pmatrix} \tilde{M}_1^e \\ \tilde{M}_2^e \end{pmatrix} = \begin{pmatrix} U^{-1} & 0 \\ 0 & I_q \end{pmatrix} M^e.$$

Then $\operatorname{Im}(\tilde{A}^e + \tilde{K}^e\tilde{C}^e)\tilde{M}^e \subseteq \operatorname{Im} \tilde{M}^e$ and $\operatorname{Im} \tilde{M} = \tilde{M}_1^e \operatorname{Ker}(\tilde{M}_2^e)$. □

It is obvious that a corresponding statement is true for (C, A)-invariance. It will be convenient to choose a suitable basis of R^n such that conditon (1.2) can be expressed in terms of blocks of a corresponding matrix representation of A and C.

LEMMA 4.2. *There exists a basis of R^n such that*

$$\phi \cap \operatorname{Ker} C = \operatorname{Im} \begin{pmatrix} D_1 \\ 0_{(r-k)\times k} \\ 0_{(n-r)\times k} \end{pmatrix}, \quad D_1 \in R^{k\times k}, \det D_1 \neq 0, \tag{4.2}$$

and

$$\phi = \operatorname{Im} \begin{pmatrix} D_1 & D_{12} \\ 0 & D_2 \\ 0 & 0 \end{pmatrix}, \quad D_2 \in R^{(r-k)\times(r-k)}, \det D_2 \neq 0. \tag{4.3}$$

If C is partitionend according to (4.3), *then*

$$C = (0 \quad C_2 \quad C_3), \quad C_2 \in R^{l\times(r-k)}, \quad \operatorname{Ker} C_2 = 0, \tag{4.4}$$

and

$$VC_2 = \begin{pmatrix} \tilde{C}_2 \\ 0 \end{pmatrix}, \ \tilde{C}_2 \in R^{(r-k)\times(r-k)}, \ \det \tilde{C}_2 \neq 0 \tag{4.5}$$

for some unimodular V.

A basis of R^n can be chosen in such a way that

$$\mathcal{S}\,\phi = \operatorname{Im} \begin{pmatrix} I_k & 0 \\ 0 & I_{r-k} \\ 0 & 0 \end{pmatrix}, \tag{4.6}$$

and

$$\mathcal{S}\,\phi \cap \operatorname{Ker} C = \operatorname{Im} \begin{pmatrix} I_k \\ 0 \\ 0 \end{pmatrix}. \tag{4.7}$$

If $A(\phi \cap \operatorname{Ker} C) \subseteq \phi$, then

$$A = \begin{pmatrix} A_{11} & A_{12} & A_{13} \\ 0 & A_{22} & A_{23} \end{pmatrix}, \ A_{11} \in R^{r\times k}, \tag{4.8}$$

and

$$\operatorname{Im} A_{11}D_1 \subseteq \operatorname{Im} \begin{pmatrix} D_1 & D_{12} \\ 0 & D_2 \end{pmatrix}. \tag{4.9}$$

If $\phi = \mathcal{S}\,\phi$, then (4.9) *is trivially satisfied.*

Proof. Since $\mathrm{Ker}C$ is a pure submodule there is a basis of R^n such that $\mathrm{Ker}C = \mathrm{Im}\binom{I_p}{0}$. We can modify this basis in such a way that $\phi = \mathrm{Im}\, M$ with

$$M = \begin{pmatrix} M_1 \\ M_2 \\ 0 \end{pmatrix} \in R^{n\times r},\ M_1 \in R^{p\times r},\ M_2 \in R^{t\times r},$$

where M_2 has full row rank. In particular, we can choose M_2 such that $M_2 = (0 \quad M_{22})$, $M_{22} \in R^{t\times t}$, $\det M_{22} \neq 0$. Hence

$$M = \begin{pmatrix} M_{11} & M_{12} \\ 0 & M_{22} \\ 0 & 0 \end{pmatrix},$$

and we obtain

$$\phi \cap \mathrm{Ker}C = \left\{ M\begin{pmatrix} x \\ y \end{pmatrix},\ M_{22}y = 0 \right\} = \left\{ M\begin{pmatrix} x \\ y \end{pmatrix},\ y = 0 \right\} = \mathrm{Im} \begin{pmatrix} M_{11} \\ 0 \\ 0 \end{pmatrix}.$$

Since M_{11} has full column rank we have

$$\text{column rank}(M_{11}) = k = \mathrm{rank}(\phi \cap \mathrm{Ker}C).$$

A suitable change of basis of R^n yields $M_{11} = \binom{D_1}{0}$ with D_1 as in (4.2), which implies (4.3). The remaining parts of the lemma are easy to verify. □

We now characterize extensions of submodules in terms of matrices. Let $\phi \subseteq R^n$ be given as

$$\phi = \mathrm{Im}\, M,\ M = \begin{pmatrix} D \\ 0 \end{pmatrix},\ D \in R^{r\times r},\ \det D \neq 0.$$

Consider $\phi^e \in E(\phi, q)$ with

$$\phi^e = \mathrm{Im}\, M^e = \mathrm{Im} \left(\begin{array}{c} M_1 \\ M_2 \\ \hline M_3 \end{array}\right) \in R^{[r+(n-r)+q]\times t}$$

and

$$\mathrm{Ker}\, M^e = 0. \tag{4.10}$$

Then $\phi^e \in E(\phi, q)$ is equivalent to

$$\mathrm{Im} \begin{pmatrix} D \\ 0 \end{pmatrix} = \begin{pmatrix} M_1 \\ M_2 \end{pmatrix} \mathrm{Ker}\, M_3. \tag{4.11}$$

We can assume $M_3 = (0_{\rho\times\rho}, M_{32})$, $\mathrm{Ker}\, M_{32} = 0$. Then (4.11) yields $M_2 = (0_{(n-r)\times\rho}, M_{22})$. If $M_1 = (M_{11}\, M_{12})$ is partitioned accordingly, then (4.11) is equivalent to $\mathrm{Im}\, D = \mathrm{Im}\, M_{11}$. Hence we have $\rho \geq r$. Now $M_{11} \in R^{r\times\rho}$ and (4.10) imply $\rho = r$. We can take $M_{11} = D$ and conclude that $\phi^e \in E(\phi)$ if and only if $\phi^e = \mathrm{Im}\, M^e$ for some M^e of the form

$$M^e = \left(\begin{array}{cc} D & N \\ 0 & P \\ \hline 0 & Q \end{array}\right),\ \mathrm{Ker}\, Q = 0. \tag{4.12}$$

If $\phi = \mathcal{S}\,\phi$, then $M = \binom{I_r}{0}$ and we can choose $N = 0$ in (4.12) such that

$$M^e = \left(\begin{array}{cc} I_r & 0 \\ 0 & P \\ \hline 0 & Q \end{array}\right). \tag{4.13}$$

5. Criteria for nonpure submodules. In this section we shall first give a proof of Proposition 1.5 and then present results on submodules that are not pure.

Proof of Proposition 1.5. Assume $A^e + K^e C^e \in R^{(n+q)\times(n+q)}$ and

$$\operatorname{Im}(A^e + K^e C^e) M^e \subseteq \operatorname{Im} M^e \tag{5.1}$$

for some

$$K^e = \begin{pmatrix} K_{11} & K_{12} \\ K_{21} & K_{22} \end{pmatrix} \in R^{(n+q)\times(l+q)},$$

and $M^e = \binom{P}{Q}$, $Q \in R^{q\times r}$, with $\operatorname{Im} M^e \in E(\phi, q)$. From (5.1) follows

$$(A + K_{11}C)P + K_{12}Q = PT, \tag{5.2}$$

$$K_{21}CP + K_{22}Q = QT \tag{5.3}$$

for some $T \in R^{r\times r}$. We have to show that $y \in \phi \cap \operatorname{Ker} C$ implies $Ay \in \phi$. Because of $\operatorname{Im}\binom{P}{Q} \in E(\phi, q)$ we have $y = Pw$ for some w with $Qw = 0$. Then (5.2), (5.3), and $Cy = 0$ yield $APw = PTw$, $QTw = 0$. Hence $Ay = PTw \in \phi$. □

Examining the preceding proof we find that a slightly more general result is true. Namely, if K^e is a matrix over $R[1/d]$ for some $d \in R$, $d \neq 0$, and if the elements of $A^e + K^e C^e$ are in R, then (1.6) implies (C, A)-invariance of ϕ. We also note that in Example 1.2 a weaker form of injection-invariance can be achieved with an observer matrix K over the ring $\mathbb{Q}[\tau, 1/\tau]$. It can be an advantage to embed R into a larger ring $R[1/d]$, e.g., in order to solve coefficient assignment problems [1]. We shall use that approach below. The following definition will be needed. For a submodule η of R^l with $\operatorname{rank}\eta = t$ and invariant factors $f_t \mid \cdots \mid f_1$, define $\Delta(\eta) = f_1$. Note that the submodule η is pure if and only if $\Delta(\eta) = 1$.

PROPOSITION 5.1. *Set $d = \Delta(C\phi)$. The submodule ϕ is (C, A)-invariant if and only if there exists a matrix $K \in R[1/d]^{n\times l}$ such that $A + KC \in R^{n\times n}$ and $(A + KC)\phi \subseteq \phi$.*

Proof. For the "if" part we refer to the remark on the proof of Proposition 1.5. For the "only if" part, assume ϕ as in Lemma 4.2 such that

$$C = \begin{pmatrix} 0 & \tilde{C}_2 & * \\ 0 & 0 & * \end{pmatrix}, \quad \tilde{C}_2 \in R^{(r-k)\times(r-k)},\ \det \tilde{C}_2 \neq 0.$$

We know that $A(\phi \cap \operatorname{Ker} C) \subseteq \phi$ implies

$$A = \begin{pmatrix} A_{11} & A_{12} & * \\ 0 & A_{22} & * \end{pmatrix}$$

and (4.9). Then

$$C\phi = \operatorname{Im}\begin{pmatrix} \tilde{C}_2 D_2 \\ 0 \end{pmatrix}.$$

Using the Smith form it is easy to see that the elements of $(\tilde{C}_2 D_2)^{-1}$ are in $R[1/d]$. Therefore, it is possible to choose

$$K = \begin{pmatrix} K_{11} & * \\ K_{21} & * \end{pmatrix} \in R[1/d]^{r+(n-r)\times(k+(l-k))}$$

in such a way that

$$A_{11}D_{12} + (A_{12} + K_{11}\tilde{C}_2)D_2 = 0 \quad \text{and} \quad (A_{22} + K_{21}\tilde{C}_2)D_2 = 0.$$

Hence K has the desired properties. □

We know from Proposition 2.4 that for pure submodules (A, B)-invariance is equivalent to (A, B)-feedback invariance. The corresponding result for (C, A)-invariance involves the submodule $C\phi$.

COROLLARY 5.2. *If $C\phi$ is a pure submodule, then ϕ is (C, A)-invariant if and only if ϕ is (C, A)-injection-invariant.*

The next proposition again deals with (C, A)-injection-invariance.

PROPOSITION 5.3. *Let ϕ be a (C, A)-invariant submodule that satisfies*

$$C\,\mathcal{S}(\phi) = \mathcal{S}(C\phi). \tag{5.4}$$

If

$$\phi = (\phi \cap \operatorname{Ker} C) + (\phi \cap \operatorname{Im} C^T), \tag{5.5}$$

in particular if $\phi = \mathcal{S}\,\phi$, then ϕ is (C, A)-injection-invariant.

Proof. Let ϕ, $\mathcal{S}\,\phi$ and C be given by (4.3), (4.6), and (4.4). Because of $C\,\mathcal{S}(\phi) = \operatorname{Im} C_2$ and (5.4) we can assume

$$C_2 = (I_{r-k}\ 0)^T \tag{5.6}$$

in $C = (0\ C_2\ C_3)$. Then

$$\operatorname{Im} C^T = \operatorname{Im} \begin{pmatrix} 0 \\ I_{r-k} & 0 \\ C_3^T \end{pmatrix}$$

and (5.5) imply $\phi = \operatorname{Im} M$ with

$$M = \begin{pmatrix} D_1 & 0 \\ 0 & D_2 \\ 0 & 0 \end{pmatrix}. \tag{5.7}$$

Recall that in the given setting (C, A)-invariance of ϕ is equivalent to (4.8) and (4.9). Because of (5.7) we have (4.9) in the form

$$\operatorname{Im} A_{11}D_1 \subseteq \operatorname{Im} \begin{pmatrix} D_1 & 0 \\ 0 & D_2 \end{pmatrix}. \tag{5.8}$$

Let

$$K = \begin{pmatrix} K_1 \\ K_2 \end{pmatrix}$$

be partitioned conforming to A in (4.8). Then

$$(A + KC)M = \begin{pmatrix} A_{11} & A_{12} + K_1C_2 \\ 0 & A_{22} + K_2C_2 \end{pmatrix} \begin{pmatrix} D_1 & 0 \\ 0 & D_2 \end{pmatrix}.$$

Because of (5.6) we can choose K_1 and K_2 in such a way that $A_{i2} + K_iC_2 = 0$, $i = 1, 2$. Then (5.8) implies $\operatorname{Im}(A + KC)M \subseteq \operatorname{Im} M$, which proves (C, A)-injection-invariance. □

We now extend Theorem 1.6 to submodules ϕ where all invariant factors are equal.

THEOREM 5.4. *Let $\phi \subseteq R^n$ be a submodule of rank r that is isomorphic to $dR \oplus \cdots \oplus dR$. Then ϕ is dynamically (C, A)-injection-invariant if and only if $\phi^\perp$ contains an (A^T, C^T)-invariant submodule of rank $n - r$.*

Proof. "⇐" We can assume

$$\phi = \operatorname{Im} \begin{pmatrix} dI_r \\ 0 \end{pmatrix}.$$

Then $\mathcal{S}\,\phi = \binom{I_r}{0}$. Because of $(\mathcal{S}\,\phi)^\perp = (\phi)^\perp$ we see from Theorem 1.6 that $\mathcal{S}\,\phi$ is dynamically (C, A)-injection-invariant. Let $(\mathcal{S}\,\phi)^e \in E(\phi, q)$ and K^e be such that

$$(A^e + K^eC^e)(\mathcal{S}\,\phi)^e \subseteq (\mathcal{S}\,\phi)^e.$$

By (4.13) we have

$$(\mathcal{S}\,\phi)^e = \operatorname{Im} \begin{pmatrix} I_r & 0 \\ 0 & P \\ 0 & Q \end{pmatrix}, \ \operatorname{Ker} Q = 0,$$

for some P and Q. Hence $\phi^e = d(\mathcal{S}\,\phi)^e \in E(\phi, q)$ and $(A^e + K^eC^e)\phi^e \subseteq \phi^e$.

"⇒" Since $\mathcal{S}\,\phi$ together with ϕ is dynamically (C, A)-injection-invariant we can use Theorem 1.6 again and complete the proof. □

6. Appendix on pure submodules and closure. According to Kaplanski [20] the concept of pure submodules can be attributed to Prüfer (1923). It has been introduced in systems theory by Conte and Perdon [7]. In this section we collect some basic facts on pure submodules and a related closure operator, for which we refer to [4], [27], [7], [22], [24]. The terminology is not uniform. Sometimes pure submodules are called saturated, e.g., in [27] and [24], or closed, e.g., in [9] and [22]. As before we are dealing with a PID R and submodules of $V := R^n$. Let $L(V)$ be the lattice of submodules of V.

DEFINITION 6.1. *For $\phi \in L(V)$ set*

$$\mathcal{S}\,\phi = (\phi^\perp)^\perp.$$

The submodule ϕ is called pure if $\phi = \mathcal{S}\,\phi$.

The mapping $\phi \mapsto \mathcal{S}\,\phi$ is a closure operator on $L(V)$ such that (i) $\phi \subseteq \psi$ implies $\mathcal{S}\,\phi \subseteq \mathcal{S}\,\psi$, (ii) $\phi \subseteq \mathcal{S}\,\phi$, and (iii) $S(S\phi) = S\phi$.

LEMMA 6.2.

(i) *The closure $\mathcal{S}\,\phi$ is the smallest pure submodule of V containing ϕ. We have*

$$\mathcal{S}\,\phi = \{x \in V \mid \alpha x \in \phi \text{ for some } \alpha \in R,\ \alpha \neq 0\}.$$

(ii) *Let the columns of $M \in R^{n\times r}$ be a basis of ϕ. Let $f_r \mid \cdots \mid f_1$ be the invariant factors of ϕ such that*

$$\Sigma = \begin{pmatrix} \operatorname{diag}(f_r, \dots, f_1) \\ 0 \end{pmatrix}$$

is the Smith form of M. If U and W are unimodular and $M = U\Sigma W$, then $\phi = \operatorname{Im} U\Sigma$ and $\mathcal{S}\,\phi = \operatorname{Im} U(I_r\ 0)^T$.

(iii) $\operatorname{rank} \mathcal{S}\,\phi = \operatorname{rank} \phi$.

(iv) $\mathcal{S}(\phi^\perp) = \phi^\perp$.

We include a proof of the identity (6.1) below.

LEMMA 6.3. *For submodules μ and θ we have*

$$\mathcal{S}(\mu) \cap \mathcal{S}(\theta) = \mathcal{S}(\mu \cap \theta).$$

The intersection of pure submodules is pure. Furthermore,

$$\mathcal{S}(\mu^\perp + \theta^\perp) = (\mu \cap \theta)^\perp. \tag{6.1}$$

Proof. For $\pi, \rho \in L(V)$ we have

$$(\pi + \rho)^\perp = \pi^\perp \cap \rho^\perp.$$

Taking $\pi = \mu^\perp$ and $\rho = \theta^\perp$ we obtain

$$(\mu^\perp + \theta^\perp)^\perp = \mathcal{S}(\mu) \cap \mathcal{S}(\theta),$$

which implies

$$\mathcal{S}(\mu^\perp + \theta^\perp) = [\mathcal{S}(\mu) \cap \mathcal{S}(\theta)]^\perp = [\mathcal{S}(\mu \cap \theta)]^\perp = (\mu \cap \theta)^\perp. \qquad \square$$

Over PIDs several different ways to characterize pure submodules are known.

LEMMA 6.4. *For $\phi \in L(V)$ the following statements are equivalent.*

(i) *ϕ is a pure submodule.*

(ii) *ϕ is a direct summand of V.*

(iii) *$\phi \cap \alpha V = \alpha\phi$ for all $\alpha \in R$.*

(iv) *All invariant factors of ϕ are 1.*

(v) *V/ϕ is torsion-free.*

Acknowledgment. We would like to thank D. Flockerzi for a valuable comment.

REFERENCES

[1] J. ASSAN, J. F. LAFAY, AND A. M. PERDON, *Weak feedback cyclizability and coefficient assignment for weakly reachable systems over a principal ideal domain*, in Proceedings of the 37th IEEE Conference on Decision and Control, Tampa, FL, IEEE Computer Society, Los Alamitos, CA, 1998, pp. 3563–3568.

[2] J. ASSAN, J. F. LAFAY, AND A. M. PERDON, *Computation of maximal pre-controllability submodules over a Noetherian ring*, Systems Control Lett., 37 (1999), pp. 153–161.

[3] G. BASILE AND G. MARRO, *Controlled and conditioned invariant subspaces in linear systems theory*, J. Optim. Theory Appl., 3 (1969), pp. 306–315.

[4] N. BOURBAKI, *Éléments de Mathématique*, Livre II, Algèbre, Hermann, Paris, 1964.

[5] J. W. BREWER, J. W. BUNCE, AND F. S. VAN VLECK, *Linear Systems over Commutative Rings*, Lecture Notes in Pure and Appl. Math. 104, Marcel Dekker, New York, 1986.

[6] W. S. Ching and B. F. Wyman, *Duality and the regulator problem for linear systems over commutative rings*, J. Comput. System Sci., 14 (1977), pp. 360–368.
[7] G. Conte and A. M. Perdon, *Systems over principal ideal domains. A polynomial model approach*, SIAM J. Control Optim., 20 (1982), pp. 112–124.
[8] G. Conte and A. M. Perdon, *Problems and results in a geometric approach to the theory of systems over rings*, in Linear Algebra for Control Theory, IMA Vol. Math. Appl. 62, P. Van Dooren and B. Wyman, eds., Springer, New York, 1994, pp. 61–74.
[9] G. Conte and A. M. Perdon, *The disturbance decoupling problem for systems over a ring*, SIAM J. Control Optim., 33 (1995), pp. 750–764.
[10] J. Ferrer, F. Puerta, and X. Puerta, *Differentiable structure of the set of controllable $(A, B)^t$-invariant subspaces*, Linear Algebra Appl., 275–276 (1998), pp. 161–177.
[11] I. Gohberg, M. A. Kaashoek, and F. van Schagen, *Partially specified matrices and operators: Classification, completion, applications*, in Operator Theory Adv. Appl. 79, Birkhäuser, Basel, 1995.
[12] L. C. G. J. M. Habets, *Algebraic and Computational Aspects of Time-Delay Systems*, Ph.D. thesis, Eindhoven University of Technology, Eindhoven, The Netherlands, 1994.
[13] M. L. J. Hautus, *(A, B)-invariant and stabilizability subspaces, a frequency domain description*, Automatica J. IFAC, 16 (1980), pp. 703–707.
[14] M. L. J. Hautus, *Controlled invariance in systems over rings*, in Feedback Control of Linear and Nonlinear Systems, D. Hinrichsen and A. Isidori, eds., Lecture Notes in Control and Inform. Sci. 39, Springer, Berlin, 1982, pp. 107–122.
[15] J. A. Hermida-Alonso and T. Sanchez-Giralda, *On the duality principle for linear dynamical systems over commutative rings*, Linear Algebra Appl., 139 (1990), pp. 175–180.
[16] D. Hinrichsen, H. F. Münzer, and D. Prätzel-Wolters, *Parametrization of (C, A)-invariant subspaces*, Systems Control Lett., 1 (1981/82), pp. 192–199.
[17] N. Ito, *Decoupling Problems for Linear Systems over Commutative Rings*, Ph.D. Dissertation, Faculty of Science and Engineering, Tokyo Denki University, Tokyo, Japan, 1998.
[18] N. Ito and H. Inaba, *Dynamic feedback (A, B)-invariant submodules for linear systems over commutative Noetherian domains*, Linear Algebra Appl., 282 (1998), pp. 123–129.
[19] N. Ito, W. Schmale, and H. K. Wimmer, *Disturbance Decoupling by Dynamic Measurement Feedback for Systems Over a Principal Ideal Domain*, in preparation.
[20] I. Kaplanski, *Infinite Abelian Groups*, University of Michigan Press, Ann Arbor, MI, 1954.
[21] M. Kono, Y. Murakami, T. Morishita, and K. Mishima, *Decoupling control of a ship propelled by a controllable pitch propeller*, Trans. Soc. Instrum. Control Engin., 20 (1984), pp. 254–259 (in Japanese).
[22] B. R. McDonald, *Linear Algebra over Commutative Rings*, Monogr. Textbooks Pure Appl. Math. 87, Marcel Dekker, New York, 1984.
[23] A. S. Morse, *Ring models for delay-differential systems*, Automatica J. IFAC, 12 (1976), pp. 529–531.
[24] G. Scheja and U. Storch, *Lehrbuch der Algebra*, Teil 1, 2, Auflage, Teubner, Stuttgart, 1994.
[25] J. M. Schumacher, *(C, A)-invariant subspaces: Some facts and uses*, Rapport 110, Wiskundig Seminarium, Vrije Universiteit Amsterdam, Amsterdam, 1979.
[26] E. Sontag, *Linear systems over rings: A survey*, Ricerche Automat., 7 (1976), pp. 1–34.
[27] B. Stenström, *Rings of Quotients*, Springer, Berlin, 1975.
[28] W. M. Wonham, *Linear Multivariable Control*, Lecture Notes in Econom. and Math. Systems 101, Springer, Berlin, 1974.

SIAM J. CONTROL OPTIM.
Vol. 38, No. 6, pp. 1874–1908

APPROXIMATION AND LIMIT RESULTS FOR NONLINEAR FILTERS OVER AN INFINITE TIME INTERVAL: PART II, RANDOM SAMPLING ALGORITHMS*

AMARJIT BUDHIRAJA† AND HAROLD J. KUSHNER‡

Abstract. The paper is concerned with approximations to nonlinear filtering problems that are of interest over a very long time interval. Since the optimal filter can rarely be constructed, one needs to compute with numerically feasible approximations. The signal model can be a jump-diffusion, reflected or not. The observations can be taken either in discrete or continuous time. The cost of interest is the pathwise error per unit time over a long time interval. In a previous paper of the authors [A. Budhiraja and H.J. Kushner, *SIAM J. Control Optim.*, 37 (1999), pp. 1946–1979], it was shown, under quite reasonable conditions on the approximating filter and on the signal and noise processes that as time, bandwidth, process and filter approximation, etc. go to their limit in any way at all, the limit of the pathwise average costs per unit time is just what one would get if the approximating processes were replaced by their ideal values and the optimal filter was used. When suitable approximating filters cannot be readily constructed due to excessive computational requirements or to problems associated with a high signal dimension, approximations based on random sampling methods (or, perhaps, combinations of sampling and analytical methods) become attractive, and are the subject of a great deal of attention. The work of the previous paper is extended to a wide class of such algorithms. Under quite broad conditions, covering virtually all the cases considered to date, it is shown that the pathwise average errors converge to the same limit that would be obtained if the optimal filter was used, as time goes to infinity and the approximation parameter goes to its limit in any way at all. All the extensions (e.g., wide bandwidth observation or system driving noise) in [A. Budhiraja and H.J. Kushner, *SIAM J. Control Optim.*, 37 (1999), pp. 1946–1979] hold for our random sampling algorithms as well.

Key words. nonlinear filters, numerical approximations to nonlinear filters, robustness of filters, infinite time filtering, occupation measures, pathwise average errors, random sampling algorithms

AMS subject classifications. 93E11, 60G35

PII. S0363012998349935

1. Introduction.

This paper is an extension of the work in [3], which was concerned with the performance of a wide variety of approximations to optimal nonlinear filters over very long time intervals, where *pathwise average* errors are of interest. Let us first briefly review the motivation for that paper. Suppose that the underlying signal model is a diffusion or jump-diffusion $X(\cdot)$ (reflected or not), or a discrete time Markov chain, with white noise corrupted observations, and the dynamics and/or the observation function are nonlinear. Then, except for some few examples, one cannot construct "finite" or computable optimal filters, and some type of approximation must be used.

A very common approximation method starts by approximating the process $X(\cdot)$ by a simpler process $\tilde{X}^h(\cdot)$ for which the optimal nonlinear filter can be constructed, and then uses that filter but with the observations being those on the actual physical process $X(\cdot)$. For example, $\tilde{X}^h(\cdot)$ might be a discretized (in state and/or in time)

*Received by the editors December 31, 1998; accepted for publication (in revised form) December 21, 1999; published electronically July 11, 2000.
http://www.siam.org/journals/sicon/38-6/34993.html

†Department of Statistics, University of North Carolina, Chapel Hill, NC 27599 (budhiraj@email.unc.edu). The research of this author was supported in part by contracts N00014-96-1-0276 and N00014-96-1-0279 from the Office of Naval Research and NSF grant DMI 9812857.

‡Division of Applied Mathematics, Brown University, Providence, RI 02912 (hjk@dam.brown.edu). The research of this author was supported in part by contract DAAH04-96-1-0075 from the Army Research office and by NSF grant ECS 9703895.

form of $X(\cdot)$. It is such that, as $h \to 0$, $\tilde{X}^h(\cdot)$ converges weakly to $X(\cdot)$. Let $\Pi^h(\cdot)$ denote the actual approximating filter. For each h, it is a measure valued process, and $\Pi^h(\cdot)$ converges weakly to the true conditional distribution process as $h \to 0$, i.e., the computed expectations of any bounded and continuous function converge to the true conditional expectation [25, 23]. However, if the filter is to be used over a very long interval $[0, T]$, the most appropriate errors are often the *pathwise average* (rather than the mathematical expectation) errors per unit time, for whatever definition of "error" that is appropriate. This is the case since we work with only one long path, and the mathematical expectation over all paths might not be a useful indicator of the quality of the approximation. For specificity in this introduction, let us define the average pathwise cost or error on $[0, T]$ to be

$$G^{h,T}(\phi) = \frac{1}{T}\int_0^T f(\phi(X(t)) - \langle \Pi^h(t), \phi\rangle)dt, \tag{1.1}$$

where $\phi(\cdot)$ is a bounded and continuous function and $f(\cdot)$ is an arbitrary continuous function. Our results cover much more general forms of the cost or error function.

Now there are two parameters h and T. The convergence of the filter process $\Pi^h(\cdot)$ over any fixed finite interval says nothing about the behavior of the pathwise average errors as $h \to 0$ and $T \to \infty$ arbitrarily. Under reasonable conditions, it was shown in [3] that the pathwise errors converge in probability to an optimal deterministic limit, and this limit is exactly what one would get for the limit of the mathematical expectation $EG^{h,T}(\phi)$ if the *true optimal filter* was used instead of $\Pi^h(t)$. This is an ideal result. The convergence is independent of how $h \to 0$ or $T \to \infty$. For applications, it is important that h and T be allowed to go to their limits in an arbitrary way.

The reference [3] actually dealt with a much more general setup. The signal process was allowed to be not necessarily a diffusion but a process (possibly driven by wide bandwidth noise), which converges weakly to a diffusion as some parameter converges to its limit (e.g., the noise bandwidth goes to infinity). Wide bandwidth observation noise was also allowed. The case where the observations are taken in discrete time was also covered.

We also note the reference [25], where the pathwise average error was replaced by the expectation of the pathwise average error. In [18], the asymptotics of the filter alone were dealt with, and it was presumed that the filter was the true optimal filter, not an approximation.

We now turn to the description of this paper. In [3], the approximate filter was that for an approximating process $\tilde{X}^h(\cdot)$ but with the actual physical observations on $X(\cdot)$ used. One common and convenient example is the Markov chain approximation method, where the approximating process $\tilde{X}^h(\cdot)$ is a continuous time interpolation of a Markov chain [24, 23]. When the dimension is larger than three or four, such methods can have excessively high computational requirements. Alternatives based on random sampling or Monte Carlo then become attractive, analogous to the case of classical multidimensional integration. The topic is of considerable current interest, e.g., [5, 6, 7, 12, 13, 15, 16, 27, 28, 29]. Such methods are also of interest when the transition probabilities are very hard to compute, e.g., in discrete time problems, where the signal is the output of a system with very complex dynamics, but which can be conveniently simulated. All of the issues in [3] (which were mentioned above) arise here as well, in addition to the potentially serious errors due to the random sampling and the very large time intervals. Owing to the sampling errors as well as to the other (computational and modeling) approximations that are made in the filter

and signal processes, it is conceivable that the long term pathwise average errors per unit time will be large, even with approximations that would perform well over some bounded time interval.

This paper extends the results in [3] to such sampling based algorithms. To distinguish the algorithms in [3] from those of this paper, we refer to the former as *integration* algorithms, since the conditional distributions are computed using integrations or summations over the distributions of the approximating processes. One must keep in mind that random sampling based filters usually require a large number of samples if they are to work well.

Appropriate analogues of the occupation measure methods in [3] are employed for the proofs. Section 2 provides the standard background for the filtering problem in continuous time. Section 3 discusses the formulation of the limit problem for the continuous time case in terms of occupation measures and states a main result from [3] which will be used. Section 4 repeats this for the discrete time problem. In order to effectively exploit the past results, the problem is set up in such a way that the proofs are close to those for the "integration" algorithms of [3]. Thus only the differences in the proofs will be presented. In preparation for this, the structure of the proof in [3] is briefly outlined, and the points where there will be a difference are noted. A fundamental assumption in [3] is the *consistency* assumption which quantifies the convergence of the computational approximating process $\tilde{X}^h(\cdot)$ to $X(\cdot)$ as $h \to 0$. This is A4.1 here. It will be weakened in several ways, depending on the form of the random sampling algorithm.

Section 5 concerns a variety of forms of the sampling algorithms in discrete time. The simplest form is based on a pure random sampling of an approximating process $\tilde{X}^h(\cdot)$. The part of the convergence proof which differs from that in [3] is given. The basic scheme can be generalized in many ways. The standard variance reduction methods such as antithetic variables and stratified sampling can be used. Combinations of integration and sampling methods are often of great use, since it might be most convenient to simulate some parts of the problem but to use "integrations" over distributions of approximating processes in others. See Examples 5.6 and 5.7 in section 5. In order to cover all of the above cases, we reformulate the consistency condition so that it holds for all the above examples and in fact covers quite general algorithms. Section 6 concerns the use of importance sampling methods for the discrete time problem [5, 10, 29]. The standard form is discussed. But the most interesting form is where the measure change depends on the next observation, which is thus used to guide the simulation on the current time interval. Some such algorithms were used in [5, 29], as well as by the authors. The proofs for all of these cases differ only slightly from that given for the basic example in section 5, and thus only the differences are discussed. Finally, in section 7 we study the continuous time analogues of the various random sampling and combined random sampling-integration algorithms studied in sections 5 and 6. We begin by indicating the form of the approximating filter for the case where the random samples are mutually independent and identically distributed (i.i.d.). We then consider a general form of the approximating filter which covers not only the case of such i.i.d. samples but also various variance reduction schemes and importance sampling algorithms of the type studied for discrete time problems in sections 5 and 6.

2. Background: The optimal filter and numerical approximations: Continuous time. The optimal filter is in continuous time. For simplicity and specificity

and until further notice, suppose that the signal process is the $\mathbb{R}^r$ valued diffusion

$$dX = p(X)dt + \sigma(X)dW, \tag{2.1}$$

where $W(\cdot)$ is a standard vector-valued Wiener process and $p(\cdot)$ and $\sigma(\cdot)$ are continuous. We suppose that the solution is unique in the weak sense for each initial condition. Furthermore, suppose throughout that there is a compact set G such that $X(t) \in G$ for all t if $X(0) \in G$, and we always let $X(0) \in G$. All probability measures (random or not) on $\mathbb{R}^r$ considered hereafter will be assumed to have their support contained in G. The observation process is

$$Y(t) = \int_0^t g(X(s))ds + B(t), \tag{2.2}$$

where $g(\cdot)$ is a continuous vector-valued function and $B(\cdot)$ is a standard vector-valued Wiener process, independent of $W(\cdot)$ and $X(0)$.

As pointed out in [3], the approximation and limit results proved there continue to hold, with minor changes in the proofs, for the case where the signal process is a jump-diffusion or is a reflecting diffusion with appropriate conditions on the reflection direction. However, for the sake of simplicity, we confine our work to the model (2.1).

Let $\tilde{X}(\cdot)$ be a process satisfying (2.1), and which (loosely speaking) is conditionally independent of $(X(\cdot), W(\cdot), B(\cdot))$ given its initial condition: We formalize this as follows. $\tilde{X}(\cdot)$ is a process satisfying (2.1) (with $W(\cdot)$ replaced by some other Brownian motion) such that there exists a (possibly random) probability measure Π^* on $\mathbb{R}^r$ with the properties that are conditioned on Π^*, $\tilde{X}(\cdot)$ is independent of $(X(\cdot), W(\cdot), B(\cdot))$, and the conditional distribution of $\tilde{X}(0)$ given Π^* is Π^*. We will call Π^* the "random initial distribution" of $\tilde{X}(\cdot)$ (i.e., the distribution of $\tilde{X}(0)$). It will vary depending on the need and will be specified when needed.

For any process $U(\cdot)$, let $U_{a,b}$, $a \le b$, denote the set $\{U(s), a \le s \le b\}$. Let $E_Z f$ denote the expectation of a function f given the data (or σ-algebra) Z. Until further notice, let $\Pi(0)$ denote the distribution of $X(0)$ and $\Pi(t)$ the distribution of $X(t)$ given the data $Y_{0,t}$ and $\Pi(0)$. Define

$$R(\tilde{X}_{0,t}, Y_{0,t}) = \exp\left[\int_0^t g'(\tilde{X}(s))dY(s) - \frac{1}{2}\int_0^t |g(\tilde{X}(s))|^2 ds\right]. \tag{2.3}$$

Using the representation of the optimal filter $\Pi(\cdot)$ as it was originally developed in [20], for each bounded and measurable real-valued function $\phi(\cdot)$, we can define the evolution of the optimal filter by

$$\int \phi(x)\Pi(t)(dx) \equiv \langle \Pi(t), \phi\rangle = \frac{E_{\{\Pi(0), Y_{0,t}\}}[\phi(\tilde{X}(t))R(\tilde{X}_{0,t}, Y_{0,t})]}{E_{\{\Pi(0), Y_{0,t}\}}R(\tilde{X}_{0,t}, Y_{0,t})}. \tag{2.4}$$

The notation $E_{\{\Pi(0), Y_{0,t}\}}$ denotes the expectation conditioned on the data $Y_{0,t}$ and on $\Pi(0)$ being the initial distribution of $\tilde{X}(\cdot)$ (i.e., $\Pi(0)$ is the current value of what we generically called Π^* above). This representation is convenient for our purposes and is equivalent to the forms used subsequently which were based on measure transformations, as in [8, 14, 26].

The Markov property of $X(\cdot)$ implies that the filter defined by (2.4) satisfies the semigroup relation

$$\langle \Pi(t), \phi\rangle = \frac{E_{\{\Pi(t-s), Y_{t-s,t}\}}[\phi(\tilde{X}(s))R(\tilde{X}_{0,s}, Y_{t-s,t})]}{E_{\{\Pi(t-s), Y_{t-s,t}\}}R(\tilde{X}_{0,s}, Y_{t-s,t})},\ 0 < s \le t. \tag{2.5}$$

In (2.5), $\Pi(t-s)$ is the random initial distribution of $\tilde{X}(\cdot)$. Throughout the paper, we use the notation $E_{\{\Pi(a),Y_{a,b}\}}F(\tilde{X}_{0,s},Y_{a,b})$ for the conditional expectation, given the data $\{Y_{a,b},\Pi(a)\}$ and where the random initial distribution for $\tilde{X}(\cdot)$ is $\Pi(a)$. The analogous notation will be used when approximations to $\tilde{X}(\cdot)$ are used.

An approximating filter. Except for some special cases, $\Pi(t)$ is very hard to compute for nonlinear problems, and thus one must use some type of approximation. Perhaps the most common method of approximation is to approximate the signal process by a simpler form for which a convenient filter can be constructed. Then the approximate filter is obtained by constructing the filter for that approximating signal process, but using the actual physical observations. For example, the approximating filter might be that for a time and space discretization of the process (2.1). The key mathematical ideas behind such approximations and their convergence properties (over finite time intervals) are in [24], in connection with the Markov chain approximation method, a canonical form of this idea.

Let us formalize the above canonical approximation. Let $\tilde{X}^h(\cdot)$ denote the approximating process, which is used to construct the approximating filter, i.e., the approximating filter is constructed as though the true process was $\tilde{X}^h(\cdot)$, but in this filter, we use the actual physical observations defined by (2.2).

Let $\Pi^h(0)$ be an approximation to the true initial distribution of $X(0)$. The $\tilde{X}^h(\cdot)$ might be a Markov process, e.g., a continuous time Markov chain on a finite state space. More commonly, it is an interpolation of a discrete parameter process: i.e., there is $\delta_h > 0$ which goes to zero as $h \to 0$ such that $\tilde{X}^h(\cdot)$ is constant on the intervals $[n\delta_h, n\delta_h + \delta_h)$ and $\tilde{X}^h(n\delta_h), n = 0, \ldots,$ is Markov. When the signal process is defined in continuous time, we always assume that $\tilde{X}^h(\cdot)$ is of one of these two forms. Furthermore, we always suppose (without loss of generality) that $\tilde{X}^h(t)$ takes values in G.

Define

$$R(\tilde{X}^h_{0,t},Y_{0,t}) = \exp\left[\int_0^t g'(\tilde{X}^h(s))dY(s) - \frac{1}{2}\int_0^t |g(\tilde{X}^h(s))|^2 ds\right]. \tag{2.6}$$

For Markov $\tilde{X}^h(\cdot)$, the approximating filter $\Pi^h(\cdot)$ is defined by

$$\langle \Pi^h(t),\phi\rangle = \frac{E_{\{\Pi^h(0),Y_{0,t}\}}[\phi(\tilde{X}^h(t))R(\tilde{X}^h_{0,t},Y_{0,t})]}{E_{\{\Pi^h(0),Y_{0,t}\}}R(\tilde{X}^h_{0,t},Y_{0,t})}, \tag{2.7}$$

and $\Pi^h(\cdot)$ satisfies the semigroup equation

$$\langle \Pi^h(t+s),\phi\rangle = \frac{E_{\{\Pi^h(t),Y_{t,t+s}\}}[\phi(\tilde{X}^h(s))R(\tilde{X}^h_{0,s},Y_{t,t+s})]}{E_{\{\Pi^h(t),Y_{t,t+s}\}}R(\tilde{X}^h_{0,s},Y_{t,t+s})},\ s>0,\ t\geq 0. \tag{2.8}$$

According to our standard notation, the initial distribution of $\tilde{X}^h(\cdot)$ in (2.6) is $\Pi^h(0)$, and it is $\Pi^h(t)$ in (2.8).

When $\tilde{X}^h(\cdot)$ is piecewise constant with $\tilde{X}^h(n\delta)$ being Markov, then the approximating filter is defined by (2.7) and (2.8), but where t and s are integral multiples of δ, and $\Pi^h(\cdot)$ is constant on the intervals $[n\delta, n\delta+\delta)$. Thus the evolution of $\Pi^h(\cdot)$ can be written in recursive form in general. We see that, by Bayes' rule, (2.7) and (2.8) are filters for the $\tilde{X}^h(\cdot)$ process, but with the actual observations $Y_{n\delta,n\delta+\delta}$ used at step n. The conditions for convergence of this Markov chain approximation method are in [21, 24]. The following is the essential condition.

A2.1. *A consistency assumption.* We assume that for any sequence $\{\Pi^h\}$ of probability measures converging weakly to some probability measure Π, $\tilde{X}^h(\cdot)$ with the initial distribution Π^h converges weakly to $\tilde{X}(\cdot)$ with the initial distribution Π.

By the fact that $X(\cdot)$ is a Feller process, A2.1 is equivalent to the following: For any sequence Π^h and any $q(\cdot)$ which is a bounded, continuous, and real-valued function on the Skorohod space $D[G;0,\infty)$ (the space of G-valued functions which are right continuous and have left-hand limits and with the Skorohod topology),

$$E_{\Pi^h} q(\tilde{X}^h(\cdot)) - E_{\Pi^h} q(\tilde{X}(\cdot)) \to 0, \tag{2.9}$$

as $h \to 0$.

3. Occupation measures: Continuous time. We now provide the definitions which are needed for the formulation of the limit and robustness results. The methods are based on occupation measure arguments.

Assumptions and definitions. The measure valued process $\Pi(\cdot)$ is well defined by (2.4) no matter what the initial condition $\Pi(0)$ is, even if it is *not* the distribution of $X(0)$, or if it is random but *independent* of $(W(\cdot), B(\cdot))$. Thus we can speak of the pair $(X(\cdot), \Pi(\cdot))$ as having an arbitrary initial condition. We say that the process $(X(\cdot), \Pi(\cdot))$ is stationary if the distribution of $(X(t+\cdot), \Pi(t+\cdot))$ does not depend on t. From the Feller–Markov property of $X(\cdot)$ and the semigroup relation (2.5), it is easy to show that $(X(\cdot), \Pi(\cdot))$ is a Feller–Markov process. Since G is compact, $\mathcal{M}(G)$ is compact, and so $(X(\cdot), \Pi(\cdot))$ takes values in a compact state space. Thus there exists at least one stationary process. Let $\bar{Q}(\cdot)$ denote the measure of the *joint* process $\Psi(\cdot) = (X(\cdot), \Pi(\cdot), Y(\cdot), B(\cdot), W(\cdot))$, where $(X(\cdot), \Pi(\cdot))$ is stationary. Let $\bar{Q}_f(\cdot)$ denote the measure of the stationary joint process $(X(\cdot), \Pi(\cdot))$.

We make the following key assumption throughout.

A3.1. *A uniqueness assumption.* The process $(X(\cdot), \Pi(\cdot))$ has a unique stationary measure.

The importance of the uniqueness of the stationary joint process was shown in [25]. Some discussion of the uniqueness of $\bar{Q}_f(\cdot)$ is in [3, section 7], where there is also a discussion of the filtering interpretation of the stationary process. See also [1] for other sufficient conditions for A3.1. For each $t \geq 0$, define the shifted process $\Psi_f^h(t,\cdot) = (X(t+\cdot), \Pi^h(t+\cdot))$ and the centered and/or shifted processes

$$\Psi^h(t,\cdot) = (\Psi_f^h(t,\cdot), Y(t+\cdot) - Y(t), B(t+\cdot) - B(t), W(t+\cdot) - W(t)).$$

The path spaces. The vector-valued processes such as $X(\cdot), Y(\cdot), B(\cdot), \tilde{X}(\cdot)$, and so forth will take values in the path space $D[\mathbb{R}^k;0,\infty)$, i.e., in the space of $\mathbb{R}^k$-valued functions which are right continuous and have left-hand limits, with the Skorohod topology [2, 9] for the appropriate value of k.

Let $\mathcal{M}(G)$ denote the space of measures on G with the weak topology. The optimal filter $\Pi(t)$ and its approximations $\Pi^h(t)$ at each time t take values in $\mathcal{M}(G)$. The process $\Pi(\cdot)$ and its approximations will take values in the space $D[\mathcal{M}(G);0,\infty)$, also with the Skorohod topology used.

For a random variable Z and set A, let $I_A(Z)$ denote the indicator function of the event that $Z \in A$. Let C be a measurable set in the product path space of $\Psi^h(t,\cdot)$. Define the *occupation measure* $Q^{h,T}(\cdot)$ by

$$Q^{h,T}(C) = \frac{1}{T}\int_0^T I_C(\Psi^h(t,\cdot))dt. \tag{3.1}$$

In what follows, lower case letters $x(\cdot), \pi(\cdot)$, etc. are used for the canonical sample paths. Letters such as $x, y, \ldots$, are used to denote vectors such as $x(t), y(t)$, etc. Define $\psi_f(\cdot) = (x(\cdot), \pi(\cdot))$, $\Psi_f(\cdot) = (X(\cdot), \Pi(\cdot))$, and $\psi(\cdot) = (x(\cdot), \pi(\cdot), y(\cdot), b(\cdot), w(\cdot))$.

The random measures $Q^{h,T}(\cdot)$ defined by (3.1) take values in the space of measures on the product path space

$$\mathcal{M}(D[\mathbb{R}^k; 0, \infty) \times D[\mathcal{M}(G); 0, \infty))$$

for the appropriate value of k (which is the sum of the dimensions of x, y, b, w).

An error or cost function. Let $F(\cdot)$ be a real-valued function on $D[G; 0, \infty) \times D[\mathcal{M}(G); 0, \infty)$ which is measurable and continuous (with probability 1 (w.p.1)) with respect to the measure $\bar{Q}_f(\cdot)$). Owing to the compactness of G, we can suppose that $F(\cdot)$ is bounded. As in [3], we are concerned with the asymptotic (pathwise) behavior of the sample averages $\int_0^T F(\Psi_f^h(t, \cdot))dt/T$ as $h \to 0$ and $T \to \infty$. Let $Q_f^{h,T}(\cdot)$ denote the $(X(\cdot), \Pi^h(\cdot))-$marginal of $Q^{h,T}(\cdot)$, i.e., for arbitrary measurable set C' in the product path space of $\Psi_f^h(t, \cdot)$,

$$Q_f^{h,T}(C') = \frac{1}{T} \int_0^T I_{C'}(\Psi_f^h(t, \cdot))dt.$$

By the definition of the occupation measure, we can write

$$\frac{1}{T} \int_0^T F(\Psi_f^h(t, \cdot))dt = \int F(\psi_f(\cdot))Q_f^{h,T}(d\psi_f(\cdot)). \tag{3.2}$$

The representation (3.2) shows that the asymptotic values of the left-hand side can be obtained from the limits of the set of occupation measures $Q_f^{h,T}$, as $T \to \infty$ and $h \to 0$.

It was shown in [3] that for a broad class of approximate filters

$$\frac{1}{T} \int_0^T F(\Psi_f^h(t, \cdot))dt \to \int F(\psi_f(\cdot))\bar{Q}_f(d\psi_f(\cdot)) \tag{3.3}$$

in probability as $T \to \infty$ and $h \to 0$ in any way at all. As pointed out in [3], the arbitrariness of the way that $T \to \infty$ and $h \to 0$ is crucial in applications. It was also shown that

$$\frac{1}{T} \int_0^T F(\Psi_f^h(t, \cdot))dt \to \int F(\psi_f(\cdot))\bar{Q}_f(d\psi_f(\cdot)), \tag{3.3$'$}$$

where $\Pi(\cdot)$ in (3.3$'$) is the true optimal filter. Note that via an application of the dominated convergence theorem, we can replace the expressions on the left sides of (3.3) and (3.3$'$) by their expected values.

Let $\phi(\cdot)$ be a bounded, continuous, and real-valued function. A special case of (3.3) is the convergence of the mean square error

$$\begin{aligned} G^{h,T}(\phi) &\equiv \frac{1}{T} \int_0^T [\langle \Pi^h(t), \phi \rangle - \phi(X(t))]^2 dt \\ &\to \int [\langle \pi(0), \phi \rangle - \phi(x(0))]^2 \bar{Q}_f(d\psi_f(\cdot)) \end{aligned} \tag{3.4}$$

in the sense of probability as $h \to 0$ and $T \to \infty$ in any way at all. The right side of (3.4) is what one would also get as the limit if the true optimal filter was used

(and even with an expectation of the pathwise average used) [3]. In this sense, there is pathwise asymptotic optimality of the approximating filter over the infinite time interval.

The following is the main background theorem from [3].

THEOREM 3.1. *Let the filtering model be as in section* 2. *Assume the uniqueness condition* A3.1. *Define the approximate filter* $\Pi^h(\cdot)$ *via* (2.7) *where* $\tilde{X}^h(\cdot)$ *satisfies the consistency condition* A2.1. *Then, for every sequence* $\{h_k, T_k\}_{k\geq 1}$ *such that* $h_k \to 0$ *and* $T_k \to \infty$ *as* $k \to \infty$, *the family* $\{Q^{h_k,T_k}(\cdot); k \geq 1\}$ *is tight. Extract a weakly convergent subsequence with weak sense limit denoted by* $Q(\cdot)$, *a measure-valued random variable. Let* $Q^\omega(\cdot)$ *denote the sample values of* $Q(\cdot)$. $Q^\omega(\cdot)$ *induces a process, denoted by*

$$\Psi^\omega(\cdot) = \{X^\omega(\cdot), \Pi^\omega(\cdot), Y^\omega(\cdot), B^\omega(\cdot), W^\omega(\cdot)\}.$$

Here, the ω *indexes the process, not the sample paths of the process. For almost all* ω *the following hold. The processes* $(B^\omega(\cdot), W^\omega(\cdot))$ *are independent standard Wiener, with respect to which* $\{X^\omega(\cdot), \Pi^\omega(\cdot), Y^\omega(\cdot)\}$ *are nonanticipative.*

$$dY^\omega = g(X^\omega)dt + dB^\omega, \tag{3.5}$$

$$dX^\omega = p(X^\omega)dt + \sigma(X^\omega)dW^\omega. \tag{3.6}$$

For each bounded and measurable real-valued function $\phi(\cdot)$,

$$\langle \Pi^\omega(t), \phi\rangle = \frac{E_{\{\Pi^\omega(0), Y^\omega_{0,t}\}}[\phi(\tilde{X}(t))R(\tilde{X}_{0,t}, Y^\omega_{0,t})]}{E_{\{\Pi^\omega(0), Y^\omega_{0,t}\}}R(\tilde{X}_{0,t}, Y^\omega_{0,t})}. \tag{3.7}$$

Equivalently, for all t, s,

$$\langle \Pi^\omega(t+s), \phi\rangle = \frac{E_{\{\Pi^\omega(t), Y^\omega_{t,t+s}\}}[\phi(\tilde{X}(t+s))R(\tilde{X}_{0,s}, Y^\omega_{t,t+s})]}{E_{\{\Pi^\omega(t), Y^\omega_{t,t+s}\}}R(\tilde{X}_{0,s}, Y^\omega_{t,t+s})}. \tag{3.8}$$

$(X^\omega(\cdot), \Pi^\omega(\cdot))$ *is the unique stationary process, and hence its distribution does not depend on* ω *or on the chosen convergent subsequence. Finally,* (3.3) *holds in probability as* $h \to 0$ *and* $T \to \infty$ *in any way at all for any bounded and measurable real-valued function* $F(\cdot)$ *which is continuous almost everywhere with respect to* $\bar{Q}_f(\cdot)$.

4. The discrete time problem. Now we review the discrete time form of the results in the previous section. Let all processes be defined in discrete time. The signal process $X(\cdot) = \{X(n), n < \infty\}$ is assumed to be Feller–Markov and takes values in the compact set G. The observations are defined by $Y(0) = 0$ and

$$\delta Y_n \equiv Y(n) - Y(n-1) = g(X(n)) + \xi(n),\ n = 1, \ldots, \tag{4.1}$$

where $\{\xi(n)\}$ are mutually independent $(0, I)$-Gaussian random variables which are independent of $X(\cdot)$, and $g(\cdot)$ is continuous.

The Bayes' rule formula for the true conditional distribution of $X(n)$ given $Y_{0,n}$ can be represented in terms of an auxiliary process $\tilde{X}(\cdot)$ as for the continuous time case in section 2, where $\tilde{X}(\cdot)$ has the same evolution law as that of $X(\cdot)$ but (conditioned

on its possibly random initial distribution) is independent of all the other processes. Define

$$R(\tilde{X}_{0,n}, Y_{0,n}) = \exp\left[\sum_{i=1}^{n} g'(\tilde{X}(i))\delta Y_i - \frac{1}{2}\sum_{i=1}^{n} |g(\tilde{X}(i))|^2\right].$$

Then the optimal filter $\Pi(\cdot)$ can be defined by its moments

$$\langle \Pi(n), \phi\rangle = \frac{E_{\{\Pi(0), Y_{0,n}\}}[\phi(\tilde{X}(n))R(\tilde{X}_{0,n}, Y_{0,n})]}{E_{\{\Pi(0), Y_{0,n}\}}R(\tilde{X}_{0,n}, Y_{0,n})},$$

where $\Pi(0)$ is the distribution of $X(0)$ and $\tilde{X}(0)$. Alternatively,

$$\langle \Pi(n), \phi\rangle = \frac{E_{\{\Pi(n-1), \delta Y_n\}}[\phi(\tilde{X}(1))R(\tilde{X}(1), \delta Y_n)]}{E_{\{\Pi(n-1), \delta Y_n\}}R(\tilde{X}(1), \delta Y_n)}, \tag{4.2}$$

where with an abuse of notation we have defined

$$R(x, y) = \exp[g(x)'y - |g(x)|^2/2].$$

Analogously to the continuous time observation case, except in some special cases, one cannot evaluate (4.2), and it is generally necessary to approximate it in some way. The approximation problems and methods are similar to those in section 3 (see [3] for details). Namely, build a filter for a simpler Markov process $\tilde{X}^h(\cdot)$ (in discrete time here), which has values in the compact set G and which approximates $X(\cdot)$, but use the actual physical observations.

This procedure is formalized as follows. For $n = 1, \ldots,$ define

$$R(\tilde{X}^h_{0,n}, Y_{0,n}) = \exp\left[\sum_{i=1}^{n} g'(\tilde{X}^h(i))\delta Y_i - \frac{1}{2}\sum_{i=1}^{n} |g(\tilde{X}^h(i))|^2\right].$$

Now define the approximating filter $\Pi^h(\cdot)$ by its moments:

$$\langle \Pi^h(n), \phi\rangle = \frac{E_{\{\Pi^h(0), Y_{0,n}\}}[\phi(\tilde{X}^h(n))R(\tilde{X}^h_{0,n}, Y_{0,n})]}{E_{\{\Pi^h(0), Y_{0,n}\}}R(\tilde{X}^h_{0,n}, Y_{0,n})}. \tag{4.3}$$

Then, one has the following recursive representation for $\Pi^h(\cdot)$.

$$\langle \Pi^h(n), \phi\rangle = \frac{E_{\{\Pi^h(n-1), \delta Y_n\}}[\phi(\tilde{X}^h(1))R(\tilde{X}^h(1), \delta Y_n)]}{E_{\{\Pi^h(n-1), \delta Y_n\}}R(\tilde{X}^h(1), \delta Y_n)}. \tag{4.4}$$

Equations (4.3) and (4.4) correspond to the filter which models the signal process via $\tilde{X}^h(\cdot)$ but uses the actual observations $\delta Y_n = Y(n) - Y(n-1)$.

The process $\tilde{X}^h(\cdot)$ in (4.4) is assumed to be independent of the other processes, given its initial distribution. We also use the following analogue of the basic consistency assumption A2.1, which is the sense of approximation of $X(\cdot)$ by $\tilde{X}^h(\cdot)$.

A4.1. *A consistency assumption.* For any sequence $\{\Pi^h\}$ of probability measures converging weakly to some probability measure Π, $(\tilde{X}^h(0), \tilde{X}^h(1))$ with the initial distribution Π^h converges weakly to $(\tilde{X}(0), \tilde{X}(1))$ with the initial distribution Π.

Analogously to the situation in section 3, (4.2) is well defined even if $\Pi(0)$ is not the initial distribution of $X(\cdot)$. Allowing the initial condition $(X(0), \Pi(0))$ to be arbitrary, the discrete time process $\Psi_f(\cdot) = (X(\cdot), \Pi(\cdot))$ is Feller–Markov. We now write the discrete time analogue of the key uniqueness assumption.

A4.2. *A uniqueness assumption.* There is a unique stationary process $\Psi_f(\cdot) = (X(\cdot), \Pi(\cdot))$. Denote its measure by $\bar{Q}_f(\cdot)$.

The occupation measure. For each n, define $B(n) = \sum_{i=1}^{n} \xi^i$, and define the analogue of $\Psi^h(t, \cdot)$, namely,

$$\Psi^h(n, \cdot) = \{X(n + \cdot), \Pi^h(n + \cdot), Y(n + \cdot) - Y(n), B(n + \cdot) - B(n)\},$$

$$\Psi^h_f(n, \cdot) = \{X(n + \cdot), \Pi^h(n + \cdot)\}.$$

Define the canonical elements of the path spaces $\psi(\cdot)$ and $\psi_f(\cdot)$ analogously, as done in section 3.

The Skorohod topology is replaced by a "sequence" topology, as in [3]. The $\Pi^h(n)$ still takes values in $\mathcal{M}(G)$, and the weak topology is still used on this space. Define the occupation measure $Q^{h,N}(\cdot)$ as follows. For a Borel set C in the product sequence space

$$Q^{h,N}(C) = \frac{1}{N} \sum_{n=1}^{N} I_C(\Psi^h(n, \cdot)). \tag{4.6}$$

Analogously to the definitions in section 3, define $\Psi(\cdot) = (X(\cdot), \Pi(\cdot), Y(\cdot), B(\cdot))$. Let $F(\cdot)$ be a real-valued bounded and continuous (w.p.1 with respect to $\bar{Q}_f(\cdot)$) function of $\psi_f(\cdot)$. Then, the following discrete time analogue of Theorem 3.1 is proved in [3].

THEOREM 4.1. *Let the filtering model be as above. Assume that* A4.2 *holds. Define the approximate filter via* (4.3)*, where we assume that the auxiliary process* $\tilde{X}^h(\cdot)$ *satisfies* A4.1*. Then* $\{Q^{h,N}(\cdot); h > 0, N \geq 0\}$ *is tight. Let* $Q(\cdot)$ *denote a weak sense limit, always as* $h \to 0$ *and* $N \to \infty$*. Let* ω *be the canonical variable on the probability space on which* $Q(\cdot)$ *is defined, and denote the sample values by* $Q^\omega(\cdot)$*. Then, for each* ω*,* $Q^\omega(\cdot)$ *is a measure on the product path (sequence) space. It induces a process*

$$\Psi^\omega(\cdot) = (X^\omega(\cdot), \Pi^\omega(\cdot), Y^\omega(\cdot), B^\omega(\cdot)). \tag{4.7}$$

For almost all ω *the following hold.* $(X^\omega(\cdot), \Pi^\omega(\cdot))$ *is stationary.* $B^\omega(\cdot)$ *is the sum of mutually independent* $N(0, I)$ *random variables* $\{\xi^\omega(n)\}$ *which are independent of* $X^\omega(\cdot)$*. Also*

$$\delta Y^\omega_n \equiv Y^\omega(n) - Y^\omega(n-1) = g(X^\omega(n)) + \xi^\omega(n), \tag{4.8}$$

and $X^\omega(\cdot)$ *has the transition function of* $X(\cdot)$*. For each integer* n *and each bounded and measurable real-valued function* $\phi(\cdot)$*,*

$$\langle \Pi^\omega(n), \phi \rangle = \frac{E_{\{\Pi^\omega(0), Y^\omega_{0,n}\}}[\phi(\tilde{X}(n)) R(\tilde{X}_{0,n}, Y^\omega_{0,n})]}{E_{\{\Pi^\omega(0), Y^\omega_{0,n}\}} R(\tilde{X}_{0,n}, Y^\omega_{0,n})}. \tag{4.9}$$

Finally,

$$\frac{1}{N} \sum_{n=1}^{N} F(\Psi^h_f(n + \cdot)) = \int F(\psi_f(\cdot)) \bar{Q}^{h,N}_f(d\psi_f(\cdot)) \to \int F(\psi_f(\cdot)) \bar{Q}_f(d\psi_f(\cdot)) \tag{4.10}$$

in probability, where $h \to 0$ and $N \to \infty$ in any way at all.

Discussion of the proof. In the problems of this paper, the approximating filter will be constructed using random sampling methods or combinations of random sampling and integration methods. Since many arguments in the proofs are similar to those used in [3], in what follows we will try to use as much of the proof in [3] as possible and to concentrate on the differences. In view of that we now highlight the chief features of the proof in [3]. We comment on the discrete parameter case, but analogous remarks hold for the continuous parameter model. Further details are in the reference. In [3] (see (4.6)), the measure valued random variable $Q^{h,N}(\cdot)$ was obtained as an occupation measure connected with the processes $X(\cdot), \Pi^h(\cdot), B(\cdot), Y(\cdot)$, and the same definition will be used in what follows, but with the new definitions of $\Pi^h(\cdot)$ of this paper used.

The first step in the proof of Theorem 4.1 is to show that the sequence $\{Q^{h,N}(\cdot); h, N\}$ of measure-valued random variables is tight. For that, it suffices to show that the sequence of its expectations is tight [22, Chapter 1.6]. In order to show that, it is enough to show that the families $\{X(n+\cdot); n \geq 0\}$, $\{B(n+\cdot) - B(n); n \geq 0\}$, $\{Y(n+\cdot) - Y(n); n \geq 0\}$, and $\{\Pi^h(n+\cdot); h > 0, n \geq 0\}$ are tight. However, showing that is trivial in view of the compactness of the state space. We note that in the continuous time case the proof of tightness of these processes involves a little more work.

By the first equality in (4.10), the limit is determined by the weak sense limits of the occupation measures, as $N \to \infty$, and $h \to 0$. Thus we need to determine the sample values $Q^\omega(\cdot)$ of any weak sense limit $Q(\cdot)$. Equivalently, we need to characterize the set of processes induced by $Q^\omega(\cdot)$. The proof of the stationarity of $(X^\omega(\cdot), \Pi^\omega(\cdot))$ in [3] will work without any change for the problems of this paper. Furthermore the proofs of the representation (4.8) and that $X^\omega(\cdot)$ has the law of evolution of $X(\cdot)$ for almost all ω will be no different than the analogous arguments in [3], and similarly for the continuous parameter case. Thus, establishing the representation (4.9) becomes the only step in the proof that will be different from that in [3]. Once this step is established, (4.10) follows readily from the uniqueness assumption on the invariant measure of the joint signal and filter process.

The proof of the representation for $\Pi^\omega(\cdot)$ will differ, depending on the choice of $\Pi^h(\cdot)$. The following comments concerning a key detail in the proof of the representation (4.9) in [3] will be useful in providing a guide to the proofs for the cases of this paper.

For arbitrary $\psi(\cdot) = (x(\cdot), \pi(\cdot), y(\cdot), b(\cdot))$, and integer m define the function $A(\cdot)$ by

$$A(\psi(m)) = \langle \pi(m), \phi \rangle - \frac{E_{\{\pi(m-1), y(m)\}}[\phi(\tilde{X}(1)) R(\tilde{X}(1), y(m))]}{E_{\{\pi(m-1), y(m)\}} R(\tilde{X}(1), y(m))}. \tag{4.11}$$

The aim of the proof in [3] was to show that, for almost all ω and all m,

$$A(\Psi^\omega(m)) = 0, \qquad \text{w.p.1}, \tag{4.12}$$

which implies (4.9). This was done by showing that

$$0 = E \int Q^\omega(d\psi) \, [A(\psi(m))]_1^2, \tag{4.13}$$

where we define

$$[A]_1^2 = \min\{|A|^2, 1\}. \tag{4.14}$$

The prelimit form of the right side of (4.13) is

$$E\int Q^{h,N}(d\psi)\,[A(\psi(m))]_1^2\,, \tag{4.15}$$

which, by the definition of $Q^{h,N}(\cdot)$, equals

$$\frac{1}{N}E\sum_{n=1}^{N}[A(\Psi^h(m+n))]_1^2, \tag{4.16}$$

where

$$A(\Psi^h(n)) = \langle \Pi^h(n), \phi\rangle - \frac{E_{\{\Pi^h(n-1),\delta Y_n\}}[\phi(\tilde{X}(1))R(\tilde{X}(1),\delta Y_n)]}{E_{\{\Pi^h(n-1),\delta Y_n\}}R(\tilde{X}(1),\delta Y_n)}. \tag{4.17}$$

In order to show (4.13), it suffices to show

$$E[A(\Psi^h(n))]_1^2 \to 0$$

uniformly in n as $h \to 0$. Finally, to show the above, it suffices, in view of tightness of the families $\{\Pi^h(n); h>0, n>0\}$, $\{\delta Y_n; n>0\}$, and the consistency assumption A4.1, to show that

$$E\left[\langle \Pi^h(n), \phi\rangle - \frac{E_{\{\Pi^h(n-1),\delta Y_n\}}[\phi(\tilde{X}^h(1))R(\tilde{X}^h(1),\delta Y_n)]}{E_{\{\Pi^h(n-1),\delta Y_n\}}R(\tilde{X}^h(1),\delta Y_n)}\right]_1^2 \tag{4.18}$$

converges to 0 uniformly in n as $h \to 0$.

However, in view of the definition of $\Pi^h(n)$ via (4.4), the above expression is identically zero, which implies that (4.13) holds for any weak sense limit. An analogue of this argument will be used in the next section.

5. Some approximating filters of interest: Discrete time. In [3], the approximate filter $\Pi^h(\cdot)$ was defined by the analytical formula (4.4) for the discrete time problem, and by (2.7) for the continuous time problem. One example is the Markov chain approximation method, where the auxiliary process $\tilde{X}^h(\cdot)$ is a Markov chain approximation to $\tilde{X}(\cdot)$. When the dimension is high, such methods can have excessively high computational requirements. Alternatives, based on random sampling or Monte Carlo then become attractive, analogous to the case of classical multidimensional integration [5, 6, 7, 12, 13, 15, 16, 27, 28, 29]. In this section, several forms of this approach will be discussed. We start with the simplest form, which uses unsophisticated random sampling to evaluate the right-hand side of (4.4). The problem is set up so that much of the proof of [3, Theorem 5.1] (this is Theorem 4.1 above) can be used. After treating this simple (but canonical) case, we then move on to more general approximations, pointing out at each instance the crucial condition required for the analogue of Theorem 4.1 to hold.

EXAMPLE 5.1 (the basic "sampling" filter). *Let v^h be a sequence of integers which goes to infinity as $h \to 0$. Let $\Pi^h(n-1)$ denote the estimate of the conditional distribution of $X(n-1)$, given $Y_{0,n-1}$. Given $\Pi^h(n-1)$, we now construct $\Pi^h(n)$ based on "random sampling." Let $\{\tilde{X}^{h,l,n}(\cdot), l \le v^h\}$ be i.i.d. samples (which are independent of δY_n, conditioned on $\Pi^h(n-1)$) from $\tilde{X}^h(\cdot)$, where $\tilde{X}^h(\cdot)$ satisfies the consistency condition* A4.1 *and has the initial distribution $\Pi^h(n-1)$. One need only*

simulate samples of $\tilde{X}^h(0), \tilde{X}^h(1)$.

The filter $\Pi^h(n)$ *is defined by the sample average:*

$$\langle \Pi^h(n), \phi \rangle = \frac{\sum_{l=1}^{v^h} \phi(\tilde{X}^{h,l,n}(1)) R(\tilde{X}^{h,l,n}(1), \delta Y_n)/v^h}{\sum_{l=1}^{v^h} R(\tilde{X}^{h,l,n}(1), \delta Y_n)/v^h}, \tag{5.1}$$

which yields our estimate $\Pi^h(n)$ *of the conditional distribution of* $X(n)$, *given* $Y_{0,n}$.

THEOREM 5.2. *Under* A4.1 *and* A4.2 *and the above construction of* $\Pi^h(\cdot)$, *the conclusions of Theorem* 4.1 *hold.*

Proof. The basic steps in the proof of Theorem 4.1 were outlined after the statement of that theorem. The proof of the current theorem is similar, and we will only concern ourselves with the differences.

The set $\{Q^{h,N}(\cdot); h > 0, T < \infty\}$ is obviously tight since each of the families $\{X(n+\cdot); n \geq 0\}$,$\{B(n+\cdot) - B(n); n \geq 0\}$, $\{Y(n+\cdot) - Y(n); n \geq 0\}$, and $\{\Pi^h(n+\cdot); h > 0, n \geq 0\}$ is tight. Let $Q(\cdot) = \{Q(n), n = 0, 1, \ldots\}$ denote the limit of a weakly convergent subsequence, and denote the samples by $Q^\omega(\cdot)$. Then $Q^\omega(\cdot)$ induces a process $\Psi^\omega(\cdot)$ as in (4.7), and we need to identify the components. The stationarity of $(X^\omega(\cdot), \Pi^\omega(\cdot))$ is proved as in [3], with no change. Similarly, the characterization (4.8), the properties of $B^\omega(\cdot)$, and the fact that $X^\omega(\cdot)$ has the transition function of $X(\cdot)$ is done exactly as in [3].

The main difference is in the proof of (4.9). Proceeding as illustrated for Theorem 4.1, to identify $\Pi^\omega(\cdot)$ we need only to show (4.13). Analogously to the procedure in section 4, this is done by showing that the expression in (4.18) converges to 0 uniformly in n as $h \to 0$. By using the definition of $\Pi^h(n)$, (4.18) can be rewritten as

$$E\left[\frac{\sum_{l=1}^{v^h} \phi(\tilde{X}^{h,l,n}(1)) R(\tilde{X}^{h,l,n}(1), \delta Y_n)/v^h}{\sum_{l=1}^{v^h} R(\tilde{X}^{h,l,n}(1), \delta Y_n)/v^h} - \frac{E_{\{\Pi^h(n-1), \delta Y_n\}}[\phi(\tilde{X}^h(1)) R(\tilde{X}^h(1), \delta Y_n)]}{E_{\{\Pi^h(n-1), \delta Y_n\}} R(\tilde{X}^h(1), \delta Y_n)}\right]_1^2. \tag{5.2}$$

Owing to the properties of the $[\cdot]_1^2$ metric defined by (4.14), we can work with the numerators and denominators separately, and it is only necessary to show that, for arbitrary bounded and continuous $\phi(\cdot)$,

$$E\left[\frac{1}{v^h}\sum_{l=1}^{v^h} \phi(\tilde{X}^{h,l,n}(1)) R(\tilde{X}^{h,l,n}(1), \delta Y_n) - E_{\{\Pi^h(n-1), \delta Y_n\}} \phi(\tilde{X}^h(1)) R(\tilde{X}^h(1), \delta Y_n)\right]_1^2 \tag{5.3}$$

goes to zero uniformly in n, as $h \to 0$. But this clearly holds since for each h and n, $\{\tilde{X}^{h,l,n}(\cdot), l\}$ are mutually i.i.d. and independent of δY_n (conditioned on $\Pi^h(n-1)$), and the mean square value (conditional on $\{\Pi^h(n-1), \delta Y_n\}$) of the functional

$$\phi(\tilde{X}^{h,l,n}(1)) R(\tilde{X}^{h,l,n}(1), \delta Y_n) - E_{\{\Pi^h(n-1), \delta Y_n\}} \phi(\tilde{X}^{h,l,n}(1)) R(\tilde{X}^{h,l,n}(1), \delta Y_n)$$

has uniformly (in h, l, n) bounded expectation. □

We remark that in the above proof we found it convenient to work with the expression in (5.3), however in view of the consistency condition A4.1 on $\tilde{X}^h(\cdot)$, showing that the expression in (5.3) goes to zero, uniformly in n, as $h \to 0$ is equivalent to showing the same for the expression in (5.3′) below.

$$E\left[\frac{1}{v^h}\sum_{l=1}^{v^h}\phi(\tilde{X}^{h,l,n}(1))R(\tilde{X}^{h,l,n}(1),\delta Y_n) - E_{\{\Pi^h(n-1),\delta Y_n\}}\phi(\tilde{X}(1))R(\tilde{X}(1),\delta Y_n)\right]_1^2. \tag{5.3′}$$

EXAMPLE 5.3 (some generalizations of the filter in Example 5.1). As can be observed from the proof of Theorem 5.1, the crucial step is the establishing of convergence of the expression in (4.18) or, equivalently, of (5.3′), the form that we will use. This convergence is essentially the consequence of the consistency condition A4.1. However, as we will indicate in the following discussion, this consistency condition can be weakened considerably. This leads to many useful extensions of the basic form of the "sampling" algorithm of Example 5.1.

A weaker form of the consistency assumption A4.1. We retain the assumption of mutual independence (conditional on $\Pi^h(n-1),\delta Y_n$) of the $\{\tilde{X}^{h,l,n}(\cdot), l \le v^h\}$ for each h,n, and that the probability law of $\{\tilde{X}^{h,l,n}(0)\}$ is $\Pi^h(n-1)$, but we allow more flexibility in the choice of the individual $\tilde{X}^{h,l,n}(\cdot)$. Namely, in the construction of $\Pi^h(n)$ in (5.1), the Markov family from which $\tilde{X}^{h,l,n}(\cdot)$ is sampled may differ for different l,n. However, the initial conditions $\tilde{X}^{h,l,n}(0)$ still form an i.i.d. sample from $\Pi^h(n-1)$. To see the possibilities, write the expression in the brackets in (5.2) as the sum of the two terms:

$$\frac{\sum_{l=1}^{v^h}\phi(\tilde{X}^{h,l,n}(1))R(\tilde{X}^{h,l,n}(1),\delta Y_n)/v^h}{\sum_{l=1}^{v^h}R(\tilde{X}^{h,l,n}(1),\delta Y_n)/v^h} - \frac{E_{\{\Pi^h(n-1),\delta Y_n\}}\sum_{l=1}^{v^h}\phi(\tilde{X}^{h,l,n}(1))R(\tilde{X}^{h,l,n}(1),\delta Y_n)/v^h}{E_{\{\Pi^h(n-1),\delta Y_n\}}\sum_{l=1}^{v^h}R(\tilde{X}^{h,l,n}(1),\delta Y_n)/v^h}, \tag{5.4}$$

and

$$\frac{E_{\{\Pi^h(n-1),\delta Y_n\}}\sum_{l=1}^{v^h}\phi(\tilde{X}^{h,l,n}(1))R(\tilde{X}^{h,l,n}(1),\delta Y_n)/v^h}{E_{\{\Pi^h(n-1),\delta Y_n\}}\sum_{l=1}^{v^h}R(\tilde{X}^{h,l,n}(1),\delta Y_n)/v^h} - \frac{E_{\{\Pi^h(n-1),\delta Y_n\}}[\phi(\tilde{X}(1))R(\tilde{X}^h(1),\delta Y_n)]}{E_{\{\Pi^h(n-1),\delta Y_n\}}R(\tilde{X}(1),\delta Y_n)}. \tag{5.5}$$

Owing to the use of the $[\cdot]_1^2$ metric defined by (4.14), it is enough to work separately with the differences of the numerators and of the denominators in each of (5.4) and (5.5). Then, to handle (5.4), use the mutual independence and the uniform bounds on the expectations of the conditional variances. To handle (5.5), we will use a revised form of the consistency assumption A4.1, which is the following.

A5.1. For each (n,h), the set $\{\tilde{X}^{h,l,n}(\cdot), l\}$ is mutually independent and independent of δY_n, conditioned on $\Pi^h(n-1)$. Suppose that an arbitrary Π^h replaces

$\Pi^h(n-1)$ in the construction of the $\{\tilde{X}^{h,l,n}(0), l\}$. Then, as $h \to 0$, for any such sequence and for each bounded, continuous, and real-valued function $\Phi(\cdot)$,

$$E_{\Pi^h}\Phi(\tilde{X}^{h,l,n}(1)) - E_{\Pi^h}\Phi(\tilde{X}(1)) \to 0 \tag{5.6}$$

uniformly in n and l.

This assumption, when used for $\Phi(x) \equiv \Phi_y(x) = \phi(x)R(x,y)$ for each fixed y, leads to the desired convergence for the expression in (5.6). Observe that even though $R(\cdot)$ is not bounded, we can, without loss of generality, assume so since the family $\{\delta Y_n; n \geq 1\}$ is tight. For this reason and the fact that we use the metric (4.14), we do not need the convergence in A5.1 for $\Phi(\cdot) = \Phi_y(\cdot)$ to hold uniformly in y.

Note that we are no longer assuming that the $\tilde{X}^{h,l,n}(\cdot)$ are all samples of the same $\tilde{X}^h(\cdot)$ process. The second part of A5.1 will hold iff for all Π and any sequence $\{h_k, l_k, n_k\}_{k\geq 1}$ for which ($\mathcal{L}(X)$ denotes the probability law of X)

$$\mathcal{L}(\tilde{X}^{h_k,l_k,n_k}(0)) \Rightarrow \Pi$$

as $k \to \infty$, we have that

$$\mathcal{L}(\tilde{X}^{h_k,l_k,n_k}(0), \tilde{X}^{h_k,l_k,n_k}(1)) \Rightarrow \mathcal{L}(\tilde{X}(0), \tilde{X}(1)) \text{ as } k \to \infty,$$

where $\tilde{X}(0)$ has the law Π.

Dropping the mutual conditional independence. Return to the expression (5.2). Let $\Phi(\cdot)$ be bounded and continuous. Then the convergence in (5.2) is implied by the following even weaker consistency assumption, which can replace A4.1 and the mutual independence in Theorem 5.1.

A5.2. For each (h,n), $\{\tilde{X}^{h,l,n}(\cdot), l\}$ is independent of δY_n, conditioned on $\Pi^h(n-1)$, but they might not be independent in l. They are constructed subject to the following rule. Suppose that an arbitrary measure $\Pi^{h,n}$ (on G) takes the role of $\Pi^h(n-1)$ in the construction of the $\{\tilde{X}^{h,l,n}(\cdot), l\}$. Then the associated process $\{\tilde{X}^{h,l,n}(\cdot), l\}$ is constructed such that as $h \to 0$ for any bounded, continuous, and real-valued function $\Phi(\cdot)$,

$$\frac{1}{v^h}\sum_{l=1}^{v^h} \Phi(\tilde{X}^{h,l,n}(1)) - E_{\{\Pi^{h,n}\}}\Phi(\tilde{X}(1)) \to 0, \tag{5.7}$$

in probability, uniformly in n.

It is clear that this condition (instead of A4.1 and mutual independence of samples) suffices for Theorem 5.2 to hold for the corresponding $\{\Pi^h(n)\}$. The usefulness of this condition lies in the cases where the samples $\{\tilde{X}^{h,l,n}(\cdot), l \leq v^h\}$ for fixed h,n are not mutually independent. It is of particular value when the random sampling incorporates some variance reduction method where the samples are not mutually independent, e.g., antithetic variables or stratified sampling such as discussed next.

Variance reduction methods. The standard methods for variance reduction in Monte Carlo, such as stratified sampling and antithetic variables, can all be used here and in the subsequent algorithms and examples. We comment on one form of stratified sampling. Let $\Pi^h(n-1)$ be concentrated on points $\{x^{h,l,n}; l = 1, \ldots, v^h\}$, and let $\Pi^h_l(n-1)$ denote the weight that $\Pi^h(n-1)$ puts on $x^{h,l,n}$.

In this example, we only use variance reduction to get the samples of the initial values $\tilde{X}^{h,l,n}(0)$. Once these are given, the sample values of $\tilde{X}^{h,l,n}(1)$ are obtained by

sampling independently, using the transition probability of the approximating Markov process $\tilde{X}^{h,n}(\cdot)$. So we concentrate on the initial values for fixed n.

If the $v^h\Pi_l^h(n-1)$ were all integers, then the best sampling of the values of the $\tilde{X}^{h,l,n}(0)$ would be to take the initial point $x^{h,l,n}$ exactly $v^h\Pi_l^h(n-1)$ times, since then the variance of the sampling error of the *initial condition* would be zero. Clearly all the $v^h\Pi_l^h(n-1)$ would not usually be integers, but one tries to approximate the ideal as well as possible. One common approach is the following. First take the point $x^{h,l,n}$ exactly $[v^h\Pi_l^h(n-1)]$ (the integer part) times. After this step, the "residual" number of points remaining to be chosen is

$$\delta v^{h,n} = \sum_l \delta v_l^{h,n},$$

where

$$\delta v_l^{h,n} = (v^h\Pi_l^h(n-1) - [v^h\Pi_l^h(n-1)]).$$

The "residual frequency" of point $x^{h,l,n}$ is $\delta v_l^{h,n}/\delta v^{h,n}$. Now divide the set $\{x^{h,l,n}, l\}$ into disjoint subsets $S_i^{h,n}, i = 1, \ldots$. The set $S_i^{h,n}$ has $\delta v^{h,n,i}$ points where

$$\delta v^{h,n,i} = \sum_{l \in S_i^{h,n}} \delta v_l^{h,n}.$$

Allocate $[\delta v^{h,n,i}]$ initial points to subset i, and then select these points randomly (with replacement) from $S_i^{h,n}$, where the point $x^{h,l,n} \in S_i^{h,n}$ is given the weight $\delta v_l^{h,n}/\delta v^{h,n,i}$. Since

$$\bar{v}^{h,n} \equiv \delta v^{h,n} - \sum_i [\delta v^{h,n,i}] \geq 0,$$

we still need to allocate $\bar{v}^{h,n}$ points, if this is positive. Generally, if the division into subgroups is done properly, $\bar{v}^{h,n}/v^h$ will be either zero or small. If it is positive, either repeat the above procedure to allocate the remaining $\bar{v}^{h,n}$ points, or just select $\bar{v}^{h,n}$ points randomly from the original v^h points with appropriately modified weights.

It is easy to see that the above construction can be put in the framework of Example 5.3, and condition A5.2 holds.

The grouping into subsets might be done by dividing the points according to their "geographic location," if this is meaningful.

EXAMPLE 5.4. We would like to treat algorithms that use combinations of random sampling and integration methods in a general way. This will require an alteration in the consistency condition A5.1 or A5.2. In order to motivate the form which it will take, we first consider an example for which A5.2 is satisfied. Let $\tilde{X}^{h,n}(\cdot)$ be processes satisfying A5.1. Having defined the approximate filter $\Pi^h(j)$ for $j = 1, 2, \ldots, n-1$ suppose that $\{\tilde{X}^{h,l,n}(\cdot), l \leq v^h\}$ are samples of $\tilde{X}^{h,n}(\cdot)$ and that they are conditionally independent of δY_n given $\Pi^h(n-1)$. Define $\Pi^h(n)$ via (5.1). If the samples are mutually independent (conditioned on $\Pi^h(n-1)$) and $\tilde{X}^{h,n}(0)$ has distribution $\Pi^h(n-1)$, then condition A5.1 (and thus A5.2) is satisfied. Theorem 5.2 can be proved under weaker consistency conditions than A5.1 or A5.2, which allow great and useful flexibility in constructing the filter. To motivate a useful general form, let us first rewrite Example 5.3 in the following suggestive way.

For each h and n, define a measure (on the sample space $G \times G$) valued random variable $P^{h,n}_{\Pi^{h,n-1}}$ as follows. Let $P^{h,n}_{\Pi^h(n-1)}(A)$ be the fraction of the samples $\{\tilde{X}^{h,l,n}(\cdot), l \le v^h\}$ that are in the Borel set $A \subset G \times G$. In particular, $P^{h,n}_{\Pi^h(n-1)}\{B \times G\}$ is the fraction of the samples $\tilde{X}^{h,l,n}(0)$ which are in the set B. Since this $P^{h,n}_{\Pi^h(n-1)}$ is just the "sampling occupation measure," condition A5.2 is equivalent to

$$E\left[\int \Phi(x(1))dP^{h,n}_{\Pi^h(n-1)}(x(\cdot)) - E_{\{\Pi^h(n-1)\}}\Phi(\tilde{X}(1))\right]^2_1 \to 0 \tag{5.8}$$

uniformly in n as $h \to 0$.

Thus the crucial condition becomes the convergence of the expression in (5.8). The advantage of writing the condition in the form (5.8) is that, being written in terms of a random measure $P^{h,n}_{\Pi^h(n-1)}$, it suggests other choices of approximate filters that need not be based exclusively on Monte Carlo or random sampling. For example, as seen in Example 5.5 below, $P^{h,n}_{\Pi^h(n-1)}$ might be determined partly by random sampling and partly analytically.

A generalization of $P^{h,n}_{\Pi^h(n-1)}$ *and the approximating filter.* Motivated by the suggestiveness of (5.8), we now consider the following general form of the approximate filter and the consistency condition. Let $\{\Pi^h(n); n \ge 1\}$ be defined recursively as follows. Having defined $\Pi^h(n-1)$, let $P^{h,n}_{\Pi^h(n-1)}$ be a measure-valued random variable on the sample space $G \times G$, which is conditionally independent of δY_n given $\Pi^h(n-1)$. Define $\Pi^h(n)$ by

$$\langle \Pi^h(n), \phi\rangle = \frac{\int \phi(x(1))R(x(1), \delta Y_n)dP^{h,n}_{\Pi^h(n-1)}(x(\cdot))}{\int R(x(1), \delta Y_n)dP^{h,n}_{\Pi^h(n-1)}(x(\cdot))}. \tag{5.9}$$

We will need the following consistency condition.

A5.3. For each bounded, continuous, and real-valued function $\Phi(\cdot)$, as $h \to 0$,

$$\int \Phi(x(1))dP^{h,n}_{\Pi^h(n-1)}(x(\cdot)) - E_{\{\Pi^h(n-1)\}}\Phi(\tilde{X}(1)) \to 0, \tag{5.10}$$

in probability, uniformly in n.

We now have the following useful result whose proof follows from the above comments.

THEOREM 5.5. *Theorem* 5.2 *holds for the above constructed* $\Pi^h(\cdot)$ *if in the assumptions of that theorem the consistency condition* A5.3 *replaces* A4.1 *and the mutual independence of the samples.*

Remarks. Although not needed, it will often be the case that

$$E_{\Pi^h(n-1)}P^{h,n}_{\Pi^h(n-1)}\{B \times G\} = \Pi^h(n-1)(B). \tag{5.11}$$

The advantage of A5.3 is that it can be used for a large variety of approximation methods. For example, in the form (4.4), $P^{h.n}_{\Pi^h(n-1)}$ would be the measure of $(\tilde{X}^h(0), \tilde{X}^h(1))$ with $\tilde{X}^h(0)$ having the (random) distribution $\Pi^h(n-1)$. The conditions for the convergence for the classical Markov chain approximation, the random sampling method above, and various combinations of them, either in the same or in different time frames, can all be put into the form of (5.10) for appropriate choices

of $P^{h,n}_{\Pi^h(n-1)}$. Importance sampling methods can also be fit into the same scheme and used to improve the performance of the filter, as shown in the next section.

EXAMPLE 5.6 (an application of A5.3). A combination of random sampling and integration algorithms will be considered in this example. Consider the following commonly used model. Let $X(n) = b(X(n-1), \zeta(n-1))$, where $b(\cdot)$ is bounded and continuous and the $\{\zeta(n)\}$ are mutually i.i.d. (with distribution function P_ζ, with compact support), and independent of $X(0)$. First, suppose that $\Pi(n-1)$ is the actual conditional distribution of $X(n-1)$, given $Y_{0,n-1}$. Then the optimal $\Pi(n)$ is defined by (4.2). If the computation on the right side of (4.2) is not possible, as is usually the case, it would be approximated in some way. The difficulties in evaluating the right side might be due to the problem of computing the one step transition probability of the Markov process $\{X(n)\}$, or to the actual integrations over a possibly continuous state space that are required to evaluate (4.2). As usual, let h denote the approximation parameter for the actual practical filter, and let $\Pi^h(n)$ denote the estimate of the conditional distribution given $Y_{0,n}$.

Let $\Pi^h(n-1)$ be given. We wish to compute $\Pi^h(n)$. This can be done by a direct simulation as in Example 5.1, by combined "simulation-integration," or perhaps even by a pure "integration," method. These possibilities will be illustrated. Suppose that we approximate P_ζ by P^h_ζ, which might have a (computationally) more convenient support and is such that $P^h_\zeta \Rightarrow P_\zeta$. In addition, approximate $b(\cdot)$ by a measurable function $b_h(\cdot)$ such that

$$\lim_{h\to 0} \sup_{x,\zeta} |b(x,\zeta) - b_h(x,\zeta)| = 0.$$

If the associated integrations are convenient to carry out, one can use (4.3) to define $\Pi^h(n)$, where we define

$$\tilde{X}^h(1) = b_h(\tilde{X}^h(0), \zeta^h). \tag{5.12a}$$

In (5.12a), $\tilde{X}^h(0)$ has distribution $\Pi^h(n-1)$ and ζ^h has distribution P^h_ζ. If the support of P^h_ζ is finite and $b_h(\cdot)$ takes only finitely many values, then the integrations reduce to summations. The $\tilde{X}^h(\cdot)$ process thus defined satisfies the consistency condition A4.1. Hence Theorem 4.1 holds for $\Pi^h(\cdot)$ if A4.2 holds.

Alternatively, one can simply use Monte Carlo as in Example 5.1. All sampling below is "conditionally independent" of the past, given $\Pi^h(n-1)$. Take v^h independent samples from $\Pi^h(n-1)$ and from P^h_ζ and call them $\tilde{X}^{h,n,l}(0)$ and $\zeta^{h,n,l}$, $l \le v^h$, resp. Then use the formula

$$\tilde{X}^{h,n,l}(1) = b_h(\tilde{X}^{h,n,l}(0), \zeta^{h,n,l}) \tag{5.12b}$$

and (5.1) to get $\Pi^h(n)$.

Combinations of the above two approaches might be worthwhile also. For example, if the support of the P^h_ζ is a (not too big) finite set, then one can sample from $\Pi^h(n-1)$, but "integrate" over the noise for each sample of the initial condition by doing the necessary summations. One would normally try to choose the support of P^h_ζ such that the integrals are well approximated for an appropriate set of functions $\phi(\cdot)$. On the other hand, one might discretize the state space such that the support of the $\tilde{X}^h(0)$ (i.e., of each of the $\Pi^h(n)$) is confined to a finite set G_h, and integrate with respect to the "initial" measure $\Pi^h(n-1)$, but simulate the noise. For each of these combinations, there is a $P^{h,n}_{\Pi^h(n-1)}$ such that A5.3 holds, provided that the

discretization of the space converges to the whole space in an appropriate manner and, for the part of the computation which involves random sampling, the number of samples goes to infinity as $h \to 0$. The construction of $P^{h,n}_{\Pi^h(n-1)}$ is not hard and the details are omitted.

EXAMPLE 5.7 (a Markov chain approximation method). In this example, by considering a sampled diffusion model, we illustrate a potentially useful combination of integration and simulation. Suppose that the signal process $X(n)$ is a sample from a diffusion process $X(\cdot)$ at discrete time n. Suppose that $X(\cdot)$ solves an Itô equation with a unique weak sense solution for each initial condition and has continuous drift and diffusion coefficients. Then the exact filter (4.2) involves getting the probability distribution of $\tilde{X}(1)$ (which solves the same Itô equation) with the correct initial distribution $\Pi(n-1), n = 1, \ldots$. One could try to solve the Fokker–Planck equation by some numerical method. This is not easy when there are degeneracies. The Markov chain approximation method [24] is a general and powerful approach which converges under the specified conditions, even with quite weak (reflecting) boundary reflections and jumps added. The following discussion uses the idea of the Markov chain approximation without going into excessive details.

For each h, let $\{X^h_n\}$ be a discrete parameter Markov chain on a finite state space $G_h \subset G$, and let $p^h(x,z)$ denote the one step transition probabilities. For $\delta_h > 0$, define the continuous time interpolation $\tilde{X}^h(\cdot)$ by $\tilde{X}^h(t) = X^h_n$ on the interval $[n\delta_h, n\delta_h + \delta_h)$. Suppose, without loss of generality, that $1/\delta_h$ is an integer, and assume that $\tilde{X}^h(\cdot)$ satisfies the consistency assumption A4.1. The use of such chains in the construction of approximate filters is now quite common. See [21, 24, 23]. The references [21, 24] give straightforward and automatic ways of constructing such chains.

In [21, 24] and in current usage in applications, the process $\tilde{X}^h(\cdot)$ is used as in the algorithm (4.4). But it can also be used as the basic simulated process in (5.1). In order to demonstrate the possibilities, we now illustrate an interesting combination of these two schemes which is rather different from the combinations illustrated by Example 5.6. We work with a single observation interval at a time, and for concreteness we discuss the method for the time interval $[0,1]$. Let $\Pi^h = \Pi^h(0)$ denote the approximation to the distribution of $X(0)$. We need to approximate the distribution of $\tilde{X}(1)$. This is done by either computing or estimating the distribution of X^h_{1/δ_h}, where X^h_0 has the distribution Π^h.

To estimate or compute the distribution of X^h_{1/δ_h}, we recursively estimate or compute the distribution of the X^h_m for $m = 1, 2, \ldots, 1/\delta_h$. The motivation behind the combined integration/Monte Carlo procedure to be described is that in some regions of the state space, it might be easier to use one method and in other regions the other method. For illustrative purposes, we suppose that G is divided into disjoint subsets G_1 and G_2 and define $G_{h,i} = G_h \cap G_i$. Suppose that it is easy to compute the transition probabilities $p^h(x,z)$ for $x \in$ some neighborhood of G_1, but harder for x outside of that neighborhood. We suppose that it is feasible to run simulations of the process for any selected initial condition. For example, $p^h(x,z)$ might be given implicitly as the output of a complicated physical mechanism for which the transition probability is hard to compute when x is in G_2 but which can be simulated. We try to exploit this situation by simulating where convenient and integrating where that is convenient. The division into subsets G_i and the associated "difficulty" of some procedure in G_i is meant to be suggestive. We will sometimes be "integrating" when in G_2 and sometimes simulating when in G_1. But the major part of each type of computation will be done in the region where it is advantageous.

Let us divide the unit interval into n_h (an integer) subintervals, with $\epsilon = \delta_h k_h$. Thus $1/\delta_h = k_h n_h$. For notational simplicity, we work with the interval $[0,1]$, but the method is identical for all the intervals $[n, n+1]$. Let μ_m^h denote the estimate of the distribution of $X_{mk_h}^h$ with values $\mu_m^h(x), x \in G_h$, where we start with $\mu_0^h = \Pi^h$. The time interval ϵ should be small enough such that the paths starting in $G_{h,i}$ stay close to it with a high probability on that interval. One needs to be careful if ϵ is allowed to go to zero as $h \to 0$ (which we do not do), since it is well known from simulations that the procedure can degenerate unless v^h goes to infinity fast enough (and at a rate which depends on how fast $\epsilon \to 0$). In fact, there is little loss of generality or practicality in fixing ϵ to be a small constant.

Suppose that μ_m^h is given. Then μ_{m+1}^h is computed as follows. First, do the analytic computation using $p^h(x,z)$ to get the part of $\mu_{m+1}^h(z)$ which is due to the initial states in $G_{h,1}$. Namely, compute

$$\sum_{x \in G_{h,1}} P\{X_{(m+1)k_h}^h = z \big| X_{mk_h}^h = x\} \mu_m^h(x). \tag{5.13a}$$

The part of $\mu_{m+1}^h(z)$ which is due to initial states in $G_{h,2}$ is obtained by simulation. To simulate, we sample a total of v^h points (where $v^h \to \infty$) in $G_{h,2}$ with relative probabilities $\{\mu_m^h(x), x \in G_{h,2}\}$. Denote the samples by $\{X_0^{h,l,m}, l \le v^h\}$. The sampling of the $\{X_0^{h,l,m}, l \le v^h\}$ can be done either with replacement or, preferably, using a variance reduction method such as the one based on stratified sampling which was described at the end of Example 5.3. For each of these "initial values" $X_0^{h,l,m}, l \le v^h$, simulate at random a path of the chain for k_h steps. Let $\{X_k^{h,l,m}, k \le k_0\}$ denote the sample values.

Then the part of the estimate of $\mu_{m+1}^h(z)$ which is due to initial states in $G_{h,2}$ is

$$\left[\sum_{x \in G_{h,2}} \mu_m^h(x) \right] \frac{1}{v^h} \sum_{l=1}^{v_h} I_{\{X_{k_h}^{h,l,m} = z\}}. \tag{5.13b}$$

The sum of (5.13a) and (5.13b) is $\mu_{m+1}^h(z)$. If $h \to 0$ and Π^h converges weakly to, say, Π, it follows from the consistency condition A4.1 for the $\tilde{X}^h(\cdot)$ process that $\mu_{n_h}^h(\cdot)$ converges weakly to the distribution of $\tilde{X}(1)$, which corresponds to $\tilde{X}(0)$ having distribution Π.

To identify the measure $P_{\Pi^h(n-1)}^{h,n}$ in A5.3 (for our example $n = 1$) note the following. It is the measure on $G_h \times G_h$, with initial distribution given by our combination of $\mu_0^h(x)$ for $x \in G_{h,1}$ and the sampling distribution for $x \in G_{h,2}$. The distribution of the terminal value, conditioned on the initial distribution, is computed by repeating the updating procedure outlined above $1/\epsilon$ times.

We note that variance reduction methods can be employed in the sampling of the random paths themselves.

6. More examples: Importance sampling methods for discrete time models. Importance sampling methods are in common use to improve the performance of Monte Carlo algorithms (see, for example, [10, 11]). The basic idea can be seen from the following simple example. Suppose that we wish to estimate $Ef(X)$ via Monte Carlo, where $f(\cdot)$ is bounded and continuous and X has distribution P. The simplest estimate has the form $\sum_{i=1}^n f(X_i)/n$, where $\{X_i\}$ are mutually independent and chosen at random from P. Suppose that we know that values of X in a set A have

the dominant effect on $Ef(X)$, where A has small probability. Then large n would be needed to get a good estimate. Let Q be mutually absolutely continuous with respect to P, and where $Q(A)$ has a "moderate" value. Then for appropriate choice of Q, the unbiased estimate $\sum_{i=1}^{n} f(X_i)[dP/dQ](X_i)/n$ has a much smaller variance than the original estimate, where now the X_i are drawn at random and independently from Q [10, 11]. Importance sampling attempts to sample more (using Q) in the regions which prior information suggests are more important, and corrects for this bias via the weighing with the Radon–Nikodym derivative.

Importance sampling has also been used to improve the quality of nonlinear filtering algorithms that use random sampling [5, 29]. When used over an infinite time interval, the robustness and convergence questions raised earlier in the paper remain important. In the next example, we discuss the general idea of importance sampling and show how the associated proof of convergence is covered by what has already been done for setups such as that in Example 5.1 of section 5. In this next example, we describe the importance sampling on a typical interval $[n-1, n]$, and it does not use the next observation δY_n. This next observation can provide useful information to guide the sampling. There are many intriguing possibilities, and one form of such a use is discussed in Example 6.4. We only illustrate some possibilities. There are numerous possible variations, and the choice of the better ones is still a matter of research. With all of the variations, variance reduction methods can be used, as can combined sampling-integration methods.

EXAMPLE 6.1 (the basic idea of importance sampling in Monte Carlo filtering). Return to the setup used in Example 5.1 of section 5. Let P^h_{n-1} denote the probability law of $\tilde{X}^h(\cdot) = (\tilde{X}^h(0), \tilde{X}^h(1))$ when $\Pi^h(n-1)$ is the measure of $\tilde{X}^h(0)$. For each h and n, let $M^{h,n}$ denote a random measure which is almost surely (a.s.) mutually absolutely continuous (i.e., measure equivalent) with respect to P^h_{n-1}. For each h and n, let $\{\tilde{X}^{h,l,n}(\cdot), l \le v^h\}$ be mutually conditionally independent, conditioned on $\delta Y_n, P^h_{n-1}, M^{h,n}$, and with the distribution $M^{h,n}$. Define the likelihood ratio (the Radon–Nikodym derivative) and its value on the random path $\tilde{X}^{h,k,n}(\cdot)$ by

$$L^{h,n} = \frac{dP^h_{n-1}}{dM^{h,n}}, \qquad L^{h,k,n} = \frac{dP^h_{n-1}}{dM^{h,n}}(\tilde{X}^{h,k,n}(\cdot)). \tag{6.1}$$

We introduce the following assumption.

A6.1.

$$\sup_{h,n} E\frac{dP^h_{n-1}}{dM^{h,n}}(\tilde{X}^h(0), \tilde{X}^h(1))R^2(\tilde{X}^h(1), \delta Y_n) < \infty, \tag{6.2}$$

where $\tilde{X}^h(\cdot)$ in (6.2) has the distribution P^h_{n-1} (conditioned on $\delta Y_n, P^h_{n-1}, M^{h,n}$).

Define the approximate filter $\Pi^h(\cdot)$ to be:

$$\langle \Pi^h(n), \phi\rangle = \frac{\sum_{l=1}^{v^h} L^{h,k,n}\phi(\tilde{X}^{h,l,n}(1))R(\tilde{X}^{h,l,n}(1), \delta Y_n)/v^h}{\sum_{l=1}^{v^h} L^{h,k,n}R(\tilde{X}^{h,l,n}(1), \delta Y_n)/v^h}. \tag{6.3}$$

THEOREM 6.2. *Assume* A4.1, A4.2, A6.1, *and the filter form* (6.3). *Then Theorem* 5.1 *holds.*

Proof. To prove the theorem, it suffices to show that

$$E\left[\frac{1}{v^h}\sum_{k=1}^{v_h}(L^{h,k,n}\Phi(\tilde{X}^{h,k,n}(1), \delta Y_n) - E_{\Pi^h(n-1),\delta Y_n}\Phi(\tilde{X}^h(1), \delta Y_n))\right]^2 \tag{6.4}$$

converges to 0 uniformly in n as $h \to 0$, where $\Phi(x,y) = \phi(x)R(x,y)$ and $\phi(\cdot)$ is any bounded and continuous real-valued function.

We can write

$$\begin{aligned} &E_{M^{h,n},P^h_{n-1},\delta Y_n} L^{h,k,n}\Phi(\tilde{X}^{h,k,n}(1),\delta Y_n) \\ &= E_{M^{h,n},P^h_{n-1},\delta Y_n}\Phi(\tilde{X}^h(1),\delta Y_n) = E_{\Pi^h(n-1),\delta Y_n}\Phi(\tilde{X}^h(1),\delta Y_n), \end{aligned} \tag{6.5}$$

where as before the subscript in the expectation is the conditioning data and $\tilde{X}^h(\cdot)$ in the second and the third expression has the law (conditioned on $M^{h,n}, P^h_{n-1}, \delta Y_n$) P^h_{n-1}. The first equality follows by the definition of the Radon–Nikodym derivative. The second equality is simply a statement of the fact that all we need to know about $\tilde{X}^h(\cdot)$ to compute the expectation is its initial distribution and the one step law of evolution.

Under $M^{h,n}$, the samples are mutually independent, conditioned on $\delta Y_n, P^h_{n-1}, M^{h,n}$. By the above facts, (6.4) can be rewritten as

$$\begin{aligned} &EE_{M^{h,n},P^h_{n-1},\delta Y_n}\left[\frac{1}{v^h}\sum_{k=1}^{v_h}(L^{h,k,n}\Phi(\tilde{X}^{h,k,n}(1),\delta Y_n) - E_{\Pi^h(n-1),\delta Y_n}\Phi(\tilde{X}^h(1),\delta Y_n))\right] \\ &= \frac{1}{(v^h)^2}EE_{M^{h,n},P^h_{n-1},\delta Y_n}\sum_{k=1}^{v_h}[L^{h,k,n}\Phi(\tilde{X}^{h,k,n}(1),\delta Y_n) - E_{\Pi^h(n-1),\delta Y_n}\Phi(\tilde{X}^h(1),\delta Y_n)]^2. \end{aligned} \tag{6.6}$$

The right-hand side of (6.6) is

$$O(1)\frac{1}{v^h}EE_{M^{h,n},P^h_{n-1},\delta Y_n}[L^{h,n}R^2(\tilde{X}^h(1),\delta Y_n)].$$

The last term, in turn, can be bounded by

$$O(1)\frac{1}{v^h}\left(E\left[\frac{dP^h_{n-1}}{dM^{h,n}}(\tilde{X}^h(0),\tilde{X}^h(1))R^2(\tilde{X}^h(1),\delta Y_n)\right] + ER^2(\tilde{X}^h(1),\delta Y_n)\right),$$

where $\tilde{X}^h(\cdot)$ is as in A6.1. The above expression is easily seen to be $O(1/v^h)$ by (6.2). □

An extension: Dropping the mutual independence. As in Example 5.3 of section 5, where we relaxed the condition on mutual independence of samples by instead assuming A5.2, we can formulate a similar condition here which can be used to incorporate variance reduction methods along with importance sampling. More precisely, let $M^{h,n}$ be as before. Also let $\{\tilde{X}^{h,l,n}(\cdot); l \le v^h\}$ be as before, except that they need not be (conditionally) mutually independent. Instead of assuming A6.1, assume A6.2 below.

A6.2. Let $\Phi(x,y) = \phi(x)R(x,y)$, where $\phi(\cdot)$ is a bounded and continuous real-valued function. Then

$$E\left[\frac{1}{v^h}\sum_{k=1}^{v_h}(L^{h,k,n}\Phi(\tilde{X}^{h,k,n}(1),\delta Y_n) - E_{\Pi^h(n-1),\delta Y_n}\Phi(\tilde{X}^h(1),\delta Y_n))\right]_1^2$$

converges to 0 uniformly in n as $h \to 0$.

The following extension of Theorem 6.2 can now be stated.

THEOREM 6.3. *Theorem* 5.2 *holds for the filter defined by* (6.3) *under* A4.1, A4.2, *and* A6.2.

Remark on A6.1 *and* A6.2. Assumptions A6.1 and A6.2 are seen to hold for many examples of interest. Perhaps most commonly, the desired change of measure is such that

$$\sup_{h,n} \sup_{x,y \in G} \frac{dP^h_{n-1}}{dM^{h,n}}(x,y) < \infty.$$

In such a case A6.1 is clearly satisfied. Typically the importance sampling is with respect to initial condition only, and Π^h_{n-1} is a discrete probability measure (see Example 6.4). In such situations a natural choice of $M^{h,n}$ is the measure obtained by multiplying the probabilities of the points by appropriate weights which are uniformly bounded in h and n. An important case where the uniform boundedness does not hold, but where A6.1 can be verified, is described in Example 6.4. Note that if the measure $M^{h,n}$ only depends upon $\{\delta Y_j; j \le n-1\}$, then A6.1 can be simplified to

$$\sup_{h,n} E \frac{dP^h_{n-1}}{dM^{h,n}}(\tilde{X}^h(0), \tilde{X}^h(1)) < \infty.$$

Furthermore, for all these situations, if the sampling is done from $M^{h,n}$, using some variance reduction scheme of the types illustrated in section 5 rather than in a purely i.i.d. manner, then A6.2 holds.

As stated above, of special interest is the case where the importance sampling is with respect to the initial condition only. To illustrate this case, we consider the special case of Example 5.6 of section 5, where the sampling filter without importance sampling is given by (5.1) with $\tilde{X}^{h,n,l}(\cdot)$ defined via (5.12b). Then P^h_{n-1} can be identified with the measure $\Pi^h(n-1) \times P^h_\zeta$ in that this measure determines that of $(\tilde{X}^h(0), \tilde{X}^h(1))$. Now, let us write $M^{h,n}$ in a similar manner; namely, $M^{h,n} = M^{h,n}_0 \times P^h_\zeta$, where $M^{h,n}_0$ is a random measure on G. In this case, the measure transformation is over the initial condition only and we have

$$L^{h,n} = \frac{d\Pi^h(n-1)}{dM^{h,n}_0} \quad \text{and} \quad L^{h,k,n} = \frac{d\Pi^h(n-1)}{dM^{h,n}_0}(\tilde{X}^{h,k,n}(0)).$$

In the next example, we see that the idea of applying importance sampling to the initial condition can be enhanced by the use of the next observation δY_n, to determine the $M^{h,n}_0$.

Example 6.4 (observation dependent importance sampling). Appropriate measure transformations $M^{h,n}$ (or $M^{h,n}_0$ as defined at the end of the above example) can improve the estimates quite a bit [5, 29]. Better $M^{h,n}$ will depend on the next observation δY_n, and we will illustrate the point via the signal model of Example 5.6 of section 5, where $\tilde{X}^h(\cdot)$ is defined by (5.12). Data and examples of such a procedure can be found in [29]. Again, A6.1 and the mutual absolute continuity are the only conditions (in addition to A4.1 and A4.2) that need to be verified for Theorem 5.1 to hold for the $\Pi^h(\cdot)$ defined in this example. Keep in mind that we are illustrating only one type of procedure, and even that has many variations. Consider the following procedure.

Suppose that $\Pi^h(n-1)$ is concentrated on the v^h points $\{x^{h,l,n}, l \le v^h\}$, with the lth point having probability $\Pi^h_l(n-1)$. The paths emanating from some of the $x^{h,l,n}$ might be "poor" predictors of the observation δY_n in the sense that the conditional density

$$p\{\delta Y_n | X(n) = b_h(x^{h,l,n}, \xi^h)\}$$

is very small with a high probability. For some other points $x^{h,l,n}$, this value might be high with a reasonable probability. It seems reasonable to explore the paths emanating from the more promising initial points more fully, if this can be done without (asymptotically) biasing the procedure. The main problem is that we do not know (apart from the values of $\Pi^h(n-1)$) which are the more promising points and how much more promising they are. The "weights" for the importance sampling are to be determined by an exploratory sampling procedure, after which the sampling to get the next estimate $\Pi^h(n)$ will be done. This "double" sampling explains the complexity of the following algorithm. Nevertheless, such algorithms are sometimes useful [29] in that the total computation for a filter with comparable accuracy can be less than what is needed for a direct method such as that in Example 5.1.

The procedure starts by getting a "typical" value of $b_h(x^{h,l,n}, \xi^h)$. The word "typical" is used loosely here. The aim is to get some preliminary approximation to the "predictive values" of the trajectories emanating from the point, given the next observation. This "typical value" might be an estimate of the mean value, or it might be a simple sample value or an average of several sample values. We call these the "indicator" values and denote them by $\hat{X}^{h,l,n}(1), l \le v^h$. Then the "predictive power" of this indicator value is computed, and the associated weights used to get the importance sampling measure for the final computation of $\Pi^h(n)$. The details follow in algorithmic form.

(1) Let $\hat{X}^{h,l,n}(1)$ $(l \le v^h)$ denote an "indicator" quantity, which (hopefully) is highly correlated with the "value" of sampling with initial condition $x^{h,l,n}$. (The points for which we get such an "indicator" quantity might also be chosen by some sampling procedure.)

(2) Compute the conditional Gaussian density $p(\delta Y_n | X(n) = \hat{X}^{h,l,n}(1))$, and define the "conditional likelihood" of the observation

$$p^{h,l,n} = \frac{p(\delta Y_n | X(n) = \hat{X}^{h,l,n}(1))}{\sum_k \Pi_k^h(n-1) p(\delta Y_n | X(n) = \hat{X}^{h,k,n}(1))}, \qquad l \le v_h. \tag{6.7}$$

The numerator of $p^{h,l,n}$ up to a normalizing factor is $R(\hat{X}^{h,l,n}(1), \delta Y_n)$. Note that $p^{h,l,n}$ is not a probability. If the numerator in (6.7) is the same for all points, then $p^{h,l,n} = 1$ for all l.

(3) Sample $m^{h,n} \ge v^h$ times (with replacement) from the set

$$\{x^{h,l,n}, l \le v^h\}$$

with weights proportional to the $p^{h,l,n}\Pi_l^h(n-1)$. Note that $\sum_l p^{h,l,n}\Pi_l^h(n-1) = 1$. This yields a set which we denote by

$$\{\overline{x}^{h,l,n}, l \le m^{h,n}\}.$$

It is found in practice that the performance is often better if $m^{h,n}$ is several times v^h. This tends to assure a better spread for the support of the conditional distribution.

(4) Sample $\{\zeta^{h,k,n}, k \le m^{h,n}\}$ from P_ζ^h and compute

$$\overline{X}^{h,l,n}(1) = b_h(\overline{x}^{h,l,n}, \zeta^{h,k,n}), \qquad k \le m^{h,n}. \tag{6.8}$$

(5) If $v^h = m^{h,n}$, then set $\tilde{X}^{h,l,n}(1) = \overline{X}^{h,l,n}(1) = x^{h,l,n+1}$. If $v^h < m^{h,n}$, then resample at random (with replacement) v^h times from $\{\overline{X}^{h,l,n}(1), l \le m^{h,n}\}$ to get the set $\{\tilde{X}^{h,k,n}(1), k \le v^h\}$, and set $x^{h,k,n+1} = \tilde{X}^{h,k,n}(1)$.

In this procedure, the measure $M^{h,n}$ was defined by defining $M_0^{h,n}$ via the weight

$$M_{0,k}^{h,n} = p^{h,k,n}\Pi_l^h(n-1) \tag{6.9}$$

that it puts on the initial point $x^{h,l,n}$. Finally, we use the filter defined by (6.3) with

$$L^{h,l,n} = \frac{1}{p^{h,l,n}}. \tag{6.10}$$

The set of likelihood functions clearly satisfy A6.1, and the corresponding measures satisfy the mutual absolute continuity requirement.

Note on numerical data. Various forms of the suggested algorithmic forms have been simulated on the four-dimensional problem of [4]. The model represented a ship moving in a region limited by a shoreline, and this led to the nonlinearities. The observations were angle only, and there was little data in each observation. We tried to make comparisons with essentially the same program time used for the various cases. While no effort was made to optimize the performance of any algorithm, certain conclusions can be drawn. Wherever possible, use an "integration" method. They were more efficient with "moderately" limited computation time and generally performed well. Depending on the parameters, the Monte Carlo methods might be close. Variance reduction methods definitely help. The observation dependent importance sampling methods of the type of Example 6.4 further reduced the error variances by about 10%. Now, drop the equal time constraints. There were lower bounds to the (experimental) error variances for the integration algorithm, as its "order" increased. These bounds are usually larger than the true variances and were larger than error variances for the Monte Carlo methods for large numbers of samples. Owing to the large number of samples (hence pseudorandom numbers) needed for the Monte Carlo, a considerable effort needs to be put into the coding to achieve realistic execution times. Owing to the several layers of sampling used in the observation dependent importance sampling method, combinations of Monte Carlo and integration methods often improved on the purely Monte Carlo, but there is no room for the many details here.

7. The continuous time problem. In this section we will study the continuous time analogues of the various random sampling and combined random sampling-integration algorithms studied in sections 5 and 6. Our basic filtering model will be that in section 2. To fix ideas, we will begin by indicating the form of the approximating filter for the case where the random samples are mutually independent and identically distributed, analogously to what was done in (5.1). We will then consider a general form of the approximating filter which would cover not only the case of such i.i.d. samples but also various variance reduction schemes and importance sampling algorithms of the type studied in Examples 5.3, 6.1, and 6.4.

7.1. Example and motivation. In typical uses of the approximation (2.7), the approximating signal process $\tilde{X}^h(\cdot)$ is a piecewise constant interpolation of a discrete time process. One good example is the Markov chain approximation such as used in Example 5.7. Most current applications seem to use such Markov chain based approximations, whether they are of the explicit forms discussed in [24] or other forms which satisfy the required local consistency property, e.g., based on approximate solutions to the Fokker–Planck equation over small intervals.

Following the idea and terminology of Example 5.7, let X_n^h denote the underlying Markov chain, and let $\tilde{X}^h(\cdot)$ denote its piecewise constant interpolation, with

interpolation interval δ_h. Then one can use the approximating filter (2.7). Since the approximating process $\tilde{X}^h(\cdot)$ is piecewise constant, $R(\tilde{X}^h_{0,t}, Y_{0,t})$ equals

$$\begin{aligned}(7.1)\qquad &\exp\sum_{k=0}^{[t/\delta_h]-1}\left[g'(X^h_k)\left[Y(k\delta_h+\delta_h)-Y(k\delta_h)\right]-\frac{\delta_h}{2}\left|g(X^h_k)\right|^2\right]\\ &\times\exp\left[g'(X^h_{[t/\delta_h]})\left[Y(t)-Y([t/\delta_h]\delta_h)\right]-\tfrac{t-[t/\delta_h]\delta_h}{2}|g(X^h_{[t/\delta_h]})|^2\right].\end{aligned}$$

However, in practice observations cannot truly be taken continuously, and one would incorporate the observation into the filter at the discrete time instants $n\delta_h$ only. In fact, for such nonlinear problems the notion of continuous updating seems to be a mathematical fiction, although the times between updating might be very small. Thus one would approximate $R(\tilde{X}^h_{0,t}, Y_{0,t})$ by $R^h(\tilde{X}^h_{0,t}, Y_{0,t})$ which is defined by the following (piecewise constant) expression

$$(7.2)\qquad \exp\sum_{k=0}^{[t/\delta_h]-1}\left[g'(X^h_k)\left[Y(k\delta_h+\delta_h)-Y(k\delta_h)\right]-\frac{\delta_h}{2}\left|g(X^h_k)\right|^2\right].$$

Whatever the form of $\tilde{X}^h(\cdot)$, whether it is obtained explicitly as an interpolation of a discrete time chain X^h_n or not, the samples $\tilde{X}^h(n\delta_h)$ are always a Markov chain. For notational simplicity, we always write $\tilde{X}^h(n\delta_h) = X^h_n$.

A random sampling algorithm. The above paragraph argues that it is not a restriction to require the approximating filter process to be piecewise constant. This logic also holds for algorithms based on random sampling. It holds even if some higher-order interpolation method (say that of Milstein or other types used in [17]) is used, since even then we would use the interpolation to get a better approximation to the signal process at discrete time instants $n\delta_h$. Thus, in the approximate filters that are considered here, we approximate the conditional distribution at the instants $n\delta_h$, and we suppose that the filter is constant on $[n\delta_h, n\delta_h+\delta_h)$.

Since the estimate of the conditional distribution will be updated at each $n\delta_h$, we could try to duplicate the various methods in Examples 5.1 to 6.4, with the basic interval being δ_h, and then prove convergence as $\delta_h \to 0$. The resampling at the beginning of each interval in the various examples exploited the updated information to get more sample trajectories from the points which seemed to be more likely, in view of the information in the observations. But random sampling also loses information. There is always a chance that the better points will not be sampled. This chance is reduced as v^h increases. When resampling occurs very frequently, say at each time instant $n\delta_h$, the procedure can degenerate very fast as $\delta_h \to 0$, unless v^h increases fast enough as $\delta_h \to 0$. One can quantify such a statement. But it is also a common observation in simulations, including the ones that we have carried out. Reference [6] resamples at each discrete interval, in a "minimum variance" way, but v^h must grow as $1/\delta_h$. Such a rapid increase in v^h is an inefficient use of the computational resources, especially in view of the fact that the estimates of the conditional distribution do not change much in small intervals.

Owing to the above observations, we take the following "practical" approach. Divide time into subunits of (small, but fixed—they do not go to zero) length ϵ, and suppose that $\epsilon/\delta_h = n_h$ is always an integer. We resimulate the $\tilde{X}^h(\cdot)$ each ϵ units of time, although the observations are incorporated at the instants $n\delta$.

The general model given below is motivated by the ideas of Examples 5.4 to 6.4, and we give an analogue of condition A5.3 (namely, conditions A7.1, A7.2) which

covers many cases of interest. But, for specificity, let us first consider the case where the process $\tilde{X}^h(\cdot)$ satisfies the consistency assumption A2.1 and the samples taken on $[n\epsilon, n\epsilon+\epsilon)$ are mutually independent and independent of $Y(\cdot)$ given their initial distribution $\Pi^h(n\epsilon)$. Denote these samples by $\{\tilde{X}^{h,l,n}(\cdot), l \le v^h\}$. Define

$$R^h(\tilde{X}_{0,s}, Y_{n\epsilon,n\epsilon+s}) \tag{7.3}$$
$$= \exp \sum_{k=0}^{[s/\delta_h]-1} \left[g'(X_k^h)\left[Y(n\epsilon+k\delta_h+\delta_h) - Y(n\epsilon+k\delta_h)\right] - \frac{\delta_h}{2}\left|g(X_k^h)\right|^2 \right].$$

Thus $R^h(x_{0,s}, Y_{a,b})$ differs from $R(x_{0,s}, Y_{a,b})$ only in that in the former the function $x(\cdot)$ is replaced by the piecewise constant function with value $x(k\delta_h)$ on $[k\delta_h, k\delta_h+\delta_h)$. The basic approximating filter based on random sampling for $s = q\delta_h < \epsilon$ where q is an integer is

$$\langle \Pi^h(n\epsilon+s), \phi\rangle = \frac{\sum_{l=1}^{v^h} \phi(\tilde{X}^{h,l,n}(s)) R^h(\tilde{X}_{0,s}^{h,l,n}, Y_{n\epsilon,n\epsilon+s})/v^h}{\sum_{l=1}^{v^h} R^h(\tilde{X}_{0,s}^{h,l,n}, Y_{n\epsilon,n\epsilon+s})/v^h}. \tag{7.4}$$

This is just the "continuous time" analogue of (5.1).

As stated earlier, we will suppose that $\Pi^h(\cdot)$ is constant on the intervals $[q\delta_h, q\delta_h+\delta_h)$. Alternatively, if desired, we can interpolate, and one natural interpolation would use the form (7.4), but with the term

$$g'(X^h_{[s/\delta_h]})\left[Y(n\epsilon+s) - Y(n\epsilon+[s/\delta_h]\delta_h)\right] - \frac{s-[s/\delta_h]\delta_h}{2}|g(X^h_{[s/\delta_h]})|^2 \tag{7.5}$$

added to the sum in (7.3). The treatment of both forms is nearly identical.

As in the discrete time case, we are interested in random sampling and random sampling-integration algorithms which are more general than (7.4), with (conditional) i.i.d. samples. We would like to allow the possibility of incorporating variance reduction methods as in Example 5.3, or perhaps the sampling can be guided using some importance sampling scheme as in Examples 6.1 and 6.4. We might even be interested in algorithms which are part integration and part random sampling, say of the form discussed in Examples 5.6 and 5.7. In view of this, we work with the following general form of the approximate filter, which is our continuous time analogue of the discrete time filter defined via (5.10). The chosen general structure is motivated by the same considerations which led to (5.10) and A5.11, the desire to include many types of approximations of interest under one roof, with a general assumption which can be verified in particular cases of interest, analogously to what was done for the discrete time case. The conditions A7.1 and A7.2 hold for the independent samples case under A2.1.

Let δ_h, ϵ, n_h be as above. Define the approximating filter $\Pi^h(\cdot)$ as follows. Having defined $\Pi^h(t)$ for $0 \le t \le n\epsilon$, let $P^{h,n}_{\Pi^h(n\epsilon)}$ be conditionally independent of $\{Y_t - Y_{n\epsilon}; t \ge n\epsilon\}$ given $\Pi^h(n\epsilon)$. Define, for $1 \le j \le n_h$,

$$\langle \Pi^h(n\epsilon+j\delta_h), \phi\rangle = \frac{\int \phi(x(j\delta_h)) R^h(x_{0,j\delta_h}, Y_{n\epsilon,n\epsilon+j\delta_h}) dP^{h,n}_{\Pi^h(n\epsilon)}(x(\cdot))}{\int R^h(x_{0,j\delta_h}, Y_{n\epsilon,n\epsilon+j\delta_h}) dP^{h,n}_{\Pi^h(n\epsilon)}(x(\cdot))}. \tag{7.6}$$

For points in $[n\epsilon, (n+1)\epsilon)$ not of the form $n\epsilon + j\delta_h$, the filter is defined via the piecewise constant and right continuous interpolation. In the independent sample

case (7.4), $P^{h,n}_{\Pi^h(n\epsilon)}$ is the occupation measure defined analogously to the way it was defined above (5.9) for the discrete time case. We assume that the family $\{P^{h,n}_{\Pi^h(n\epsilon)}\}$ satisfies A7.1 and A7.2 below. Given the very general structure allowed for the filter, some condition such as A7.2 is needed for the continuous time problem in order to avoid simulated processes that are "wild." Under A2.1, condition A7.2 holds for the i.i.d. case illustrated in (7.4), since there each $\tilde{X}^{h,l,n}(\cdot)$ is a replica of the $\tilde{X}^h(\cdot)$ of A2.1 with initial conditions in the compact set G.

A7.1. For every bounded and continuous real valued function $\Phi(\cdot)$ of $x(\cdot)$ on the interval $[0,\epsilon]$ and which depends on $x(\cdot)$ only at a finite number of points,

$$\lim_{h\to 0}\sup_n E\left[\int \Phi(x(\cdot))dP^{h,n}_{\Pi^h(n\epsilon)}(x(\cdot)) - E_{\Pi^h(n\epsilon)}\Phi(\tilde{X}(\cdot))\right]^2_1 = 0.$$

We will impose another condition on the $P^{h,n}_{\Pi^h(n\epsilon)}$. For motivation, consider the case of independent samples discussed above. For $\mu > 0, \delta > 0$, define the set of paths

$$C^\delta_\mu = \left\{x(\cdot) : \sup_{s\le\delta, t+s\le\epsilon, t\le\epsilon} |x(t+s) - x(t)| \ge \mu\right\}.$$

Then our second condition is the following.

A7.2.

$$\lim_{\delta\to 0}\limsup_h \sup_n EP^{h,n}_{\Pi^h(n\epsilon)}(C^\delta_\mu) = 0 \quad \text{for each} \quad \mu > 0. \tag{7.7}$$

Remark on A7.1 *and* A7.2. The assumption A7.1 is similar to the assumption (A5.3) made for the discrete time problem. For the cases where $P^{h,n}_{\Pi^h(n\epsilon)}$ is constructed via i.i.d. sampling or some variance reduction scheme, it is verified as easily as in the discrete time case. Also, if this measure is constructed via a combination of sampling and integration in manner similar to Example 5.7, where the observations are incorporated at each time $n\delta_h$, one can again verify A7.1 by using the consistency condition A2.1. The assumption A7.1 continues to hold if the samplings of the initial conditions $\tilde{X}^{h,l,n}(0)$ are determined by importance sampling as in Example 6.4. Assumption A7.2 simply states that for small h, the sampled paths don't change much in the mean over small intervals, uniformly in n. For all the above examples this assumption can be verified by using the consistency condition A2.1 on the $\tilde{X}^h(\cdot)$ process.

The following lemma will be used in some of the tightness arguments used below.

LEMMA 7.1 [19, Theorem 2.7b]. *Let $\{Z_n(\cdot), n\}$ be a family of processes with paths in the Skorohod space $D[S_0; 0, \infty)$, where S_0 is a complete and separable metric space with metric $\gamma(\cdot)$. For each $\delta > 0$ and each t in a dense set, let there be a compact set $S_{\delta,t} \subset S_0$ such that*

$$\sup_n P\{Z_n(t) \notin S_{\delta,t}\} \le \delta.$$

Let $\mathbb{F}^n_t$ denote the minimal σ-algebra which measures $\{Z_n(u), u \le t\}$, and let $\mathcal{T}_n(T)$ denote the set of $\mathbb{F}^n_t$-stopping times which are less than $T > 0$. Suppose that for each T

$$\lim_{\delta\to 0}\limsup_n \sup_{\tau\in\mathcal{T}_n(T)} E\left[\gamma\left(Z_n(\tau+\delta), Z_n(\tau)\right) \wedge 1\right] = 0.$$

Then $\{Z^n(\cdot)\}$ is tight.

THEOREM 7.2. *Let* $(X(\cdot), Y(\cdot))$ *be as in section* 2. *Assume* A3.1, A7.1, *and* A7.2. *Then the conclusions of Theorem* 3.1 *hold for the approximate filter* $\Pi^h(\cdot)$ *defined as above.*

Proof. Many of the details of the proof are the same as in the proof of Theorem 3.1 and its variations used in the proof of Theorem 5.2 and its successors for the discrete time case. The key differences are in the proof of tightness of $\{Q^{h_k,T_k}(\cdot); k \geq 1\}$ for any sequences $h_k \to 0, T_k \to \infty$, and the proof of the representation (3.8), and we concentrate on these points. In the discrete time theorems, the tightness of $\{Q^{h_k,T_k}(\cdot); k \geq 1\}$ was essentially obvious due to the compactness of G. There was no issue of "path properties," in showing the tightness due to the discrete time parameter. In the current continuous time case, we need to deal with the path properties. Owing to the use of the weak topology, it is enough to prove the tightness of the set

$$\{\langle \Pi^{h_k}(t_k + \cdot), \phi\rangle; h_k, T_k\}$$

for each bounded and continuous real-valued function $\phi(\cdot)$ on G.

In proving the tightness of the above family, we use the criterion in the lemma. The main step is establishing that for each $\phi(\cdot)$, as above,

$$\lim_{\delta\to 0} \limsup_{h\to 0} \sup_t \sup_{\tau\in\mathcal{T}^{h,t}(\rho)} E\left|\langle \Pi^h(t+\tau+\delta), \phi\rangle - \langle \Pi^h(t+\tau), \phi\rangle\right|_1^2 = 0, \tag{7.8}$$

where $\mathcal{T}^{h,t}(\rho)$ denotes the set of stopping times bounded by ρ for the process $\Pi^h(t+\cdot)$. We can assume without loss of generality that δ in the above expression is less than ϵ. Thus $t+\tau$ and $t+\tau+\delta$ are either in the same interval of the form $[j\epsilon, (j+1)\epsilon)$ or they are in adjacent such intervals. This observation along with an application of a triangle inequality shows that, in order to prove (7.8), it suffices to prove that

$$\limsup_{h\to 0} \sup_j \sup_{0\leq K_1\leq K_2\leq n_h, |K_1-K_2|\delta_h\leq\delta} E\left|\langle \Pi^h(j\epsilon + K_1\delta_h), \phi\rangle - \langle \Pi^h(j\epsilon + K_2\delta_h), \phi\rangle\right|_1^2 \tag{7.9}$$

converges to 0 as $\delta \to 0$. Note that $j\epsilon + K_1\delta_h$ and $j\epsilon + K_2\delta_h$ are both in the interval $[j\epsilon, j\epsilon + \epsilon]$.

Now we bound the above expectation by the sum of the following three terms. The first two terms are, for $i = 1, 2$,

$$\limsup_{h\to 0} \sup_j \sup_{0\leq K_i\delta_h\leq\epsilon} F_1(j, h, K_i\delta_h), \tag{7.10}$$

where

$$F_1(j,h,K_i\delta_h) = E\left[\langle \Pi^h(j\epsilon + K_i\delta_h), \phi\rangle - \frac{E_{\Pi^h(j\epsilon), Y_{j\epsilon, j\epsilon+K_i\delta_h}} \phi(\tilde{X}(K_i\delta_h)) R^h(\tilde{X}_{0,K_i\delta_h}, Y_{j\epsilon,j\epsilon+K_i\delta_h})}{E_{\Pi^h(j\epsilon), Y_{j\epsilon,j\epsilon+K_i\delta_h}} R^h(\tilde{X}_{0,K_i\delta_h}, Y_{j\epsilon,j\epsilon+K_i\delta_h})}\right]_1^2.$$

The third term is

$$\limsup_{h\to 0} \sup_j \sup_{0\leq K_1\delta_h\leq K_2\delta_h\leq\epsilon, |K_1-K_2|\delta_h\leq\delta} F_2(j, h, K_1\delta_h, K_1\delta_h), \tag{7.11}$$

where

$$F_2(j,h,K_1\delta_h,K_1\delta_h) = E\left[\frac{E_{\Pi^h(j\epsilon),Y_{j\epsilon,j\epsilon+K_1\delta_h}}\phi(\tilde{X}(K_1\delta_h))R^h(\tilde{X}_{0,K_1\delta_h},Y_{j\epsilon,j\epsilon+K_1\delta_h})}{E_{\Pi^h(j\epsilon),Y_{j\epsilon,j\epsilon+K_1\delta_h}}R^h(\tilde{X}_{0,K_1\delta_h},Y_{j\epsilon,j\epsilon+K_1\delta_h})} - \frac{E_{\Pi^h(j\epsilon),Y_{j\epsilon,j\epsilon+K_2\delta_h}}\phi(\tilde{X}(K_2\delta_h))R^h(\tilde{X}_{0,K_2\delta_h},Y_{j\epsilon,j\epsilon+K_2\delta_h})}{E_{\Pi^h(j\epsilon),Y_{j\epsilon,j\epsilon+K_2\delta_h}}R^h(\tilde{X}_{0,K_2\delta_h},Y_{j\epsilon,j\epsilon+K_2\delta_h})}\right]_1^2.$$

In dealing with (7.11), owing to the properties of the $|\cdot|_1^2$ metric defined by (4.14), we need only work with the differences of the numerators and denominators separately. Then, (7.11) is easily dealt with using the continuity property of $\tilde{X}(\cdot)$. In particular, it follows from the fact that for any bounded and continuous function $\phi(\cdot)$

$$\lim_{\delta\to 0}\sup_{|K_1-K_2|\delta_h\le\delta}\sup_{\pi} E\Big[E_{\pi,Y_{j\epsilon,j\epsilon+K_1\delta_h}}\phi(\tilde{X}(K_1\delta_h))R^h(\tilde{X}_{0,K_1\delta_h},Y_{j\epsilon,j\epsilon+K_1\delta_h}) - E_{\pi,Y_{j\epsilon,j\epsilon+K_2\delta_h}}\phi(\tilde{X}(K_2\delta_h))R^h(\tilde{X}_{0,K_2\delta_h},Y_{j\epsilon,j\epsilon+K_2\delta_h})\Big]_1^2 = 0, \tag{7.12}$$

where $K_i\delta_h \le \epsilon$.

Now we consider (7.10). By the definition of $\Pi^h(j\epsilon)$ in terms of $P^{h,j}_{\Pi^h(j\epsilon)}$ in (7.6), the first term inside the bars in (7.10) equals

$$\frac{\int \phi(x(K_i\delta_h))R^h(x_{0,K_i\delta_h},Y_{j\epsilon,j\epsilon+K_i\delta_h})dP^{h,j}_{\Pi^h(j\epsilon)}(x(\cdot))}{\int R^h(x_{0,K_i\delta_h},Y_{j\epsilon,j\epsilon+K_i\delta_h})dP^{h,j}_{\Pi^h(j\epsilon)}(x(\cdot))}. \tag{7.13}$$

Again, we need only work with the differences of the numerators of the right-hand term inside the bars in (7.10) and that in (7.13) for arbitrary bounded and continuous $\phi(\cdot)$.

The proof that the limit of (7.10) as $\delta \to 0$ is zero will use an approximation method. For small $\Delta > 0$, with ϵ an integral multiple of Δ, define $R^\Delta(x_{0,s},Y_{a,a+s})$ by

$$\exp\left\{\sum_{i:i\Delta<s} g'(x(i\Delta))\left[Y(a+i\Delta+\Delta)-Y(a+i\Delta)\right] - \frac{\Delta}{2}\sum_{i:i\Delta<s}|g(x(i\Delta))|^2\right\}.$$

For each h, define

$$A^{h,\Delta} = \sup_{K_i\delta_h\le\epsilon}\sup_j\sup_\pi E_\pi[R^h(\tilde{X}_{0,K_i\delta_h},Y_{j\epsilon,j\epsilon+K_i\delta_h}) - R^\Delta(\tilde{X}_{0,K_i\delta_h},Y_{j\epsilon,j\epsilon+K_i\delta_h})]_1^2.$$

For each $\rho > 0$, there is $\Delta_0 > 0$ such that for $\Delta < \Delta_0$ and small $h > 0$ we have $A^{h,\Delta} \le \rho$. Define

$$B^{h,\Delta} = \sup_{K_i\delta_h\le\epsilon}\sup_j E\int\left[R^h(x_{0,K_i\delta_h},Y_{j\epsilon,j\epsilon+K_i\delta_h}) - R^\Delta(x_{0,K_i\delta_h},Y_{j\epsilon,j\epsilon+K_i\delta_h})\right]_1^2 dP^{h,j}_{\Pi^h(j\epsilon)}(x(\cdot)). \tag{7.14}$$

By A7.2, for each $\rho > 0$ there is $\Delta_0 > 0$ such that for $\Delta < \Delta_0$, and small $h > 0$, $B^{h,\Delta} \le \rho$. This assertion is proved as follows. Define $x^\Delta(t) = x(i\Delta)$ for $t \in [i\Delta, i\Delta+\Delta)$. To

prove the assertion, it is sufficient to show that

$$\limsup_h \sup_j \sup_{K_i \le n_h} \int \left[\sum_{l=0}^{K_i} \left[g(x(l\delta_h)) - g(x^\Delta(l\delta_h)) \right]' \left[Y(l\delta_h + \delta_h) - Y(l\delta_h) \right] \right]^2 d[EP^{h,j}_{\Pi^h(j\epsilon)}(x(\cdot))] \tag{7.15}$$

is arbitrarily small if Δ is small enough. By A7.2, we can suppose that the difference of the $g-$terms in the bracket in (7.15) is as small as we wish, and this implies the assertion.

Thus, to show that the limit of (7.10) is zero as $\delta \to 0$, it is sufficient to show that

$$\begin{aligned} &\lim_{\Delta\to 0} \limsup_{h\to 0} \sup_j \sup_{0\le K_i\delta_h \le \epsilon} \\ &\quad E \Bigg| \int \phi(x(K_i\delta_h)) R^\Delta(x_{0,K_i\delta_h}, Y_{j\epsilon, j\epsilon + K_i\delta_h}) dP^{h,j}_{\Pi^h(j\epsilon)}(x(\cdot)) \\ &\qquad - E_{\Pi^h(j\epsilon), Y_{j\epsilon, j\epsilon+K_i\delta_h}} \phi(\tilde{X}(K_i\delta_h)) R^\Delta(\tilde{X}_{0,K_i\delta_h}, Y_{j\epsilon, j\epsilon+K_i\delta_h}) \Bigg|_1^2 = 0. \end{aligned} \tag{7.16}$$

By A7.2, it is sufficient to show (7.16) if the $K_i\delta_h$ in the $\phi(x(K_i\delta_h))$ and $\phi(\tilde{X}(K_i\delta_h))$ are replaced by the closest integral multiple of Δ for small Δ. Now, all the $K_i\delta_h$ in (7.16) are integral multiples of Δ for some fixed Δ. Hence, the $\sup_{0\le K_i\delta_h\le\epsilon}$ in (7.16) is redundant and can be dropped. We would like to use A7.1 at this point. But A7.1 holds only for each function $\Phi(\cdot)$. In (7.16), Δ is fixed, and we can suppose without loss of generality that $\epsilon = k_0\Delta$ for some integer k_0. Given any small $\rho > 0$, there is a bounded set B_ρ such that the values of $\{Y(j\epsilon + t) - Y(j\epsilon), t \le \epsilon\}$ are confined to B_ρ with at least probability $1 - \rho$ for all j. Because of this and the continuity of $R^\Delta(\cdot)$, we need only verify (7.16) for some finite set of values of the $Y-$variables. Due to this and to the fact that $k_0 < \infty$, we need only evaluate (7.16) for each $Y(\cdot)$ and $K_i\delta_h \le \epsilon$ being some arbitrary multiple of Δ. Then A7.1 can be used and guarantees (7.16). This completes the proof of tightness of $\{Q^{h_k,T_k}(\cdot), k \ge 1\}$.

We now prove the representation in (3.8). The general scheme used in Theorem 5.2 for this characterization will be followed. Let $\psi(\cdot) = (x(\cdot), \pi(\cdot), y(\cdot), b(\cdot))$ denote the canonical paths of the signal process, the conditional probability process, the observation process, and the observation noise process. They are connected by $y(t) = \int_0^t g'(x(s))ds + b(t)$. For arbitrary bounded and continuous $\phi(\cdot)$, arbitrary $\psi(\cdot)$, and times t, s, define the function $A(\psi(\cdot); t, s))$ by

$$A(\psi(\cdot); t, s) = \langle \pi(t+s), \phi \rangle - \frac{E_{\{\pi(t), y_{t,t+s}\}}[\phi(\tilde{X}_{0,s}) R(\tilde{X}_{0,s}, y_{t,t+s})]}{E_{\{\pi(t), y_{t,t+s}\}} R(\tilde{X}_{0,s}, y_{t,t+s})}. \tag{7.17}$$

Recall the definition of $\Psi^\omega(\cdot)$ from Theorem 3.1. We will also use other notations from section 3. The aim of the proof of Theorem 3.2 in [3], which is our Theorem 3.1, was to show that, for almost all ω and all t, s,

$$A(\Psi^\omega(\cdot); t, s) = 0, \qquad \text{w.p.1}, \tag{7.18}$$

which implies (3.8). In fact it suffices to consider $s \le \epsilon$. Hereafter we will consider only such values of s without any further comment.

The statement in (7.18) will be proved by showing that

$$0 = E \int Q^{\omega}(d\psi)\,[A(\psi(\cdot);t,s)]_1^2\,. \tag{7.19}$$

The prelimit form of the right side of (7.19) is

$$E \int Q^{h,T}(d\psi)\,[A(\psi(\cdot);t,s)]_1^2\,,$$

which, by the definition of $Q^{h,T}(\cdot)$, equals

$$\frac{1}{T}\int_0^T E\left[A(\Psi^h(\cdot);u+t,s)\right]_1^2 du, \tag{7.20}$$

where

$$A(\Psi^h(\cdot);t,s) = \langle \Pi^h(t+s),\phi\rangle - \frac{E_{\{\Pi^h(t),Y_{t,t+s}\}}[\phi(\tilde{X}(s))R(\tilde{X}_{0,s},Y_{t,t+s})]}{E_{\{\Pi^h(t),Y_{t,t+s}\}}R(\tilde{X}_{0,s},Y_{t,t+s})}. \tag{7.21}$$

In order to show (7.19), it suffices to show that

$$\lim_h \sup_t E[A(\Psi^h(\cdot);t,s)]_1^2 \to 0. \tag{7.22}$$

Furthermore, in view of the properties of the $\tilde{X}(\cdot)$ process and the tightness of the set $\{\Pi^h(t);h,t\}$, to prove (7.22) it is clearly sufficient to show that

$$\sup_t E\left[\langle \Pi^h(t+s),\phi\rangle - \frac{E_{\{\Pi^h(t),Y_{t,t+s}\}}[\phi(\tilde{X}(s))R^h(\tilde{X}_{0,s},Y_{t,t+s})]}{E_{\{\Pi^h(t),Y_{t,t+s}\}}R^h(\tilde{X}_{0,s},Y_{t,t+s})}\right]_1^2 \to 0 \tag{7.23}$$

as $h \to 0$.

Since $s < \epsilon$, the points t and $t+s$ are either in the same subinterval of the form $[j\epsilon,(j+1)\epsilon]$ or are in adjacent intervals of this form. We consider below the case of adjacent intervals. The arguments required for the same interval case are simpler versions of the former case and thus are omitted. Let now $t \in [j\epsilon + i\delta_h, j\epsilon + i\delta_h + \delta_h)$ and $t+s \in [(j+1)\epsilon + i'\delta_h, (j+1)\epsilon + i'\delta_h + \delta_h)$. Showing (7.23) is equivalent to proving that, for each s,

$$E\left[\langle \Pi^h(t+s),\phi\rangle - \frac{E_{\{\Pi^h(j\epsilon+i\delta_h),Y_{j\epsilon+i\delta_h,t+s}\}}\left[\phi(\tilde{X}(\alpha\delta_h))R^h(\tilde{X}_{0,\alpha\delta_h},Y_{j\epsilon+i\delta_h,(j+1)\epsilon+i'\delta_h})\right]}{E_{\{\Pi^h(j\epsilon+i\delta_h),Y_{j\epsilon+i\delta_h,t+s}\}}R^h(\tilde{X}_{0,\alpha\delta_h},Y_{j\epsilon+i\delta_h,(j+1)\epsilon+i'\delta_h})}\right]_1^2 \tag{7.24}$$

converges to 0 as $h \to 0$, uniformly in t, where $\alpha = n_h + i' - i$. Thus, $|\alpha\delta_h - s| \le \delta_h$.

The expectation in (7.24) can be bounded above by the sum of

$$E\left[\langle \Pi^h(t+s),\phi\rangle - \frac{E_{\{\Pi^h((j+1)\epsilon),Y_{(j+1)\epsilon,t+s}\}}\left[\phi(\tilde{X}(i'\delta_h))R^h(\tilde{X}_{0,i'\delta_h},Y_{(j+1)\epsilon,(j+1)\epsilon+i'\delta_h})\right]}{E_{\{\Pi^h((j+1)\epsilon),Y_{(j+1)\epsilon,t+s}\}}R^h(\tilde{X}_{0,i'\delta_h},Y_{(j+1)\epsilon,(j+1)\epsilon+i'\delta_h})}\right]_1^2 \tag{7.25}$$

and

$$E\left[\frac{E_{\{\Pi^h((j+1)\epsilon),Y_{(j+1)\epsilon,t+s}\}}[\phi(\tilde{X}(i'\delta_h))R^h(\tilde{X}_{0,i'\delta_h},Y_{(j+1)\epsilon,(j+1)\epsilon+i'\delta_h})]}{E_{\{\Pi^h((j+1)\epsilon),Y_{(j+1)\epsilon,t+s}\}}R^h(\tilde{X}_{0,i'\delta_h},Y_{(j+1)\epsilon,(j+1)\epsilon+i'\delta_h})} - \frac{E_{\{\Pi^h(j\epsilon+i\delta_h),Y_{j\epsilon+i\delta_h,t+s}\}}[\phi(\tilde{X}(\alpha\delta_h))R^h(\tilde{X}_{0,\alpha\delta_h},Y_{j\epsilon+i\delta_h,(j+1)\epsilon+i'\delta_h})]}{E_{\{\Pi^h(j\epsilon+i\delta_h),Y_{j\epsilon+i\delta_h,t+s}\}}R^h(\tilde{X}_{0,\alpha\delta_h},Y_{j\epsilon+i\delta_h,(j+1)\epsilon+i'\delta_h})}\right]_1^2. \tag{7.26}$$

Using the definition from (7.6) of $\Pi^h(t+s)$ in terms of $P^{h,j+1}_{\Pi^h((j+1)\epsilon)}$ in (7.25) and working with numerators and denominators separately, as we may, it follows that showing the convergence to zero of the $\sup_t$ of (7.25) as $h \to 0$ to zero is equivalent to showing the same for

$$E\Bigg[\int \phi(x(i'\delta_h))R^h(x_{0,i'\delta_h}, Y_{(j+1)\epsilon,(j+1)\epsilon+i'\delta_h})dP^{h,j+1}_{\Pi^h((j+1)\epsilon)}(x(\cdot)) - E_{\{\Pi^h((j+1)\epsilon),Y_{(j+1)\epsilon,t+s}\}}[\phi(\tilde{X}(i'\delta_h))R^h(\tilde{X}_{0,i'\delta_h}, Y_{(j+1)\epsilon,(j+1)\epsilon+i'\delta_h})]\Bigg]^2_1. \tag{7.27}$$

Let $\Delta > 0$ be small. Now, repeat the logic which led to (7.16). By A7.2 and the continuity properties of $\tilde{X}(\cdot)$, it is sufficient to prove the result if both of the $R^h(\cdot)$ in (7.27) are replaced by $R^\Delta(\cdot)$, and the $i'\delta_h$ in $\tilde{X}(i'\delta_h)$ and $x(i'\delta_h)$ are replaced by the nearest integral multiple of Δ. Then A7.1 yields the desired convergence. This takes care of (7.25).

We now turn to (7.26). This time we do not work with the numerators and denominators separately. For motivation, note that if $\Pi^h(\cdot)$ were the true conditional distribution for the discrete time signal process $X(n\delta_h)$, then (7.26) is identically zero. Let $\Delta > 0$ be small, and let ϵ be an integral multiple of Δ. Owing to the properties of $\tilde{X}(\cdot)$ and the tightness of the set $\{\Pi^h(t+\cdot); h, t\}$, it is sufficient to show that the $\lim_h \sup_t$ of (7.26) is zero if $R^h(\cdot)$ were replaced by $R^\Delta(\cdot)$ and the $i\delta_h$ and $i'\delta_h$ were integral multiples of Δ. Thus we can write $t = j\epsilon + k_1\Delta$ and $t+s = (j+1)\Delta + k_2\Delta$, where $k_i\Delta \le \epsilon$. Using the fact that the k_i have only finitely many values, it is sufficient to show that

$$\lim_h \sup_j E\Bigg[\frac{E_{\{\Pi^h((j+1)\epsilon),Y_{(j+1)\epsilon,(j+1)\epsilon+k_2\Delta}\}}[\phi(\tilde{X}(k_2\Delta))R^\Delta(\tilde{X}_{0,k_2\Delta}, Y_{(j+1)\epsilon,(j+1)\epsilon+k_2\Delta})]}{E_{\{\Pi^h((j+1)\epsilon),Y_{(j+1)\epsilon,(j+1)\epsilon+k_2\Delta}\}}[R^\Delta(\tilde{X}_{0,k_2\Delta}, Y_{(j+1)\epsilon,(j+1)\epsilon+k_2\Delta})]} - \frac{E_{\{\Pi^h(j\epsilon+k_1\Delta),Y_{j\epsilon+k_1\Delta,(j+1)\epsilon+k_2\Delta}\}}[\phi(\tilde{X}(\epsilon-k_1\Delta+k_2\Delta))R^\Delta(\tilde{X}_{0,\epsilon-k_1\Delta+k_2\Delta}, Y_{j\epsilon+k_1\Delta,(j+1)\epsilon+k_2\Delta})]}{E_{\{\Pi^h(j\epsilon+k_1\Delta),Y_{j\epsilon+k_1\Delta,(j+1)\epsilon+k_2\Delta}\}}[R^\Delta(\tilde{X}_{0,\epsilon-k_1\Delta+k_2\Delta}, Y_{j\epsilon+k_1\Delta,(j+1)\epsilon+k_2\Delta})]}\Bigg]^2_1 = 0. \tag{7.28}$$

The difficulty in treating this term is that the initial times are different, being $(j+1)\epsilon$ in the first term and $j\epsilon + i\delta_h$ in the second. Because of this, we need to represent both initial measures in terms of the same quantity, namely, in terms of $P^{h,j}_{\Pi^h(j\epsilon)}$, and the details will now be given. Define the function

$$\Theta\left(\phi, k_2\Delta, Y_{(j+1)\epsilon,(j+1)\epsilon+k_2\Delta}, x\right) = E\left[\phi(\tilde{X}(k_2\Delta))R^\Delta(\tilde{X}_{0,k_2\Delta}, Y_{(j+1)\epsilon,(j+1)\epsilon+k_2\Delta})\middle|\tilde{X}(0) = x, Y_{(j+1)\epsilon,(j+1)\epsilon+k_2\Delta}\right]. \tag{7.29}$$

If $\phi(\cdot)$ is equal to the constant function with value unity, we simply write 1 for ϕ in (7.29). Then, using the definition (7.6) of $\Pi^h((j+1)\epsilon)$ in terms of $P^{h,j}_{\Pi^h(j\epsilon)}$, the numerator of the first term inside the bars in (7.28) can be rewritten as

$$\frac{\int \Theta\left(\phi, k_2\Delta, Y_{(j+1)\epsilon,(j+1)\epsilon+k_2\Delta}, x(\epsilon)\right) R^\Delta(x_{0,\epsilon}, Y_{j\epsilon,(j+1)\epsilon})dP^{h,j}_{\Pi^h(j\epsilon)}(x(\cdot))}{\int R^\Delta(x_{0,\epsilon}, Y_{j\epsilon,(j+1)\epsilon})dP^{h,j}_{\Pi^h(j\epsilon)}(x(\cdot))}. \tag{7.30}$$

The denominator of the left-hand term inside the bars in (7.28) has the same representation, but with 1 replacing ϕ. Thus, that left-hand term can be written as

$$\frac{\int \Theta\left(\phi, k_2\Delta, Y_{(j+1)\epsilon,(j+1)\epsilon+k_2\Delta}, x(\epsilon)\right) R^\Delta(x_{0,\epsilon}, Y_{j\epsilon,(j+1)\epsilon})dP^{h,j}_{\Pi^h(j\epsilon)}x(\cdot))}{\int \Theta\left(1, k_2\Delta, Y_{(j+1)\epsilon,(j+1)\epsilon+k_2\Delta}, x(\epsilon)\right) R^\Delta(x_{0,\epsilon}, Y_{j\epsilon,(j+1)\epsilon})dP^{h,j}_{\Pi^h(j\epsilon)}(x(\cdot))}. \tag{7.31}$$

By A7.1, without changing the limits in (7.28), this fraction can be replaced by

$$\frac{E_{\{\Pi^h(j\epsilon),Y_{j\epsilon,(j+1\epsilon)+k_2\Delta}\}}\Theta\left(\phi,k_2\Delta,Y_{(j+1)\epsilon,(j+1)\epsilon+k_2\Delta},x(\epsilon)\right)R^{\Delta}(x_{0,\epsilon},Y_{j\epsilon,(j+1)\epsilon})}{E_{\{\Pi^h(j\epsilon),Y_{j\epsilon,(j+1\epsilon)+k_2\Delta}\}}\Theta\left(1,k_2\Delta,Y_{(j+1)\epsilon,(j+1)\epsilon+k_2\Delta},x(\epsilon)\right)R^{\Delta}(x_{0,\epsilon},Y_{j\epsilon,(j+1)\epsilon})}. \tag{7.32}$$

In turn, using the Markov property of $\tilde{X}(\cdot)$, the definition of $\Theta(\cdot)$ as a conditional expectation, and the fact that $R^{\Delta}(\cdot)$ is the exponential of a sum, this equals

$$\frac{E_{\{\Pi^h(j\epsilon),Y_{j\epsilon,(j+1)\epsilon)+k_2\Delta}\}}\phi(\tilde{X}(\epsilon+k_2\Delta))R^{\Delta}(\tilde{X}_{0,\epsilon+k_2\Delta},Y_{j\epsilon,(j+1)\epsilon+k_2\Delta})}{E_{\{\Pi^h(j\epsilon),Y_{j\epsilon,(j+1)\epsilon)+k_2\Delta}\}}R^{\Delta}(\tilde{X}_{0,\epsilon+k_2\Delta},Y_{j\epsilon,(j+1)\epsilon+k_2\Delta})}. \tag{7.33}$$

Now we turn our attention to the second term inside the bars in (7.28). This is treated in essentially the same way as was the first term. Consider the numerator of that term. The expectation, conditioned on

$$\{\tilde{X}(\epsilon-k_1\Delta)=x,\tilde{X}_{0,\epsilon-k_1\Delta},Y_{j\epsilon+k_1\Delta,(j+1)\epsilon+k_2\Delta}\},$$

is just

$$\Theta(\phi,k_2\Delta,Y_{(j+1)\epsilon,(j+1)\epsilon+k_2\Delta},x)R^{\Delta}(X_{0,\epsilon-k_1\Delta},Y_{j\epsilon+k_1\Delta,(j+1)\epsilon}).$$

We proceed as we did above with the first term. Using the definition (7.6) yields an expression analogous to (7.29). Then applying first A7.1 and then the Markov property of $\tilde{X}(\cdot)$ to that expression yields that we can replace the second term in (7.28) by (7.33) as well without changing the limit. We omit the details, which are nearly the same as for the first term. Thus the term in the bars in (7.28) can be replaced by zero without changing the limit.

The proof of (7.22) is now complete. □

REFERENCES

[1] A.G. Bhatt, A. Budhiraja, and R.L. Karandikar, *Markov property and ergodicity of the nonlinear filter*, SIAM J. Control Optim., submitted, 1999.

[2] P. Billingsley, *Convergence of Probability Measures*, Wiley, New York, 1968.

[3] A. Budhiraja and H.J. Kushner, *Approximation and limit results for nonlinear filters over an infinite time interval*, SIAM J. Control Optim., 37 (1999), pp. 1946–1979.

[4] A. Budhiraja and H.J. Kushner, *A nonlinear filtering algorithm based on an approximation of the conditional distribution*, IEEE Trans. Automat. Control, to appear.

[5] J. Carpenter, P. Clifford, and P. Fearnhead, *An Improved Particle Filter for Non-Linear Problems*, preprint, Statistics Department, University of Oxford, Oxford, UK, 1998.

[6] D. Crisan and T.Lyons, *Nonlinear filtering and measure valued processes*, Probab. Theory Related Fields, 109 (1997), pp. 217–244.

[7] P. Del Moral and G. Salat, *Filtrage non-linéaire résolution particuliere à la monte carlo*, C.R. Acad. Sci. Paris Ser. I Math, 320 (1997) pp. 1147–1152.

[8] R. Elliot, *Stochastic Calculus and Applications*, Springer-Verlag, Berlin, New York, 1982.

[9] S.N. Ethier and T.G. Kurtz, *Markov Processes: Characterization and Convergence*, Wiley, New York, 1986.

[10] G.S. Fishman, *Monte Carlo*, Springer, Berlin, New York, 1995.

[11] W.R. Gilks, S. Richardson, and D.J. Spiegelhalter, *Markov Chain Monte Carlo in Practice*, Chapman and Hall, London, 1996.

[12] N.J. Gordon, D.J. Salmond, and C. Ewing, *Bayesian state estimation for tracking and guidance using the bootstrap filter*, J. Guidance Control Dynamics, 18 (1995), pp. 1434–1443.

[13] N.J. Gordon, D.J. Salmond, and A.F.M. Smith, *Novel approach to nonlinear/nonGaussian Baysian state estimation*, IEE Proceedings-F, 140 (1993), pp. 107–113.

[14] G. Kallianpur, H. Fujisaki, and H. Kunita, *Stochastic differential equations for the nonlinear filtering problem*, Osaka J. Math., 9 (1972), pp. 19–40.
[15] M. Isard and A. Blake, *Contour tracking by stochastic propagation of conditional density*, in Proceedings of the 4th European Conference on Computer Vision, Cambridge, UK, 1996, pp. 343–356.
[16] G. Kitagawa, *Monte carlo filter and smoother for non Gaussian nonlinear state space models*, J. Comput. Graph. Statist., 5 (1996), pp. 1–25.
[17] P.E. Kloeden and E. Platen, *Numerical Solution of Stochastic Differential Equations*, Springer-Verlag, Berlin, New York, 1992.
[18] H. Kunita, *Asymptotic behavior of the nonlinear filtering errors of a Markov process*, J. Multivariate Anal., 1 (1971), pp. 365–393.
[19] T.G. Kurtz, *Approximation of Population Processes*, CBMS-NSF Regional Conf. Ser. in Appl. Math., 36, SIAM, Philadelphia, 1981.
[20] H.J. Kushner, *Dynamical equations for nonlinear filtering*, J. Differential Equations, 3 (1967), pp. 179–190.
[21] H.J. Kushner, *Probability Methods for Approximations in Stochastic Control and for Elliptic Equations*, Academic Press, New York, 1977.
[22] H.J. Kushner, *Weak Convergence Methods and Singularly Perturbed Stochastic Control and Filtering Problems*, Systems Control, Fourd. Appl. 3, Birkhäuser Boston, Boston, 1990.
[23] H.J. Kushner, *Robustness and convergence of approximations to nonlinear filters for jump–diffusions*, Comput. Appl. Math., 16 (1997), pp. 153–183.
[24] H.J. Kushner and P. Dupuis, *Numerical Methods for Stochastic Control Problems in Continuous Time*, Springer-Verlag, Berlin, New York, 1992.
[25] H.J. Kushner and H. Huang, *Approximation and limit results for nonlinear filters with wide bandwidth observation noise*, Stochastics Stochastics Rep., 16 (1986), pp. 65–96.
[26] R. Liptser and A.N. Shiryaev, *Statistics of Random Processes*, Springer-Verlag, Berlin, New York, 1977.
[27] P. Müller, *Monte Carlo integration in general dynamic models*, in Statistical Multiple Integration (Arcata, CA, 1989), Contemp. Math. 115, AMS, Providence, RI, 1991, pp. 145–163.
[28] M-S. Oh, *Monte Carlo integration via importance sampling: Dimensionality effect and an adaptive algorithm*, in Statistical Multiple Integration (Arcata, CA, 1989), Contemp. Math. 115, AMS, Providence, RI, 1991, pp. 165–187.
[29] M.K. Pitt and N. Shephard, *Filtering via Simulation: Auxiliary Particle Filters*, preprint, Mathematics Department, University of Oxford, Oxford, UK, 1997.

SIAM J. CONTROL OPTIM.
Vol. 38, No. 6, pp. 1909–1937

AN APPROXIMATION THEORY FOR STRONGLY STABILIZING SOLUTIONS TO THE OPERATOR LQ RICCATI EQUATION*

J. C. OOSTVEEN†, R. F. CURTAIN†, AND K. ITO‡

Abstract. The linear-quadratic (LQ) control problem is considered for a class of infinite-dimensional systems with bounded input and output operators, that are not exponentially stabilizable, but only strongly stabilizable. A sufficient condition for the existence of a minimizing control and of a stabilizing solution to the associated LQ Riccati equation is given. The main contribution of this paper is the convergence of the stabilizing solutions of a sequence of finite-dimensional Riccati equations to the strongly stabilizing solution of the infinite-dimensional Riccati equation. The result is applied to a model of propagation of sound waves in a one-dimensional wave-guide.

Key words. linear quadratic control problem, algebraic Riccati equation on Hilbert space, dissipative systems, collocated actuators and sensors

AMS subject classifications. 49N10, 65P05, 93C20

PII. S0363012998339691

1. Introduction.

Controllers designed on the basis of an infinite-dimensional model of the plant are very often infinite-dimensional as well. In particular, this is the case for LQ-, LQG-, and H_∞-controllers. Of course, when one wants to actually compute or even implement such a controller, one has to approximate it by a finite-dimensional controller. Therefore, over the years much research has been directed towards developing approximation schemes for these controllers and, in particular, approximation schemes for numerical solutions of algebraic Riccati equations. Many of the papers use an approach based on approximation results for C_0-semigroups, i.e., they use a version of the Trotter–Kato theorem in the proof of convergence of the solution of the Riccati equation (see, for instance, Banks and Burns [1], Gibson [9], Ito [11], Kappel and Salamon [14], and the references therein).

In all previous papers it was assumed that A generates an exponentially stable semigroup or that (A, B) is exponentially stabilizable. This is not always the case. In the literature, there are many examples of systems with dissipative generators that cannot be exponentially stabilized by a compact feedback operator (see Gibson [8]). In this paper, we consider the linear-quadratic (LQ) Riccati equation for systems $\Sigma(A, B, C)$ with bounded input and output operator, but we assume that $\Sigma(A, B, C)$ is strongly stabilizable and not exponentially stabilizable. Such systems occur often as models of flexible structures and in problems of wave propagation and scattering. The lack of exponential stability makes the analysis more difficult and our main contribution is the development of an approximation theory for stabilizing solutions of the LQ Riccati equation under the relaxed stabilizability assumption of strong stabilizability. Our present contribution is motivated by recent results on the existence

*Received by the editors June 1, 1998; accepted for publication (in revised form) January 28, 2000; published electronically August 3, 2000.
http://www.siam.org/journals/sicon/38-6/33969.html

†Department of Mathematics and Computer Science, University of Groningen, P.O. Box 800, 9700 AV Groningen, The Netherlands (j.c.oostveen@math.rug.nl, r.f.curtain@math.rug.nl). Current address for J.C. Oostveen: Philips Research Laboratories, Prof. Holstlaan 4, 5656 AA Eindhoven, The Netherlands (oostvn@natlab.research.philips.com).

‡Center for Research in Scientific Computation, North Carolina State University, Raleigh, NC 27695-8205 (kito@math.ncsu.edu).

of stabilizing solutions to Riccati equations for strongly stabilizable systems with bounded inputs and outputs (see Oostveen and Curtain [20]).

In section 2, we introduce some of the basic notions and we derive sufficient conditions for the existence and uniqueness of a strongly stabilizing solution to the LQ Riccati equation for systems that are strongly stabilizable and detectable. Existence results for (strongly) stabilizing solutions of more general Riccati equations can be found in Staffans [23] and Mikkola [18]. We prove a sharper existence result, but more importantly, our approximation results depend on the construction in our proof. Section 3 contains results about the convergence of finite-dimensional approximations of an infinite-dimensional system. The approximation result for the strongly stabilizing solution of the algebraic Riccati equation then follows in section 4. This result is specialized to dissipative systems with collocated actuators and sensors in section 5. These systems have a state-space description $\Sigma(A, B, B^*)$, where A generates a contraction semigroup (i.e., A is a dissipative operator). The terminology "collocated" for the condition $C = B^*$ comes from the fact that this condition arises when actuators and sensors are implemented at the same location. Although mathematically very special, this class is very important from an applications point of view. They are the typical example of systems that are strongly stabilizable, but not exponentially stabilizable. Finally, in section 6, we illustrate the approximation result with an example of LQ control for an acoustical problem. The model for this example was described in Morse and Ingard [19] and Beale [2], and the approximation of the system was taken from Ito and Propst [12].

2. The standard LQ Riccati equation for strongly stabilizable systems. Consider the system

$$\begin{aligned} \dot{z}(t) &= Az(t) + Bu(t), \quad z(0) = z_0, \\ y(t) &= Cz(t), \end{aligned} \tag{2.1}$$

where A is the infinitesimal generator of a C_0-semigroup $T(t)$ on the separable Hilbert space Z, $B \in \mathcal{L}(U, Z)$, $C \in \mathcal{L}(Z, Y)$, where U and Y are separable Hilbert spaces as well. In what follows we denote this system by $\Sigma(A, B, C)$. We want to find $\overline{u} \in \mathbf{L}_2(0, \infty; U)$ that minimizes the quadratic cost criterion

$$\mathcal{J}(u, z_0) = \int_0^\infty (\|y(t)\|^2 + \|u(t)\|^2)dt. \tag{2.2}$$

Before we give conditions for the existence of a solution for this problem, we give a number of definitions related to the stability of systems $\Sigma(A, B, C, D)$ (i.e., systems as in (2.1) where the second equation is replaced by $y(t) = Cz(t) + Du(t)$ and $D \in \mathcal{L}(U, Y)$). The following concept of a system was introduced in Weiss [24] in the more general context of well-posed linear systems.

DEFINITION 2.1. *Consider the system $\Sigma(A, B, C, D)$. We introduce the following maps for this system.*

- *The* input map $\Phi_t : \mathbf{L}_2(0, t; U) \to Z$ *is given by*

$$\Phi_t u = \int_0^t T(s)Bu(s)ds.$$

If for all u $\lim_{t\to\infty} \Phi_t u$ exists, then we define the extended input map Φ *by*

$$\Phi u = \lim_{t\to\infty} \Phi_t u.$$

- *The* output map $\Psi_t : Z \to \mathbf{L}_2(0,t;Y)$ *defined by*

$$(\Psi_t z)(\cdot) = CT(\cdot)z.$$

This map has a limit for $t \to \infty$ *as an operator from* Z *to* $\mathbf{L}_2^{loc}(0,\infty;Y)$. *If the limit exists as an operator from* Z *to* $\mathbf{L}_2(0,\infty;Y)$, *then it is called the* extended output map *and denoted by* Ψ. *It is then given by*

$$(\Psi z)(\cdot) = CT(\cdot)z.$$

- *The* extended input-output map $\mathbb{F} : \mathbf{L}_2(0,\infty;U) \to \mathbf{L}_2(0,\infty;Y)$ *is given by*

$$(\mathbb{F}u)(\cdot) = Du(\cdot) + C\int_0^\infty T(\cdot - s)Bu(s)ds.$$

It is defined via a similar limiting procedure as the extended output map.

- *For* $\alpha \in \mathbb{R}$, *let* $\mathbb{C}_\alpha^+ = \{s \in \mathbb{C} | \mathrm{Re}(s) > \alpha\}$. *The* transfer function $G(s) : \mathbb{C}_\alpha^+ \to \mathcal{L}(U,Y)$ *is given by*

$$G(s) = D + C(sI - A)^{-1}B.$$

We remark that if $T(t)$ is exponentially stable, then the extended input, output, and input-output maps are well defined and $G \in \mathbf{H}_\infty$. For our applications exponential stability is too strong and so we introduce the following weaker notions.

DEFINITION 2.2. *Consider the system* $\Sigma(A,B,C,D)$ *and the associated operators defined above.*

- $T(t)$ *is a* strongly stable C_0-semigroup *if* $\lim_{t\to\infty} T(t)z = 0$ *for all* $z \in Z$;
- $\Sigma(A,B,C,D)$ *is* input stable *if* $\Phi \in \mathcal{L}(\mathbf{L}_2(0,\infty;U),Z)$;
- $\Sigma(A,B,C,D)$ *is* output stable *if* $\Psi \in \mathcal{L}(Z,\mathbf{L}_2(0,\infty;Y))$;
- $\Sigma(A,B,C,D)$ *is* input-output stable *if* $\mathbb{F} \in \mathcal{L}(\mathbf{L}_2(0,\infty;U),\mathbf{L}_2(0,\infty;Y))$.

If $\Sigma(A,B,C,D)$ *satisfies all the above notions, then we call it a* strongly stable system.

These stability notions could have been defined in a different way. For instance, output stability is known to be equivalent to the condition $C(sI-A)^{-1}z \in \mathbf{H}_2(Y)$ for all $z \in Z$. It is also equivalent to C being an *infinite-time admissible observation operator* for $T(t)$. Input stability is equivalent to the condition that $B^*(sI-A^*)^{-1}z \in \mathbf{H}_2(U)$ for all $z \in Z$ and to B being an *infinite-time admissible control operator* for $T(t)$ (see, for instance, Hansen and Weiss [10]). Finally, input-output stability is equivalent to $G(s) \in \mathbf{H}_\infty(\mathcal{L}(U,Y))$.

The intuition behind these definitions is that a system is input stable if every $\mathbf{L}_2$-input results in a bounded state, a system is output stable if every initial condition results in an $\mathbf{L}_2$-output and a system is input-output stable if an $\mathbf{L}_2$-input results in an $\mathbf{L}_2$-output. A class of examples of strongly stable systems can be found in section 5.

Next we prove a result on the existence of a unique solution to the LQ control problem.

THEOREM 2.3. *Consider the system* $\Sigma(A,B,C)$ *with cost criterion* (2.2).

a. *Suppose that there exists an operator* $F \in \mathcal{L}(Z,U)$ *such that* $A+BF$ *generates a strongly stable semigroup* $T_F(t)$ *and* $\Sigma(A+BF,B,\left[\begin{smallmatrix}F\\C\end{smallmatrix}\right])$ *is output stable. Then there exists a minimizing control* $\overline{u}$ *which is given by* $\overline{u}(t) = -B^*X_0z(t)$, *where* X_0 *is the minimal self-adjoint nonnegative solution to the algebraic Riccati equation*

$$A^*Xz + XAz - XBB^*Xz + C^*Cz = 0 \tag{2.3}$$

for all $z \in D(A)$. Moreover, the corresponding minimal cost is given by

$$\mathcal{J}(\overline{u}, z_0) = \langle X_0 z_0, z_0 \rangle.$$

b. *If, in addition, there exists an operator $L \in \mathcal{L}(Y, Z)$ such that $A + LC$ generates a strongly stable C_0-semigroup $T_L(t)$ and $\Sigma(A + LC, \begin{bmatrix} L & B \end{bmatrix}, C)$ is input stable, then X_0 is the unique self-adjoint nonnegative solution to* (2.3). *Moreover, $A - BB^*X_0$ generates a strongly stable semigroup.*

Proof. a. Choosing $u = Fz$, we obtain

$$\begin{aligned}\mathcal{J}(Fz, z_0) &= \int_0^\infty (\|CT_F(t)z_0\|^2 + \|FT_F(t)z_0\|^2)dt \\ &\leq \text{const.}\|z_0\|^2,\end{aligned}$$

since $\Sigma(A + BF, B, \begin{bmatrix} F \\ C \end{bmatrix})$ is output stable. So the system satisfies the conditions of Theorem 6.2.4 in Curtain and Zwart [5] and there exists a minimal self-adjoint nonnegative solution X_0 of (2.3), such that the minimal cost is $\mathcal{J}(\overline{u}, z_0) = \langle X_0 z_0, z_0 \rangle$. From this we immediately obtain the following estimates, which we will use in the proof of part b.

$$\int_0^\infty \|B^*X_0 T_{X_0}(t)z_0\|^2 dt \leq \|X_0\|\|z_0\|^2 \tag{2.4}$$

and

$$\int_0^\infty \|CT_{X_0}(t)z_0\|^2 dt \leq \|X_0\|\|z_0\|^2, \tag{2.5}$$

where $T_{X_0}(t)$ is the semigroup generated by $A - BB^*X_0$.

b. We show that (2.4) and (2.5) hold if X_0 is any nonnegative solution of the Riccati equation (2.3). For clarity we write Π for an arbitrary nonnegative solution and reformulate (2.3) as

$$A_\Pi^*\Pi z + \Pi A_\Pi z + \Pi BB^*\Pi z + C^*Cz = 0, \tag{2.6}$$

where $A_\Pi = A - BB^*\Pi$ generates the semigroup $T_\Pi(t)$. Let $z = T_\Pi(t)z_0$ for $z_0 \in D(A)$ and take the inner product with $T_\Pi(t)z_0$ to obtain

$$\langle \Pi T_\Pi(t)z_0, A_\Pi T_\Pi(t)z_0 \rangle + \langle \Pi A_\Pi T_\Pi(t)z_0, T_\Pi(t)z_0 \rangle + \|B^*\Pi T_\Pi(t)z_0\|^2 + \|CT_\Pi(t)z_0\|^2 = 0.$$

Integrating with respect to t gives

$$\langle \Pi T_\Pi(t)z_0, T_\Pi(t)z_0 \rangle + \int_0^t \|B^*\Pi T_\Pi(s)z_0\|^2 + \int_0^t \|CT_\Pi(s)z_0\|^2 = \langle \Pi z_0, z_0 \rangle, \tag{2.7}$$

and this proves the claim.

c. We generalize Lemma 6.2.6 of Curtain and Zwart [5] to the strongly stable case. Let Π be a nonnegative solution to (2.3) such that $T_\Pi(t)$ is strongly stable. Define the following subset of controls:

$$\begin{aligned}U_{stab}(z_0) = \Big\{ & u \in \mathbf{L}_2(0, \infty; U) | \\ & z(t) = T(t)z_0 + \int_0^t T(t-s)Bu(s)ds \text{ and } z(t) \to 0 \text{ as } t \to \infty \Big\}.\end{aligned}$$

This set is nonempty, since $\overline{u}(t) = -B^*\Pi T_\Pi(t)z_0$ is in $U_{stab}(z_0)$; the closed-loop system has the state $z(t) = T_\Pi(t)z_0$ and $\overline{u} \in \mathbf{L}_2(0,\infty;U)$ follows from part b, above. Let $\tilde{\Pi}$ be any other nonnegative self-adjoint solution to (2.3). As in Lemma 6.2.6, it is easily shown that for any $u \in U_{stab}$ the cost on the finite-interval $[0,t_e]$ satisfies

$$\begin{aligned}\int_0^{t_e} (\|y(t)\|^2 + \|u(t)\|^2)dt &= \langle z_0, \tilde{\Pi}z_0\rangle - \langle z(t_e), \tilde{\Pi}z(t_e)\rangle \\ &\quad + \int_0^{t_e} \langle (u(s) + B^*\tilde{\Pi}z(s)), (u(s) + B^*\tilde{\Pi}z(s))\rangle ds \\ &\geq \langle z_0, \tilde{\Pi}z_0\rangle - \langle z(t_e), \tilde{\Pi}z(t_e)\rangle.\end{aligned}$$

Since $\tilde{\Pi} \geq 0$ and $z(t_e) \to 0$ as $t_e \to \infty$, we arrive at

$$\begin{aligned}\langle z_0, \tilde{\Pi}z_0\rangle &\leq \int_0^\infty (\|y(t)\|^2 + \|u(t)\|^2)dt \\ &= \mathcal{J}(u, z_0)\end{aligned}$$

for all $u \in U_{stab}$. Now $\overline{u} = -B^*\Pi T_\Pi(\cdot)z_0 \in U_{stab}$ and so

$$\langle z_0, \tilde{\Pi}z_0\rangle \leq \mathcal{J}(\overline{u}, z_0) = \langle z_0, \Pi z_0\rangle,$$

where we have used (2.7) and the fact that T_Π is strongly stable. Thus, Π is the maximal self-adjoint solution of (2.3).

d. We see that it remains to prove that $A - BB^*X_0$ generates a strongly stable C_0-semigroup. To do this, we write

$$\begin{aligned}\dot{z} &= (A - BB^*X_0)z \\ &= (A + LC)z + (-LC - BB^*X_0)z\end{aligned}$$

and so

$$z(t) = T_L(t)z_0 - \int_0^t T_L(t-s) \begin{bmatrix} L & B \end{bmatrix} \begin{bmatrix} Cz(s) \\ B^*X_0z(s) \end{bmatrix} ds.$$

Now, by the strong stability of $T_L(t)$, $T_L(t)z_0 \to 0$ as $t \to \infty$. Furthermore, $u_1(t) = Cz(t) \in \mathbf{L}_2(0,\infty;Y)$ by (2.5) and $u_2(t) = B^*X_0z(t) \in \mathbf{L}_2(0,\infty;U)$ by (2.4). So Lemma 12 in Oostveen and Curtain [20] applies to show that the integral term tends to zero as t tends to infinity. So, we obtain that $z(t)$ tends to zero as t goes to infinity, and $T_{X_0}(t)$ is strongly stable. □

We remark that a slightly less general version of this theorem was proven in Oostveen and Curtain [20], in which it was assumed that one could take $F = KC$ and $L = BK$ for some $K \in \mathcal{L}(Y,U)$. In Staffans [23] and Mikkola [18] abstract results on more general Riccati equations are proven. However, Mikkola proved the existence of a unique stabilizing solution, whereas we prove that the solution is unique in the class of self-adjoint nonnegative operators. For interpretations of Theorem 2.3 in terms of strong stabilizability and strong detectability, we refer to Curtain and Oostveen [4].

Next, we show how to reformulate the problem as an equivalent one for a strongly stable system. This was also done in Oostveen and Curtain [20] under the assumption that $F = KC$ and $L = BK$. Our assumptions in this paper are weaker.

THEOREM 2.4. *Consider the problem of minimizing the cost functional* (2.2) *subject to* $\Sigma(A,B,C)$ *under the assumption that there exist* $F \in \mathcal{L}(Z,U)$ *and* $L \in \mathcal{L}(Y,Z)$ *such that*

A1. *$A+BF$ and $A+LC$ generate strongly stable C_0-semigroups $T_F(t)$ and $T_L(t)$, respectively;*
A2. *$\Sigma(A+BF, B, \left[\begin{smallmatrix} F \\ C \end{smallmatrix}\right])$ is input-output stable and output stable;*
A3. *$\Sigma(A+LC, \left[\begin{smallmatrix} L & B \end{smallmatrix}\right], F)$ is input stable, output stable, and input-output stable.*

Then this problem is equivalent to that of minimizing the cost functional

$$\mathcal{J}_F(v, z_0) = \int_0^\infty (\|v(t) + Fz(t)\|^2 + \|y(t)\|^2)dt \tag{2.8}$$

subject to

$$\begin{aligned} \dot{z}(t) &= (A + BF)z(t) + Bv(t), \quad z(0) = z_0, \\ y(t) &= Cz(t). \end{aligned} \tag{2.9}$$

The optimal control for (2.8), (2.9) *is given by*

$$\overline{v}(t) = -(\mathbb{F}_F^*\mathbb{F}_F)^{-1}\mathbb{F}_F^*\Psi_F z_0, \tag{2.10}$$

where Ψ_F and $\mathbb{F}_F$ are the extended output map and the extended input-output map, respectively, of the system $\Sigma(A + BF, B, [\begin{smallmatrix} F \\ C \end{smallmatrix}], [\begin{smallmatrix} I \\ 0 \end{smallmatrix}])$. Moreover, there exists a unique nonnegative solution to the algebraic Riccati equation (2.3), *which is given by*

$$X = \Psi_F^* \left[I - \mathbb{F}_F(\mathbb{F}_F^*\mathbb{F}_F)^{-1}\mathbb{F}_F^*\right] \Psi_F. \tag{2.11}$$

Proof. a. Note that with $u = v + Fz$, the systems (2.1) (with $D = 0$) and (2.9) have the same solution

$$z(t) = T_F(t)z_0 + \int_0^t T_F(t-s)Bv(s)ds$$

and $\mathcal{J}_F(v, z_0) = \mathcal{J}(u, z_0)$. Therefore, the two optimization problems are equivalent if $\mathcal{J}(u, z_0) < \infty$ for $u \in \mathbf{L}_2(0, \infty; U)$ if and only if $v = u - Fz \in \mathbf{L}_2(0, \infty; U)$. Suppose first that $v \in \mathbf{L}_2(0, \infty; U)$. Since $\Sigma(A + BF, B, F)$ is input-output stable, $Fz(t) - FT_F(t)z_0 \in \mathbf{L}_2(0, \infty; U)$ and since $\Sigma(A+BF, B, F)$ is output stable, $FT_F(t)z_0$ and hence $Fz(t) \in \mathbf{L}_2(0, \infty; U)$. Thus $u \in \mathbf{L}_2(0, \infty; U)$. The output stability and input-output stability of $\Sigma(A + BF, B, C)$ imply that $Cz(t) \in \mathbf{L}_2(0, \infty; Y)$ which, together with $u \in \mathbf{L}_2(0, \infty; U)$, shows that $\mathcal{J}(u, z_0) < \infty$.

Conversely, suppose that $\mathcal{J}(u, z_0) < \infty$ and write

$$z(t) = T_L(t)z_0 + \int_0^t T_L(t-s)(Bu(s) - Ly(s))ds.$$

Then $Fz(t) - FT_L(t)z_0 \in \mathbf{L}_2(0, \infty; U)$ since $\Sigma(A+LC, \left[\begin{smallmatrix} B & L \end{smallmatrix}\right], F)$ is input-output stable. Also, $FT_L(t)z_0 \in \mathbf{L}_2(0, \infty; U)$, since $\Sigma(A + LC, B, F)$ is output stable. So $Fz(t) \in \mathbf{L}_2(0, \infty; U)$ and therefore also $v = u - Fz \in \mathbf{L}_2(0, \infty; U)$.

b. We now obtain an explicit solution of the minimization of (2.8) for the stable system (2.9) in terms of the bounded maps $\mathbb{F}_F$, Ψ_F given by

$$\begin{aligned} \Psi_F z_0 &= \left[\begin{smallmatrix} F \\ C \end{smallmatrix}\right] T_F(t)z_0, \\ (\mathbb{F}_F v)(t) &= \left[\begin{smallmatrix} v(t) \\ 0 \end{smallmatrix}\right] + \left[\begin{smallmatrix} F \\ C \end{smallmatrix}\right] \int_0^t T_F(t-s)Bv(s)ds. \end{aligned}$$

Since $\mathcal{J}_F(v, z_0) < \infty$ for all $v \in \mathbf{L}_2(0,\infty;U)$ we can write (2.8) equivalently as

$$\mathcal{J}_F(v, z_0) = \|\Psi_F z_0 + \mathbb{F}_F v\|^2.$$

Next we show that the cost functional $\mathcal{J}_F$ is coercive in v. Let the operator $\mathbb{F}_{LF}$ denote the extended input-output map of the system $\Sigma(A+LC, \left[\begin{array}{cc} B & -L \end{array}\right], -F, \left[\begin{array}{cc} I & 0 \end{array}\right])$. By assumption A3, $\mathbb{F}_{LF} \in \mathcal{L}(\mathbf{L}_2(0,\infty;U\times Y), \mathbf{L}_2(0,\infty;U))$. We show that $\mathbb{F}_{LF}\mathbb{F}_F = I_{L_2(0,\infty;U)}$, using a frequency domain argument. Let $\widehat{\ }$ denote Laplace transforms and let $v \in \mathbf{L}_2(0,\infty;U)$, $y \in \mathbf{L}_2(0,\infty;Y)$. Then,

$$\begin{aligned}&\widehat{(\mathbb{F}_{LF}(v \oplus y))}(s)\\ &= \left(\left[\begin{array}{cc} I & 0 \end{array}\right] - F(sI - A - LC)^{-1}\left[\begin{array}{cc} B & -L \end{array}\right]\right)\left(\begin{array}{c} \hat{v}(s) \\ \hat{y}(s) \end{array}\right)\end{aligned}$$

and

$$\widehat{(\mathbb{F}_F v)}(s) = \left(\left[\begin{array}{c} I \\ 0 \end{array}\right] + \left[\begin{array}{c} F \\ C \end{array}\right](sI - A - BF)^{-1}B\right)\hat{v}(s).$$

Now,

$$\begin{aligned}&\widehat{(\mathbb{F}_{LF}\mathbb{F}_F v)}(s)\\ &= \left\{F(sI - A - LC)^{-1}(LC - BF)(sI - A - BF)^{-1}B\right.\\ &\qquad \left. -F(sI - A - LC)^{-1}B + F(sI - A - BF)^{-1}B + I\right\}\hat{v}(s)\\ &= F(sI - A - LC)^{-1}\left\{-sI + A + BF + sI - A - LC + LC - BF\right\}\\ &\quad \cdot(sI - A - BF)^{-1}B\hat{v}(s) + \hat{v}(s)\\ &= \hat{v}(s).\end{aligned}$$

Consequently, $(\mathbb{F}_{LF}\mathbb{F}_F v)(t) = v(t)$ for all $v \in \mathbf{L}_2(0,\infty;U)$.

We deduce

$$\begin{aligned}\|v\|^2 &= \|\mathbb{F}_{LF}\mathbb{F}_F v\|^2\\ &\le \|\mathbb{F}_{LF}\|^2\|\mathbb{F}_F v\|^2\\ &= \|\mathbb{F}_{LF}\|^2\langle \mathbb{F}_F v, \mathbb{F}_F v\rangle\\ &\le \|\mathbb{F}_{LF}\|^2\|\mathbb{F}_F^*\mathbb{F}_F v\|\|v\|.\end{aligned}$$

Note that because of the identity $\mathbb{F}_{LF}\mathbb{F}_F = I$ we cannot have $\|\mathbb{F}_{LF}\| = 0$. Thus,

$$\|\mathbb{F}_F^*\mathbb{F}_F v\| \ge \|\mathbb{F}_{LF}\|^{-2}\|v\|. \tag{2.12}$$

We can now apply the theory of paragraphs 1.1.1–1.1.3 in Lions [17] to show that there exists a unique minimizing control $\overline{v} \in \mathbf{L}_2(0,\infty;U)$ that is characterized by

$$\mathbb{F}_F^*\mathbb{F}_F\overline{v} + \mathbb{F}_F^*\Psi_F z_0 = 0. \tag{2.13}$$

Because (2.12) implies that $\mathbb{F}_F^*\mathbb{F}_F$ is boundedly invertible, we can solve (2.13) for $\overline{v}$ to obtain (2.10).

From Theorem 2.3 we know that

$$\mathcal{J}_F(\overline{v}, z_0) = \mathcal{J}(\overline{u}, z_0) = \langle X_0 z_0, z_0\rangle$$

and (2.11) follows. □

It is interesting to note that although the strategy in Oostveen and Curtain [20] is also to express the general control problem via a feedback in terms of an equivalent stable one, the feedbacks are different and the formulas for X are very different. In the stable case ($F = 0$, $L = 0$), however, the formulas for X coincide. For proving convergence results for the standard LQ Riccati equation the approach in this paper is more convenient.

It turns out that A1, A2, A3, below, are key conditions in proving our main approximation result in section 4. Consequently, we introduce the following definition.

DEFINITION 2.5. *The system* $\Sigma(A, B, C)$ *is* strongly stabilizable-detectable *if there exist* $F \in \mathcal{L}(Z, U)$ *and* $L \in \mathcal{L}(Y, Z)$ *such that*

A1. $A+BF$ *and* $A+LC$ *generate strongly stable* C_0*-semigroups* $T_F(t)$ *and* $T_L(t)$*, respectively;*

A2. $\Sigma(A+BF, B, \left[\begin{smallmatrix} F \\ C \end{smallmatrix}\right])$ *is input-output stable and output stable;*

A3. $\Sigma(A+LC, \left[\begin{smallmatrix} L & B \end{smallmatrix}\right], F)$ *is input-output stable, input stable, and output stable.*

We remark that if both $A + BF$ and $A + LC$ generate exponentially stable C_0-semigroups on Z, then the conditions A1, A2, and A3 hold automatically. For a more detailed analysis of these conditions, we refer to Curtain and Oostveen [4].

3. Approximation for strongly stable systems. Consider again the system on the Hilbert spaces Z, U, Y

$$\begin{aligned} \dot{z}(t) &= Az(t) + Bu(t), \\ y(t) &= Cz(t) \end{aligned} \tag{3.1}$$

and a sequence of approximating systems on the finite-dimensional spaces $Z^N = \mathbb{R}^{k(N)}$, $U^N = \mathbb{R}^{m(N)}$, $Y^N = \mathbb{R}^{p(N)}$,

$$\begin{aligned} \dot{z}^N(t) &= A^N z^N(t) + B^N u^N(t), \\ y^N(t) &= C^N z^N(t). \end{aligned} \tag{3.2}$$

We assume the existence of injective linear maps

$$\begin{aligned} i^N &: \mathbb{R}^{k(N)} \to Z, \\ j^N &: \mathbb{R}^{m(N)} \to U, \\ k^N &: \mathbb{R}^{p(N)} \to Y \end{aligned}$$

and surjective linear maps

$$\begin{aligned} \pi^N &: Z \to \mathbb{R}^{k(N)}, \\ \rho^N &: U \to \mathbb{R}^{m(N)}, \\ \sigma^N &: Y \to \mathbb{R}^{p(N)}, \end{aligned}$$

such that $\pi^N i^N$, $\rho^N j^N$, and $\sigma^N k^N$ are identity maps and $i^N \pi^N$, $j^N \rho^N$, and $k^N \sigma^N$ are orthogonal projections. Note that it is not necessary that $(i^N)^* = \pi^N$, $(j^N)^* = \rho^N$, or $(k^N)^* = \sigma^N$. On the spaces $\mathbb{R}^{k(N)}$, $\mathbb{R}^{m(N)}$, and $\mathbb{R}^{p(N)}$ we will always consider the induced inner products

$$\begin{aligned} \langle z_1, z_2 \rangle_{k(N)} &= \langle i^N z_1, i^N z_2 \rangle_Z, \\ \langle u_1, u_2 \rangle_{m(N)} &= \langle j^N u_1, j^N u_2 \rangle_U, \\ \langle y_1, y_2 \rangle_{p(N)} &= \langle k^N y_1, k^N y_2 \rangle_Y. \end{aligned}$$

$(A^N)^*$, $(B^N)^*$, $(C^N)^*$ will denote the adjoint matrices with respect to the induced inner products. As we use the norms corresponding to these inner products as well,

it is obvious that $\|i^N\| = \|j^N\| = \|k^N\| = \|\pi^N\| = \|\rho^N\| = \|\sigma^N\| = 1$. Define $Z^N = \text{range}(i^N\pi^N)$, $U^N = \text{range}(j^N\rho^N)$, and $Y^N = \text{range}(k^N\sigma^N)$. Then

$$\|i^N B^N \rho^N\|_{\mathcal{L}(U^N, Z^N)} = \|B^N\|_{\mathcal{L}(\mathbb{R}^{m(N)}, \mathbb{R}^{k(N)})},$$
$$\|k^N C^N \pi^N\|_{\mathcal{L}(Z^N, Y^N)} = \|C^N\|_{\mathcal{L}(\mathbb{R}^{k(N)}, \mathbb{R}^{p(N)})},$$

and similar equalities hold for all operators defined in an analogous way.

Associated with $\Sigma(A, B, C)$, we define the operators $T, \Phi, \Psi, \mathbb{F}, G(s)$ as in Definition 2.1. The operators $T^N, \Phi^N, \Psi^N, \mathbb{F}^N, G^N(s)$ are defined in the same way, based on the system $\Sigma(A^N, B^N, C^N)$. In the latter case, we are dealing with a finite-dimensional system. Hence, $T^N(t) = e^{A^N t}$ and $(\Psi^N z)(t) = C^N e^{A^N t} z$.

DEFINITION 3.1. $\Sigma(A^N, B^N, C^N)$ converges strongly *to* $\Sigma(A, B, C)$ *if for all* $z \in Z$,

$$T(t)z = \lim_{N\to\infty} i^N e^{A^N t}\pi^N z,$$
$$T^*(t)z = \lim_{N\to\infty} i^N e^{(A^N)^* t}\pi^N z$$

uniformly on compact time intervals, and for all $z \in Z$, $u \in U$, $y \in Y$,

$$Bu = \lim_{N\to\infty} i^N B^N \rho^N u,$$
$$B^* z = \lim_{N\to\infty} j^N (B^N)^* \pi^N z,$$
$$Cz = \lim_{N\to\infty} k^N C^N \pi^N z,$$
$$C^* y = \lim_{N\to\infty} i^N (C^N)^* \sigma^N y,$$
$$u = \lim_{N\to\infty} j^N \rho^N u,$$
$$y = \lim_{N\to\infty} k^N \sigma^N y.$$

Note that the convergence of the semigroup in the above definition implies, by taking $t = 0$, that for all $z \in Z$,

$$z = \lim_{N\to\infty} i^N \pi^N z. \tag{3.3}$$

In connection with the convergence of a sequence of finite-dimensional systems to an infinite-dimensional system, we introduce a notion of uniform stability of the sequence of finite-dimensional systems.

DEFINITION 3.2. $\Sigma(A^N, B^N, C^N)$ *is* uniformly output stable *if there exists* $c > 0$ *such that for all* $z \in Z$,

$$\sup_N \int_0^\infty \|k^N C^N e^{A^N t}\pi^N z\|^2 dt = \sup_N \int_0^\infty \|k^N (\Psi^N \pi^N z)(t)\|^2 dt \le c\|z\|^2.$$

$\Sigma(A^N, B^N, C^N)$ *is* uniformly input-output stable *if there exists* $c > 0$ *such that*

$$\sup_N \|k^N \mathbb{F}^N \rho^N\| = \sup_N \|k^N G^N(s)\rho^N\|_{H_\infty} = \sup_N \|k^N C^N (sI - A^N)^{-1} B^N \rho^N\|_{H_\infty} \le c.$$

In the proof of the following lemma, as well as in all other convergence proofs in this paper, we will use the fact that if the sequences of operators R^N and S^N converge strongly to R and S, respectively, then $R^N S^N$ converges strongly to RS.

LEMMA 3.3. *Suppose that $T(t)$ is strongly stable, $\Sigma(A,B,C)$ is output stable, the finite-dimensional systems $\Sigma(A^N,B^N,C^N)$ converge strongly to $\Sigma(A,B,C)$, and $\Sigma(A^N,B^N,C^N)$ is uniformly output stable. Then for all $z \in Z$,*

$$\lim_{N\to\infty} k^N \Psi^N \pi^N z = \Psi z$$

in $\mathbf{L}_2(0,\infty;Y)$ and

$$\lim_{N\to\infty} i^N (\Psi^N)^* \sigma^N y = \Psi^* y$$

in Z for all $y \in \mathbf{L}_2(0,\infty;Y)$.

Proof. For any $t_1 > 0$, we obtain

$$\begin{aligned}
&\|\Psi z - k^N \Psi^N \pi^N z\|^2_{L_2(0,\infty;Y)} \\
&= \int_0^\infty \|CT(t)z - k^N C^N e^{A^N t}\pi^N z\|^2 dt \\
&= \int_0^{t_1} \|CT(t)z - k^N C^N e^{A^N t}\pi^N z\|^2 dt \\
&\quad + \int_{t_1}^\infty \|CT(t)z - k^N C^N e^{A^N t}\pi^N z\|^2 dt \\
&= \int_0^{t_1} \|CT(t)z - k^N C^N e^{A^N t}\pi^N z\|^2 dt \\
&\quad + \int_0^\infty \|CT(t+t_1)z - k^N C^N e^{A^N (t+t_1)}\pi^N z\|^2 dt.
\end{aligned}$$

Now, the integrand of the second integral satisfies

$$\begin{aligned}
&\|CT(t+t_1)z - k^N C^N e^{A^N (t+t_1)}\pi^N z\|^2 \\
&= \|CT(t+t_1)z - (k^N C^N e^{A^N (t+t_1)}\pi^N z - k^N C^N e^{A^N t}\pi^N T(t_1)z) \\
&\quad - k^N C^N e^{A^N t}\pi^N T(t_1)z\|^2 \\
&\le (\|CT(t+t_1)z\| + \|k^N C^N e^{A^N t}\{e^{A^N (t_1)}\pi^N z - \pi^N T(t_1)z\}\| \\
&\quad + \|\|k^N C^N e^{A^N t}\pi^N T(t_1)z\|)^2 \\
&\le 3\|CT(t+t_1)z\|^2 + 3\|k^N C^N e^{A^N t}\{e^{A^N (t_1)}\pi^N z - \pi^N T(t_1)z\}\|^2 \\
&\quad + 3\|k^N C^N e^{A^N t}\pi^N T(t_1)z\|^2.
\end{aligned}$$

Hence,

$$\begin{aligned}
&\|\Psi z - k^N \Psi^N \pi^N z\|^2_{L_2(0,\infty;Y)} \\
&\le \int_0^{t_1} \|CT(t)z - k^N C^N e^{A^N t}\pi^N z\|^2 dt + 3\int_0^\infty \|CT(t+t_1)z\|^2 \\
&\quad + 3\int_0^\infty \|k^N C^N e^{A^N t}\{e^{A^N t_1}\pi^N z - \pi^N T(t_1)z\}\|^2 \\
&\quad + 3\int_0^\infty \|\|k^N C^N e^{A^N t}\pi^N T(t_1)z\|^2 \\
&=: \alpha_1 + \alpha_2 + \alpha_3 + \alpha_4.
\end{aligned}$$

Now we will derive estimates for each of the integral terms $\alpha_1, \alpha_2, \alpha_3, \alpha_4$. For α_1 we have

$$\alpha_1 \leq 2\int_0^{t_1} \|(C - k^N C^N \pi^N)T(t)z\|^2 dt + 2\sup_N \|k^N C^N \pi^N\|^2 \int_0^{t_1} \|T(t)z - i^N e^{A^N t}\pi^N z\|^2 dt.$$

The right-hand side tends to zero as N tends to infinity by the Lebesgue dominated convergence theorem because the systems $\Sigma(A^N, B^N, C^N)$ are strongly convergent to $\Sigma(A, B, C)$. α_2 satisfies

$$\alpha_2 = 2\int_0^{\infty} \|CT(t_1 + t)z\|^2 dt = 2\|\Psi T(t_1)z\| \leq 2\|\Psi\|\|T(t_1)z\|,$$

which tends to zero as t_1 tends to infinity, independently of N, because of the output stability of $\Sigma(A, B, C)$ and the strong stability of $T(t)$. Since the systems $\Sigma(A^N, B^N, C^N)$ are uniformly output stable, there exists a positive constant c, such that

$$\alpha_3 \leq 3c\|i^N e^{A^N t_1}\pi^N z - T(t_1)z\|^2.$$

The right-hand side tends to zero as N tends to infinity because $\Sigma(A^N, B^N, C^N)$ converges strongly to $\Sigma(A, B, C)$. For the same constant c, we have

$$\alpha_4 \leq 3c\|T(t_1)z\|^2,$$

the right-hand side of which tends to zero, independently of N, because $T(t)$ is strongly stable. These estimates together show that $k^N \Psi^N \pi^N z \to \Psi z$ strongly as $N \to \infty$. Because $\|i^N(\Psi^N)^*\sigma^N\| = \|\Psi^N\| \leq c$, we can use Lemma 3.5 (page 151) of Kato [15] to argue that it is sufficient for the proof of convergence of $i^N(\Psi^N)^*\sigma^N y$ to $\Psi^* y$ to consider $y \in \mathbf{L}_2(0, \infty; Y)$ with compact support (the argument being that the class $\mathbf{L}_2(0, \infty)$-functions with compact support is dense in $\mathbf{L}_2(0, \infty)$). So, let $y \in \mathbf{L}_2(0, \infty; Y)$ have support on $[0, t_1]$, $t_1 > 0$. Then

$$\begin{aligned}\|\Psi^* y - i^N(\Psi^N)^*\sigma^N y\| &\leq \int_0^{t_1} \|(T^*(t) - i^N e^{(A^N)^* t}\pi^N)C^* y(t)\| dt \\ &\quad + \int_0^{t_1} \|i^N e^{(A^N)^* t}\pi^N\|\|(C^* - i^N(C^N)^*\sigma^N)y(t)\| dt.\end{aligned}$$

Both integrals tend to zero as $N \to \infty$ because the systems $\Sigma(A^N, B^N, C^N)$ are strongly convergent to $\Sigma(A, B, C)$. □

COROLLARY 3.4. *Under the assumptions of Lemma* 3.3,

$$\lim_{N\to\infty} \|k^N C^N (sI - A^N)^{-1}\pi^N z - C(sI - A)^{-1} z\|_{H_2(Y)} = 0$$

for all $z \in Z$.

Proof. This result follows directly by taking Laplace transforms in Lemma 3.3 and applying the Paley–Wiener theorem (see, e.g., Theorem A.6.21 in Curtain and Zwart [5]). □

LEMMA 3.5. *Suppose that the assumptions of Lemma* 3.3 *hold and, in addition,* $\Sigma(A,B,C)$ *is input-output stable and* $\Sigma(A^N,B^N,C^N)$ *is uniformly input-output stable. Then, for all* $u \in \mathbf{L}_2(0,\infty;U)$,

$$\lim_{N\to\infty} k^N \mathbb{F}^N \rho^N u = \mathbb{F}u$$

in $\mathbf{L}_2(0,\infty;Y)$. *Similarly, for all* $y \in \mathbf{L}_2(0,\infty;Y)$,

$$\lim_{N\to\infty} j^N (\mathbb{F}^N)^* \sigma^N y = \mathbb{F}^* y$$

in $\mathbf{L}_2(0,\infty;U)$.

Proof. By the uniform input-output stability of the systems $\Sigma(A^N,B^N,C^N)$, the operators $k^N\mathbb{F}^N\rho^N$ are uniformly bounded in N. Therefore it is sufficient to prove the convergence $\mathbb{F}^N u \to \mathbb{F}u$ for input functions u with compact support. Let $u \in \mathbf{L}_2(0,\infty;U)$ with support in $[0,t_1]$, and let y, y^N be defined by $y = \mathbb{F}u$, $y^N = k^N\mathbb{F}^N\rho^N u$ and define $M_1 = \sup_N \sup_{t\in[0,t_1]} \|i^N e^{A^N t}\pi^N\|$. Then

$$\begin{aligned}
&\|y(t) - y^N(t)\|^2 \\
&\le 3\int_0^t \|(C - k^N C^N \pi^N) T(s) B u(t-s)\|^2 ds \\
&\quad +3 \sup_N \|k^N C^N \pi^N\| \int_0^t \|(T(s) - i^N e^{A^N s}\pi^N) B u(t-s)\|^2 ds \\
&\quad +3\, M_1 \sup_N \|k^N C^N \pi^N\| \int_0^t \|(B - i^N B^N \rho^N) u(t-s)\|^2 ds,
\end{aligned}$$

for $0 \le t \le t_1$. By the strong convergence of $\Sigma(A^N,B^N,C^N)$ to $\Sigma(A,B,C)$ and the Lebesgue dominated convergence theorem, it now follows that $\|y - y^N\|_{L_2(0,t_1;Y)} \to 0$ as $N \to \infty$. Furthermore, defining

$$\begin{aligned}
z_u &= \int_0^{t_1} T(t_1 - s) B u(s) ds, \\
z_u^N &= \int_0^{t_1} e^{A^N(t_1-s)} B^N \rho^N u(s) ds,
\end{aligned}$$

we have, since u has support in $[0,t_1]$,

$$\begin{aligned}
y(t+t_1) &= (\Psi z_u)(t), \\
y^N(t+t_1) &= k^N(\Psi^N z_u^N)(t).
\end{aligned}$$

Lemma 3.3 now applies to prove that $\|y - y^N\|_{L_2(t_1,\infty;Y)} \to 0$ as $N \to \infty$.

To prove the convergence of $(\mathbb{F}^N)^*$, note that because of the uniform boundedness of $(\mathbb{F}^N)^*$ it is again sufficient to consider only y with compact support $[0,t_1]$. The proof then proceeds completely analogously to the case above. ∎

COROLLARY 3.6. *Under the assumptions of Lemma* 3.5,

$$\lim_{N\to\infty} \|k^N G^N(s)\rho^N \hat{u}(s) - G(s)\hat{u}(s)\|_{H_2(Y)} = 0$$

for all $\hat{u}(s) \in \mathbf{H}_2(U)$. *Similarly, for all* $\hat{y}(s) \in \mathbf{H}_2(Y)$,

$$\lim_{N\to\infty} j^N (G^N(s))^* \sigma^N \hat{y}(s) - G(s)^* \hat{y}(s)\|_{\mathbf{H}_2(Y)} = 0.$$

Proof. This result follows directly by taking Laplace transforms in Lemma 3.5 and applying the Paley–Wiener theorem (see, e.g., Theorem A.6.21 in Curtain and Zwart [5]). □

We remark that the above properties even hold for $\hat{u}(s) \equiv u$, $\hat{y}(s) \equiv y$ (see Corollary 3.4). However, in general we do not have that

$$\lim_{N\to\infty} \|k^N C^N (sI - A^N)^{-1} B^N \rho^N - C(sI - A)^{-1} B\|_{H_\infty} = 0.$$

See Proposition 1 and the counterexample on page 1146 in Kappel and Salamon [14].

4. Approximation for strongly stabilizable-detectable systems. In this section we will prove our main convergence result for Riccati equations for strongly stabilizable-detectable systems. The idea will be to transform the problem to an equivalent stable problem as in Theorem 2.4, and to apply the results for stable systems from section 3 to this stable problem. Before we prove this result in Theorem 4.2, we state and prove the following lemma, which relates the strong convergence of a system to which a feedback law is applied to the strong convergence of the original system.

LEMMA 4.1. *Let the sequence of systems $\Sigma(A^N, B^N, C^N)$ converge strongly to $\Sigma(A, B, C)$ and let the sequence of matrices $F^N \in \mathbb{R}^{m(N)\times k(N)}$ be chosen such that the operator sequences $j^N F^N \pi^N \in \mathcal{L}(Z, U)$ and $i^N (F^N)^* \rho^N \in \mathcal{L}(U, Z)$ converge strongly to F and F^*, respectively. Then $\Sigma(A^N + B^N F^N, B^N, [\begin{smallmatrix} F^N \\ C^N \end{smallmatrix}])$ converges strongly to $\Sigma(A + BF, B, [\begin{smallmatrix} F \\ C \end{smallmatrix}])$ as N tends to infinity.*

Proof. The only nontrivial part of the proof is the proof of the strong convergence of the sequence $i^N e^{(A^N + B^N F^N)t} \pi^N z$ to $T_F(t)z$, uniformly on compact time intervals, as N tends to infinity. By Theorem 3.1.7 of Davies [7], this is equivalent to proving that for all $z \in D(A)$, there exists a sequence $z^N \in \mathbb{R}^{k(N)}$ such that

$$\begin{aligned} i^N z^N &\to z, \\ i^N (A^N + B^N F^N) z^N &\to (A + BF) z. \end{aligned} \tag{4.1}$$

Because $i^N e^{A^N t} \pi^N z$ converges strongly to $T(t)z$, there exists a sequence of z^N for every $z \in D(A)$ such that

$$\begin{aligned} i^N z^N &\to z, \\ i^N A^N z^N &\to Az. \end{aligned} \tag{4.2}$$

Now taking this sequence z^N, we obtain

$$i^N (A^N + B^N F^N) z^N = i^N A^N z^N + i^N B^N F^N z^N.$$

From (4.2) it follows that the first term converges to Az. For the convergence of the second term, note that the assumptions state that

$$\begin{aligned} &i^N B^N \rho^N u \to Bu \text{ for all } u \in U \text{ and} \\ &j^N F^N \pi^N z \to Fz \text{ for all } z \in Z, \end{aligned}$$

and so, $i^N B^N \rho^N j^N F^N \pi^N z = i^N B^N F^N \pi^N z \to BFz$. Now,

$$\begin{aligned} \|i^N B^N F^N z^N - BFz\| &\le \|i^N B^N F^N (z^N - \pi^N z)\| + \|i^N B^N F^N \pi^N z - BFz\| \\ &\le \|i^N B^N F^N\| \|(z^N - \pi^N z)\| + \|i^N B^N F^N \pi^N z - BFz\|. \end{aligned}$$

Evidently, the second term converges to zero. The same holds for the first term, because $\|i^N B^N F^N\|$ is uniformly bounded by the uniform boundedness theorem and

$$\begin{aligned}\|(z^N - \pi^N z)\| &= \|\pi^N i^N z^N - \pi^N i^N \pi^N z\| \\ &\le \|\pi^N\| \cdot \|i^N z^N - i^N \pi^N z\| \\ &\le \|\pi^N\| \left(\|i^N z^N - z\| + \|z - i^N \pi^N z\|\right) \\ &\to 0. \qquad \square\end{aligned}$$

Now, we can prove our main result of this section, which is the approximation of a strongly stabilizing solution of the Riccati equation by a sequence of solutions to finite-dimensional Riccati equations.

THEOREM 4.2. *Consider the algebraic Riccati equation* (2.3)*, and assume that*

1. *the system $\Sigma(A, B, C)$ is strongly stabilizable-detectable (i.e., there exist $F \in \mathcal{L}(Z, U)$ and $L \in \mathcal{L}(Y, Z)$ for which the properties* A1–A3 *in Definition* 2.5 *are satisfied);*
2. *there exists a sequence of systems $\Sigma(A^N, B^N, C^N)$ which is strongly convergent to $\Sigma(A, B, C)$;*
3. *there exists a sequence of matrices $F^N \in \mathcal{L}(\mathbb{R}^{k(N)}, \mathbb{R}^{m(N)})$ such that*
   a. *$j^N F^N \pi^N \to F$ strongly and $i^N (F^N)^* \rho^N \to F^*$ strongly, where F is the operator from part* 1*.;*
   b. *$A^N + B^N F^N$ is a stable matrix;*
   c. *$\Sigma(A^N + B^N F^N, B^N, \left[\begin{smallmatrix} F^N \\ C^N \end{smallmatrix}\right])$ is uniformly output stable;*
   d. *$\Sigma(A^N + B^N F^N, B^N, \left[\begin{smallmatrix} F^N \\ C^N \end{smallmatrix}\right])$ is uniformly input-output stable;*
4. *there exists a sequence of matrices $L^N \in \mathcal{L}(\mathbb{R}^{m(N)}, \mathbb{R}^{p(N)})$ such that*
   a. *$A^N + L^N C^N$ is a stable matrix;*
   b. *$\Sigma(A^N + L^N C^N, \left[\begin{smallmatrix} B^N & L^N \end{smallmatrix}\right], F^N)$ is uniformly input-output stable.*

Then for every $z \in Z$,

$$Xz = \lim_{N \to \infty} i^N X^N \pi^N z,$$

where $X \in \mathcal{L}(Z)$ is the unique self-adjoint nonnegative solution of the algebraic Riccati equation (2.3) *and $X^N \in \mathcal{L}(\mathbb{R}^{k(N)})$ are the unique self-adjoint nonnegative solutions of the sequence of algebraic Riccati equations*

$$(A^N)^* X^N + X^N A^N - X^N B^N (B^N)^* X^N + (C^N)^* C^N = 0. \tag{4.3}$$

Moreover, denoting $F_X = -B^* X$ and $F^N_{X^N} = -(B^N)^* X^N$, the approximating closed-loop systems

$$\Sigma\left(A^N + B^N F^N_{X^N}, B^N, \begin{bmatrix} F^N_{X^N} \\ C^N \end{bmatrix}\right)$$

converge strongly to the infinite-dimensional closed-loop system

$$\Sigma\left(A + BF_X, B, \begin{bmatrix} F_X \\ C \end{bmatrix}\right).$$

Proof. We apply Theorem 2.4 to the approximating finite-dimensional systems $\Sigma(A^N, B^N, C^N)$. Since $A^N + B^N F^N$ and $A^N + L^N C^N$ are stable matrices, all of the assumptions A1–A3 are satisfied and the solutions to (4.3) are given by

$$X^N = (\Psi^N_F)^* \left[I - \mathbb{F}^N_F ((\mathbb{F}^N_F)^* \mathbb{F}^N_F)^{-1} (\mathbb{F}^N_F)^*\right] \Psi^N_F, \tag{4.4}$$

where Ψ_F^N and $\mathbb{F}_F^N$ are the output map and extended input-output map, respectively, for the system $\Sigma(A^N + B^N F^N, B^N, \left[\begin{smallmatrix} F^N \\ C^N \end{smallmatrix}\right], \left[\begin{smallmatrix} I \\ 0 \end{smallmatrix}\right])$. We apply Lemma 3.3 to obtain

$$\lim_{N\to\infty} (j^N \oplus k^N)\Psi_F^N \pi^N z = \Psi z \text{ in } \mathbf{L}_2(0,\infty; U \oplus Y), \tag{4.5}$$

$$\lim_{N\to\infty} i^N (\Psi_F^N)^* (\rho^N \oplus \sigma^N) w = \Psi_F^* w \tag{4.6}$$

for all $z \in Z$ and $w \in \mathbf{L}_2(0,\infty; U \oplus Y)$, where $j^N \oplus k^N$ denotes the direct sum of j^N and k^N, defined by $(j^N \oplus k^N)(u^N, y^N) = (j^N u^N, k^N y^N)$.

Similarly, Lemma 3.5 yields

$$(j^N \oplus k^N)\mathbb{F}_F^N \rho^N \to \mathbb{F}_F \text{ strongly}, \tag{4.7}$$

$$j^N (\mathbb{F}_F^N)^* (\rho^N \oplus \sigma^N) \to \mathbb{F}_F^* \text{ strongly} \tag{4.8}$$

as $N \to \infty$. Analogously to the definition of $\mathbb{F}_{LF}$ in the proof of Theorem 2.4, we define $\mathbb{F}_{LF}^N$ to be the extended input-output map of the system

$$\Sigma(A^N + L^N C^N, \left[\begin{smallmatrix} B^N & -L^N \end{smallmatrix}\right], -F^N, \left[\begin{smallmatrix} I & 0 \end{smallmatrix}\right]).$$

Assumption 4a of the present theorem tells us that $\|F_{LF}^N\|$ is uniformly bounded. Now from assumption 3d, $\sup \|\mathbb{F}_F^N\| < \infty$, and from (2.12), we have

$$\|((\mathbb{F}_F^N)^* \mathbb{F}_F^N)^{-1}\| \le \|\mathbb{F}_{LF}^N\|^2.$$

This shows that $\|((\mathbb{F}_F^N)^* \mathbb{F}_F^N)^{-1}\|$ is bounded from above uniformly in N. We also deduce that $\|j^N((\mathbb{F}_F^N)^* \mathbb{F}^N)^{-1}\rho^N\|$ are uniformly bounded from above, the reason being that we use the induced inner products on $\mathbb{R}^{m(N)}$ and so $\|((\mathbb{F}_F^N)^* \mathbb{F}_F^N)^{-1}\| = \|j^N((\mathbb{F}_F^N)^* \mathbb{F}^N)^{-1}\rho^N\|$. Thus,

$$\begin{aligned}
&\|j^N((\mathbb{F}_F^N)^* \mathbb{F}_F^N)^{-1}\rho^N v - (\mathbb{F}_F^* \mathbb{F}_F)^{-1} v\| \\
&\quad \le \|j^N((\mathbb{F}_F^N)^* \mathbb{F}_F^N)^{-1}\rho^N v - j^N \rho^N (\mathbb{F}_F^* \mathbb{F}_F)^{-1} v\| \\
&\qquad + \|j^N \rho^N (\mathbb{F}_F^* \mathbb{F}_F)^{-1} v - (\mathbb{F}_F^* \mathbb{F}_F)^{-1} v\| \\
&\quad = \|j^N((\mathbb{F}_F^N)^* \mathbb{F}_F^N)^{-1}\rho^N \left\{\mathbb{F}_F^* \mathbb{F}_F - j^N (\mathbb{F}_F^N)^* \mathbb{F}_F^N \rho^N\right\} (\mathbb{F}_F^* \mathbb{F}_F)^{-1} v\| \\
&\qquad + \|(j^N \rho^N - I)(\mathbb{F}_F^* \mathbb{F}_F)^{-1} v\| \\
&\quad \to 0
\end{aligned} \tag{4.9}$$

as $N \to \infty$.

The representations (2.11) and (4.4), together with the convergence results in (4.5)–(4.9), show that $i^N X^N \pi^N \to X$ strongly as $N \to \infty$. It is an easy consequence of the strong convergence of the stabilizing solutions that the optimal feedback operator $F_{X_N}^N$ and its adjoint converge strongly to their infinite-dimensional counterparts. The convergence of the approximating closed-loop systems now follows from the remark before Theorem 4.2. □

In the above theorem we have established the strong convergence of the approximating closed-loop systems to the infinite-dimensional closed-loop system. In practice, however, one would like to apply the finite-dimensional feedback to the infinite-dimensional plant. It is easy to see that the resulting closed-loop system

$$\Sigma\left(A + Bj^N F_{X^N}^N \pi^N, B, \left[\begin{smallmatrix} F_{X^N}^N \\ C \end{smallmatrix}\right]\right)$$

does converge strongly to the infinite-dimensional closed-loop system, as N tends to infinity. However, we have not been able to show that $A + Bj^N F^N \pi^N$ generates a strongly stable semigroup. Indeed, we doubt that this is true.

5. Applications to dissipative systems with collocated actuators and sensors. The class considered in this section, although mathematically quite special, is often used to model large flexible structures with collocated sensors and actuators (see, for instance, Joshi [13]). Systems in this class are the classical examples of systems that are strongly stabilizable but not exponentially stabilizable. The main feature is that they may have infinitely many unstable poles, which means that they are not exponentially stabilizable using a finite-rank, bounded input operator: exponentially stabilizable systems with a finite-rank, bounded input operator have at most finitely many unstable poles (see Curtain and Zwart [5, Theorem 5.2.6]). The state-space description is $\Sigma(A, B, C)$, where A is a dissipative operator on the Hilbert space Z, $B \in \mathcal{L}(U, Z)$ and $B = C^*,\ D = D^* \geq 0$. The terminology "collocated" for the condition $B = C^*$ comes from the fact that the condition arises if the actuators and sensors are implemented at the same location. This class was analyzed in Curtain and van Keulen [3], where they showed that the transfer function was positive-real and that the controller $K = -I$ robustly stabilized the system with respect to coprime factor perturbations with a robustness margin of at least $1/\sqrt{2}$.

We need to define notions of approximate controllability and approximate observability. We will not give the most intuitive definitions, but instead, we give an equivalent technical condition. For a more detailed treatment, see, e.g., Curtain and Zwart [5]. The system $\Sigma(A, B, C)$ is *approximately observable* if the operator $\Psi : Z \to \mathbf{L}_2(0, \infty; Y)$, given by $(\Psi z)(\cdot) = CT(\cdot)z$, satisfies $\ker(\Psi) = \{0\}$. The system $\Sigma(A, B, C)$ is *approximately controllable* if $\Sigma(A^*, C^*, B^*)$ is approximately observable. For a finite-dimensional system, these notions coincide with the usual controllability and observability definitions.

We consider linear systems $\Sigma(A, B, B^*)$ under the following assumptions.

- H1. A generates a C_0-semigroup of contractions $T(t)$ on the separable Hilbert space Z (i.e., A is a dissipative operator: $\langle Az, z\rangle + \langle z, Az\rangle \leq 0$ for all $z \in D(A)$).
- H2. U is a separable Hilbert space and $B \in \mathcal{L}(U, Z)$.
- H3. $\Sigma(A, B, B^*)$ is approximately observable.
- H4. $\Sigma(A, B, B)$ is approximately controllable.
- H5. A has compact resolvent.

The algebraic Riccati equation associated with this system is

$$A^*Xz + XAz - XBB^*Xz + BB^*z = 0 \tag{5.1}$$

for all $z \in D(A)$. The following properties of $\Sigma(A, B, B^*)$ were proven under the assumptions H1–H5 in [3, 6, 20, 22], respectively (see Oostveen and Curtain [21] for more details).

- P1. $A - BB^*$ generates the strongly stable C_0-semigroup $T_B(t)$ and $A^* - BB^*$ generates the strongly stable C_0-semigroup $T_B^*(t)$.
- P2. $\int_0^\infty \|B^*T_B(t)z\|^2 dt \leq \frac{1}{2}\|z\|^2$.
- P3. $\int_0^\infty \|B^*T_B^*(t)z\|^2 dt \leq \frac{1}{2}\|z\|^2$.
- P4. $B^*(sI - A + BB^*)^{-1}B \in \mathbf{H}_\infty(\mathcal{L}(U))$ and $\|B^*(sI - A + BB^*)^{-1}B\|_\infty \leq 1$.
- P5. There exists a unique self-adjoint nonnegative solution X to the algebraic Riccati equation (5.1) and $\|X\| \leq 1$.

The properties P1–P4 above can be summarized by saying that, choosing $F = -B^*$, $L = -B$, the system $\Sigma(A, B, B^*)$ is strongly stabilizable-detectable, i.e., $\Sigma(A - BB^*, B, B^*)$ is input stable, output stable, and input-output stable and $A - BB^*$ generates a strongly stable semigroup.

The claims above that $\|B^*(sI - A + BB^*)^{-1}B\|_\infty \leq 1$ and $\|X\| \leq 1$ were not found in [20]. Therefore we will provide proofs for them here. First, note that for $G(s) = B^*(sI - A + BB^*)^{-1}B$

$$\begin{aligned} &G(j\omega)^*G(j\omega) + (I - G(j\omega))^*(I - G(j\omega)) \\ &= 2G(j\omega)^*G(j\omega) + I - G(j\omega)^* - G(j\omega) \\ &= I + B^*(-j\omega I - A^* + BB^*)^{-1}(A + A^*)(j\omega I - A + BB^*)^{-1}B \\ &\leq I, \end{aligned}$$

where the inequality follows from the fact that A is dissipative. Thus,

$$I - G(j\omega)^*G(j\omega) \geq (I - G(j\omega))^*(I - G(j\omega)) \geq 0.$$

So, $G(j\omega)^*G(j\omega) \leq I$, which implies that $\|G\|_\infty \leq 1$.

For the estimate of the norm of X, note that

$$\langle Xz_0, z_0\rangle \leq \mathcal{J}(-B^*z(t), z_0) = 2\int_0^\infty \|B^*T_B(t)z_0\|^2 dt \leq \|z_0\|^2,$$

which implies that $\|X\| \leq 1$.

We now suppose that U and Y are finite-dimensional, $U = Y = \mathbb{R}^m$, and that there exists a sequence of injective linear maps $i^N : \mathbb{R}^{n(N)} \to Z$ and a sequence of surjective linear maps $\pi^N : Z \to \mathbb{R}^{n(N)}$ such that $\pi^N i^N = I_{n(N)}$ and $i^N\pi^N$ is an orthogonal projection on Z. We define a sequence of approximating systems $\Sigma(A^N, B^N, C^N)$ on $\mathbb{R}^{n(N)}$, where $A^N : \mathbb{R}^{n(N)} \to \mathbb{R}^{n(N)}N$ is such that

$$A^N + (A^N)^* \leq 0 \tag{5.2}$$

and B^N, C^N are chosen such that

$$\begin{aligned} B^N &= \pi^N B, \\ C^N &= B^* i^N. \end{aligned} \tag{5.3}$$

Now, we can give sufficient conditions for $\Sigma(A^N, B^N, C^N)$ to converge strongly to $\Sigma(A, B, B^*)$.

LEMMA 5.1. *Assume that the following conditions are satisfied.*

a1. *The sequence of injective maps $i^N : \mathbb{R}^{n(N)} \to Z$ and the sequence of surjective maps $\pi^N : Z \to \mathbb{R}^{n(N)}$ satisfy $\pi^N i^N = I_{n(N)}$, $i^N\pi^N$ is an orthogonal projection on Z, and $i^N\pi^N z \to z$ as $N \to \infty$ for all $z \in Z$.*

a2. *For all $z \in D(A)$ there exists a sequence $z^N \in \mathbb{R}^{n(N)}$ such that*

$$\begin{aligned} \|i^N z^N - z\| &\to 0 \quad \textit{and} \\ \|i^N A^N z^N - Az\| &\to 0 \ \textit{as}\ N \to \infty. \end{aligned} \tag{5.4}$$

a3. *For all $z \in D(A^*)$ there exists a sequence $z^N \in \mathbb{R}^{n(N)}$ such that*

$$\begin{aligned} \|i^N z^N - z\| &\to 0 \quad \textit{and} \\ \|i^N (A^N)^* z^N - A^* z\| &\to 0 \ \textit{as}\ N \to \infty. \end{aligned} \tag{5.5}$$

Then $\Sigma(A^N, B^N, C^N)$ defined by (5.2), (5.3) *converges strongly to $\Sigma(A, B, B^*)$. Moreover, $C^N = (B^N)^*$.*

Proof. We first show that $C^N = (B^N)^*$. For $u \in \mathbb{R}^m$, $z \in \mathbb{R}^{n(N)}$,

$$\begin{aligned}
&\langle B^N u, z\rangle_{\mathbb{R}^{n(N)}} \\
&= \langle \pi^N B u, z\rangle_{\mathbb{R}^{n(N)}} \\
&= \langle i^N \pi^N B u, i^N z\rangle_Z \\
&= \langle u, B^*(i^N\pi^N)^* i^N z\rangle_{\mathbb{R}^m} \\
&= \langle u, B^* i^N \pi^N i^N z\rangle_{\mathbb{R}^m} \\
&= \langle u, B^* i^N z\rangle_{\mathbb{R}^m} \\
&= \langle u, C^N z\rangle_{\mathbb{R}^m},
\end{aligned}$$

using the facts that $i^N\pi^N$ is an orthogonal projection and that $\pi^N i^N = I_{n(N)}$.

From Theorem 3.1.7 on page 80 in Davies [7], we have that (5.4) is equivalent to

$$i^N e^{A^N t}\pi^N z \to T(t)z$$

uniformly on compact time-intervals for all $z \in Z$, and (5.5) is equivalent to

$$i^N e^{(A^N)^* t}\pi^N z \to T^*(t)z$$

uniformly on compact time intervals for all $z \in Z$. Next, $i^N B^N = i^N\pi^N B \to B$ as $N \to \infty$ and $C^N\pi^N = (B^N)^*\pi^N = B^* i^N \pi^N \to B^*$ as $N \to \infty$, which completes the proof. ∎

LEMMA 5.2. *Let $\Sigma(A^N, B^N, C^N)$ satisfy the conditions of Lemma 5.1. Then $C^N = (B^N)^*$. Moreover, $\Sigma(A^N - B^N(B^N)^*, B^N, C^N)$ is uniformly output stable and uniformly input-output stable. If $\Sigma(A^N, B^N, (B^N)^*)$ is observable, then $A^N - B^N(B^N)^*$ is a stable matrix.*

Proof. We have $(B^N)^* = C^N$ and $A^N + (A^N)^* \le 0$. Therefore, the system $\Sigma(A^N - B^N(B^N)^*, B^N, (B^N)^*)$ satisfies properties P3 and P4. Hence it is uniformly output stable and uniformly input-output stable. The stability of $A^N - B^N(B^N)^*$ follows as in the infinite-dimensional case from the observability of the pair $(A^N, (B^N)^*)$ and $A^N + (A^N)^* \le 0$. Consider solutions $z(t) \in \mathbb{R}^{n(N)}$ of

$$\dot z(t) = (A^N - B^N(B^N)^*)z(t), \quad z(0) = z_0,$$

for arbitrary $z_0 \in \mathbb{R}^{n(N)}$. We introduce as a Lyapunov function for this differential equation $V(z) = \|z\|^2$. Then, if we differentiate V along solutions of the differential equation, we obtain

$$\begin{aligned}
&\frac{d}{dt}V(z(t)) \\
&= \langle z(t), \dot z(t)\rangle + \langle \dot z(t), z(t)\rangle \\
&= \langle z(t), (A^N - B^N(B^N)^*)z(t)\rangle + \langle (A^N - B^N(B^N)^*)z(t), z(t)\rangle \\
&= \langle (A^N + (A^N)^* - 2B^N(B^N)^*)z(t), z(t)\rangle \\
&\le -2\|(B^N)^* z(t)\|^2, \qquad (5.6)
\end{aligned}$$

where the inequality follows from the dissipativity of A^N. Now by Lasalle's invariance principle (see Lasalle [16]) all solutions $z(t)$ converge to the largest $A^N - B^N(B^N)^*$-invariant subset of the set

$$\mathcal{S} = \{z \in \mathbb{R}^{n(N)} |\ \langle (A^N + (A^N)^* - 2B^N(B^N)^*)z, z\rangle = 0\}.$$

Now let $0 \neq z_0 \in \mathcal{S}$, then for all $t \geq 0$, $z(t) = e^{(A^N - B^N(B^N)^*)t} z_0 \in \mathcal{S}$. And so, by (5.6)

$$0 = \frac{d}{dt} V(z(t)) \leq -2\|(B^N)^* z(t)\|^2 = -2\|(B^N)^* e^{(A^N - B^N(B^N)^*)t} z_0\|^2.$$

Hence, for all $t \geq 0$, $(B^N)^* e^{(A^N - B^N(B^N)^*)t} z_0 = 0$, which by the observability assumption implies that $z_0 = 0$. Thus $\mathcal{S} = \{0\}$ and consequently $A^N - B^N(B^N)^*$ is stable. □

Remark 5.3. If we assume that $i^N \pi^N z \to z$ and we choose $A^N = \pi^N A i^N$, then A^N satisfies the assumptions of Lemma 5.1, i.e., $A^N + (A^N)^* \leq 0$ and for all $z \in D(A)$, there is a sequence z^N which converges to z and for which $i^N A^N z^N$ converges to Az.

Indeed, if we choose the sequence $z^N = \pi^N z$, then

$$i^N z^N = i^N \pi^N z \ \to \ z$$

and

$$i^N A^N z^N = i^N \pi^N A i^N \pi^N z \ \to \ Az,$$

because $i^N \pi^N$ converges strongly to I, as N tends to infinity.

Furthermore, for $z_1, z_2 \in \mathbb{R}^{n(N)}$,

$$\begin{aligned}
&\langle A^N z_1, z_2 \rangle_{\mathbb{R}^{n(N)}} \\
&= \langle i^N \pi^N A i^N z_1, i^N z_2 \rangle_Z \\
&= \langle A i^N z_1, i^N \pi^N i^N z_2 \rangle_Z \text{ since } i^N \pi^N \text{ is an orthogonal projection} \\
&= \langle A i^N z_1, i^N z_2 \rangle_Z \text{ since } \pi^N i^N = I_{N(n)} \\
&= \langle i^N z_1, A^* i^N z_2 \rangle_Z \\
&= \langle i^N \pi^N i^N z_1, A^* i^N z_2 \rangle_Z \text{ since } \pi^N i^N = I_{N(n)} \\
&= \langle i^N z_1, i^N \pi^N A^* i^N z_2 \rangle_Z \text{ since } i^N \pi^N \text{ is an orthogonal projection} \\
&= \langle z_1, \pi^N A^* i^N z_2 \rangle_{\mathbb{R}^{n(N)}},
\end{aligned}$$

and so $(A^N)^* = \pi^N A^* i^N$. Now,

$$\begin{aligned}
&\langle (A^N + (A^N)^*) z, z \rangle_{\mathbb{R}^{n(N)}} \\
&= \langle \pi^N (A + A^*) i^N z, z \rangle_{\mathbb{R}^{n(N)}} \\
&= \langle i^N \pi^N (A + A^*) i^N z, i^N z \rangle_Z \\
&= \langle (A + A^*) i^N z, i^N \pi^N i^N z \rangle_Z \text{ since } i^N \pi^N \text{ is an orthogonal projection} \\
&= \langle (A + A^*) i^N z, i^N z \rangle_Z \text{ since } \pi^N i^N = I_{N(n)} \\
&\leq 0 \text{ since } A \text{ is dissipative.}
\end{aligned}$$

Thus $A^N + (A^N)^* \leq 0$.

The following corollary is now a direct consequence of Theorem 4.2, Lemmas 5.1 and 5.2, Remark 5.3, and the properties of systems $\Sigma(A, B, B^*)$.

COROLLARY 5.4. *Let the system* $\Sigma(A, B, B^*)$ *satisfy the assumptions* H1–H4*, and furthermore, assume that* $U = Y = \mathbb{R}^m$*. Let the sequence of approximating systems* $\Sigma(A^N, B^N, C^N)$ *as in* (5.2) *and* (5.3) *satisfy the assumptions of Lemma* 5.2*. Then* $X^N \to X$ *strongly as* $N \to \infty$*, where* X *and* X^N *are the solutions of the Riccati equation associated with* $\Sigma(A, B, B^*)$ *and* $\Sigma(A^N, B^N, C^N)$*, respectively.*

The class of dissipative systems with collocated actuators and sensors has a very special structure. However, one can also allow hybrid systems composed of a distributed system of the structure $\Sigma(A, B, B^*)$ coupled with a finite-dimensional system $\Sigma(A_f, B_f, C_F)$. The composite system

$$\Sigma\left(\begin{bmatrix} A & 0 \\ 0 & A_f \end{bmatrix}, \begin{bmatrix} B \\ B_f \end{bmatrix}, \begin{bmatrix} B^* & C_f \end{bmatrix}\right)$$

will satisfy the assumptions of our theory, provided that $\Sigma(A^N, B^N, C^N)$ are chosen as before and $\Sigma(A_f, B_f, C_f)$ is minimal. This broadens the class of possible applications considerably.

6. A numerical example: The one-dimensional wave equation with boundary oscillators. In this section, we apply the approximation result to a model of propagation of sound in a one-dimensional wave-guide of length 1, where the ends react as linear oscillators to the acoustic pressure. Actuating and sensing take place through the two ends.

Let subscripts x and t denote partial derivatives with respect to these variables. $\Psi(x,t)$ denotes the velocity potential, so that $\Psi_x(x,t)$ is the particle velocity in the fluid and $\Psi_t(x,t)$ is the acoustic pressure. Now, Ψ satisfies the wave equation

$$\Psi_{tt}(x,t) = \Psi_{xx}(x,t), \quad x \in (0,1).$$

The displacements of the two ends $\eta_0(t)$ at $x = 0$ and $\eta_1(t)$ at $x = 1$ satisfy

$$\frac{d^2\eta_0}{dt^2}(t) + d_0\frac{d\eta_0}{dt}(t) + k_0\eta_0(t) = -\rho_0\Psi_t(0,t) + f_1(t),$$
$$\frac{d^2\eta_1}{dt^2}(t) + d_1\frac{d\eta_1}{dt}(t) + k_1\eta_1(t) = -\rho_1\Psi_t(1,t) + f_2(t),$$

where f_1 and f_2 are external forces at the ends. From continuity of velocity at the boundary, we obtain

$$\frac{d\eta_0}{dt}(t) = -\frac{\partial\Psi}{\partial x}(0,t),$$
$$\frac{d\eta_1}{dt}(t) = \frac{\partial\Psi}{\partial x}(1,t).$$

The measurements are taken to be proportional to the velocities at the endpoints

$$y_1(t) = \frac{1}{2\rho_0}\frac{d\eta_0}{dt}(t),$$
$$y_2(t) = \frac{1}{2\rho_1}\frac{d\eta_1}{dt}(t).$$

The model can be brought into first-order form straightforwardly, introducing $z = \mathrm{col}(z_1, z_2, z_3, z_4, z_5, z_6)$ as the state, where $z_1 = \Psi$, $z_2 = \Psi_t$, $z_3 = \eta_0$, $z_4 = \eta_1$, $z_5 = \dot{\eta}_0$ and $z_6 = \dot{\eta}_1$. The state-space is then endowed with the norm corresponding to the energy of the system (see Beale [2]). For the approximation that we want to perform, a different state-space realization, introduced by Ito and Propst [12], is more convenient. An important difference between the approach of Ito and Propst and the one of Beale is that the former obtain an A-operator that has compact resolvent, but Beale does

not. The approach of Ito and Propst is to decompose Ψ into a part w^+ propagating in the positive direction and a part w^- propagating in the negative direction:

$$w^+(x,t) = \frac{1}{2}(\Psi_t(x,t) - \Psi_x(x,t)),$$
$$w^-(x,t) = \frac{1}{2}(\Psi_t(x,t) + \Psi_x(x,t)).$$

Defining $\phi_0 = \eta_0$, $\phi_1 = \eta_1$, $\psi_0 = \dot{\eta_0}$, $\psi_1 = \dot{\eta_1}$, we obtain

$$\frac{d}{dt}\begin{bmatrix} w^-(x,t) \\ w^+(x,t) \\ \phi_0(t) \\ \phi_1(t) \\ \psi_0(t) \\ \psi_1(t) \end{bmatrix} = \begin{bmatrix} w_x^-(x,t) \\ -w_x^+(x,t) \\ \psi_0(t) \\ \psi_1(t) \\ -\rho_0 w^-(0,t) - \rho_0 w^+(0,t) - k_0\phi_0(t) - d_0\psi_0(t) \\ -\rho_1 w^-(1,t) - \rho_1 w^+(1,t) - k_1\phi_1(t) - d_1\psi_1(t) \end{bmatrix},$$

with boundary conditions

$$\psi_0(t) = -w^-(0,t) + w^+(0,t),$$
$$\psi_1(t) = w^-(1,t) - w^+(1,t).$$

The state-space is taken to be $Z_0 = \mathbf{L}_2(0,1)\times\mathbf{L}_2(0,1)\times\mathbb{R}^4$ with inner product (writing $z = \mathrm{col}(w^-(x), w^+(x), \phi_0, \phi_1, \psi_0, \psi_1)$)

$$\langle z, \tilde{z}\rangle = \langle w^-, \tilde{w}^-\rangle_{L_2} + \langle w^+, \tilde{w}^+\rangle_{L_2} + \frac{k_0}{2\rho_0}\phi_0\tilde{\phi_0} + \frac{k_1}{2\rho_1}\phi_1\tilde{\phi_1} + \frac{1}{2\rho_0}\psi_0\tilde{\psi_0} + \frac{1}{2\rho_1}\psi_1\tilde{\psi_1}. \tag{6.1}$$

The input and output spaces are $U = Y = \mathbb{R}^2$. Let us define the operators A_0, B_0, and C_0 as

$$D(A_0) = \{z \in Z_0 \mid w^- \in H^1(0,1),\ w^+ \in H^1(0,1), \\ \psi_0 = -w^-(0) + w^+(0),\ \psi_1 = w^-(1) - w^+(1)\}, \tag{6.2}$$

$$A_0 z = \begin{bmatrix} w_x^-(x,t) \\ -w_x^+(x,t) \\ \psi_0(t) \\ \psi_1(t) \\ -\rho_0 w^-(0,t) - \rho_0 w^+(0,t) - k_0\phi_0(t) - d_0\psi_0(t) \\ -\rho_1 w^-(1,t) - \rho_1 w^+(1,t) - k_1\phi_1(t) - d_1\psi_1(t) \end{bmatrix}, \tag{6.3}$$

$$B_0 = \begin{bmatrix} 0 & 0 \\ 0 & 0 \\ 0 & 0 \\ 0 & 0 \\ 1 & 0 \\ 0 & 1 \end{bmatrix}, \quad C_0 = \begin{bmatrix} 0 & 0 & 0 & 0 & \frac{1}{2\rho_0} & 0 \\ 0 & 0 & 0 & 0 & 0 & \frac{1}{2\rho_1} \end{bmatrix}. \tag{6.4}$$

With these definitions, the system can be written as

$$\dot{z} = A_0 z + B_0 u,$$
$$y = C_0 z.$$

It is an easy computation to see that $C_0 = B_0^*$.

In Ito and Propst [12], it was shown that A_0 generates a C_0-semigroup of contractions $T_0(t)$ on Z_0 and A_0 has compact resolvent. Because A_0 has compact resolvent, A_0 has only point spectrum and it can be computed that $\lambda \in \sigma(A_0)$ if and only if λ satisfies

$$0 = (\lambda^2 + (d_0 + \rho_0)\lambda + k_0)(\lambda^2 + (d_1 + \rho_1)\lambda + k_1)e^{\lambda} - (\lambda^2 + (d_0 - \rho_0)\lambda + k_0)(\lambda^2 + (d_1 - \rho_1)\lambda + k_1)e^{-\lambda}. \tag{6.5}$$

Except $\lambda = 0$, all eigenvalues have negative real part and occur in complex conjugate pairs. The associated eigenvectors are given by

$$z_\lambda = \begin{bmatrix} (\lambda^2 + (d_0 + \rho_0)\lambda + k_0)(\lambda^2 + (d_1 - \rho_1)\lambda + k_1)e^{\lambda x} \\ \lambda^2 + (d_0 - \rho_0)\lambda + k_0)(\lambda^2 + (d_1 - \rho_1)\lambda + k_1)e^{-\lambda x} \\ -2\rho_0(\lambda^2 + (d_1 - \rho_1)\lambda + k_1) \\ -2\rho_1(\lambda^2 + (d_0 + \rho_0)\lambda + k_0)e^{\lambda} \\ -2\rho_0(\lambda^2 + (d_1 - \rho_1)\lambda + k_1)\lambda \\ -2\rho_1(\lambda^2 + (d_0 + \rho_0)\lambda + k_0)\lambda e^{\lambda} \end{bmatrix}.$$

The vectors $\{z_\lambda |\ \lambda \in \sigma(A)\}$ form a complete orthogonal set in Z_0, and all eigenvalues are simple.

Because in acoustics, one is interested only in variations of the pressure, we have to add an extra condition to "filter out" states that correspond to the hydrostatic pressure, i.e., the part of pressure that is constant in the spatial variable. It can be checked that this corresponds to states that are a multiple of $z_0 = \mathrm{col}(k_0k_1, k_0k_1, -2\rho_0k_1, -2\rho_1k_0, 0, 0)$. This leads to the extra condition

$$\int_0^1 w^-(x)dx + \int_0^1 w^+(x)dx - \phi_0 - \phi_1 = 0.$$

Therefore, we do not use Z_0 as our state-space, but the quotient space

$$Z = \left\{ z \in Z_0 | \int_0^1 w^-(x)dx + \int_0^1 w^+(x)dx - \phi_0 - \phi_1 = 0 \right\}.$$

Z is a closed linear subspace of Z_0 and thus a Hilbert space with the same inner product. Define $A = A_0|_Z$. A is the generator of a C_0-semigroup of contractions $T(t)$ in Z because $T_0(t)$ maps Z into Z (this follows from the fact that z_0 is an eigenvector of A and Z is the orthogonal complement of z_0 in Z_0).

A also inherits the properties that it is a Riesz-spectral operator and that it has compact resolvent from A_0. So A also has only a point spectrum and $\lambda \in \sigma(A)$ if and only if $\lambda \neq 0$ and λ satisfies (6.5). Define B and C as the restrictions of B_0 and C_0 to Z. Using the controllability and observability test in Theorem 4.2.3 of Curtain and Zwart [5], it can be computed that $\Sigma(A, B, C)$ is approximately controllable and approximately observable. We show only the controllability, as the observability is completely analogous. By Theorem 4.2.3 in Curtain and Zwart [5], $\Sigma(A, B, C)$ is controllable if for all $\lambda \in \sigma_p(A)$

$$(\langle b_1, z_\lambda \rangle, \langle b_2, z_\lambda \rangle) \neq 0,$$

where b_1, b_2 are the two columns of B. In our case,

$$(\langle b_1, z_\lambda \rangle, \langle b_2, z_\lambda \rangle) = (-2\rho_0(\lambda^2 + (d_1 - \rho_1)\lambda + k_1)\lambda,\ -2\rho_1(\lambda^2 + (d_0 + \rho_0)\lambda + k_0)\lambda e^{\lambda}).$$

As $0 \in \rho(A)$, $\langle b_1, z_\lambda \rangle = 0$ implies that $\lambda^2 + (d_1 - \rho_1)\lambda + k_1 = 0$. Similarly, $\langle b_2, z_\lambda \rangle = 0$ implies that $\lambda^2 + (d_0 + \rho_0)\lambda + k_0 = 0$. But these two conditions together imply that $z_\lambda = 0$, which contradicts the fact that λ is an eigenvalue of A.

We can summarize the previous results by saying that $C = B^*$ and the system $\Sigma(A, B, B^*)$ satisfies the assumptions H1–H5 as in section 5. In particular, $A - BB^*$ generates a strongly stable semigroup $T_B(t)$. It is interesting to note that $T_B(t)$ cannot be an exponentially stable semigroup for the following reason. Let $\tilde{A}$ denote A for the specific case that $d_0 = d_1 = 0$. This operator $\tilde{A}$ satisfies $\mathrm{Re}\langle \tilde{A}z, z \rangle = 0$ and so $\tilde{A}$ generates a unitary semigroup $\tilde{T}(t)$. Moreover,

$$A - BB^* = \tilde{A} - BB^* - \begin{bmatrix} 0 & & \\ & d_0 & 0 \\ & 0 & d_1 \end{bmatrix}.$$

So $A - BB^*$ is a compact perturbation of $\tilde{A}$ and the same holds for $A + BF$ for arbitrary bounded F. Now, because $A - BB^*$ generates a strongly stable contraction semigroup $T_B(t)$, we can apply Theorem 2 of Gibson [8] to show that the semigroup $T_F(t)$ generated by $A + BF$ is exponentially stable if and only if both $T_B(t)$ and $\tilde{T}(t)$ are exponentially stable. This is impossible, because $\tilde{T}(t)$ is unitary.

Next, we construct a sequence of approximating finite-dimensional systems, and we will show that it satisfies the conditions that guarantee convergence of the approximating solutions of the finite-dimensional LQ Riccati equations to the solution of the infinite-dimensional LQ Riccati equation. Ito and Propst proposed an approximation scheme for $\Sigma(A_0, B_0, C_0)$. We explain this scheme and then make an obvious modification of it to obtain an approximation scheme for $\Sigma(A, B, C)$. We use a spectral method based on Legendre polynomials. This method is very well suited for our problem: it is accurate and converges very fast (faster than the polynomial rate). Moreover, it is known that this scheme does not produce extraneous eigenvalues, which standard finite element and finite difference methods may do.

Let $z = \mathrm{col}(w^-(x), w^+(x), \phi_0, \phi_1, \psi_0, \psi_1) \in Z_0$. We define z_0^N by

$$z_0^N = \pi_0^N z = (w_0^-, \ldots, w_N^-, w_0^+, \ldots, w_{N-1}^+, \phi_0^N, \phi_1^N, \psi_0^N, \psi_1^N)^T \in \mathbb{R}^{2N+5},$$

where $\phi_0^N = \phi_0$, $\phi_1^N = \phi_1$, $\psi_0^N = \psi_0$, $\psi_1^N = \psi_1$, and w_k^-, w_m^+, $k = 0, \ldots, N$, $m = 0, \ldots, N-1$, are defined as follows. Let $P_k(x)$ be the Legendre polynomial of degree k. The set $\{P_k(x),\ k \in \mathbb{N}\}$ forms an orthogonal basis of $\mathbf{L}_2(-1, 1)$ and $\|P_k\|^2_{L_2(-1,1)} = 2/(2k+1)$. Hence the polynomials $p_k(x) = \sqrt{2k+1}P_k(2x-1)$ form an orthonormal basis for $\mathbf{L}_2(0, 1)$. We expand $w^-(x)$ and $w^+(x)$ with respect to the basis $\{p_k(x),\ k \in \mathbb{N}\}$ and truncate these expressions as shown below:

$$w_N^-(x) = \sum_{k=0}^{N} w_k^- p_k(x),$$

$$w_{N-1}^+(x) = \sum_{k=0}^{N-1} w_k^+ p_k(x).$$

Note the difference in notation between the approximating function $w_N^-(x)$ and the coefficient w_N^-. Let i_0^N denote the corresponding embedding which associates with a vector v in $\mathbb{R}^{2N+5}$ an element in Z_0, defined by

$$i_0^N v = \mathrm{col}\left(\sum_{k=0}^{N} v_{k+1} p_k(x), \sum_{k=0}^{N-1} v_{N+2+k} p_k(x), v_{2N+2}, v_{2N+3}, v_{2N+4}, v_{2N+5} \right).$$

Next, we introduce

$$\tilde{w}_N^-(x) = \sum_{k=0}^{N} w_k^- p_k(x) + a p_{N+1}(x),$$
$$\tilde{w}_{N-1}^+(x) = \sum_{k=0}^{N-1} w_k^+ p_k(x) + b p_N(x),$$

where a and b are chosen such that $\tilde{w}_N^-(x)$ and $\tilde{w}_{N-1}^+(x)$ satisfy the boundary conditions

$$\psi_0 = -\tilde{w}_N^-(0) + \tilde{w}_{N-1}^+(0),$$
$$\psi_1 = \tilde{w}_N^-(1) - \tilde{w}_{N-1}^+(1).$$

We will now use $\tilde{w}_N^-(x)$ and $\tilde{w}_{N-1}^+(x)$ to define the approximation A_0^N of A_0.

$$A_0^N \begin{bmatrix} w_0^- \\ \vdots \\ w_N^- \\ w_0^+ \\ \vdots \\ w_{N-1}^+ \\ \phi_0 \\ \phi_1 \\ \psi_0 \\ \psi_1 \end{bmatrix} = \begin{bmatrix} a_0^- \\ \vdots \\ a_N^- \\ a_0^+ \\ \vdots \\ a_{N-1}^+ \\ \psi_0 \\ \psi_1 \\ -k_0\phi_0 - d_0\psi_0 - \rho_0\tilde{w}_N^-(0) - \rho_0\tilde{w}_{N-1}^+(0) \\ -k_1\phi_1 - d_1\psi_1 - \rho_1\tilde{w}_N^-(1) - \rho_1\tilde{w}_{N-1}^+(1) \end{bmatrix},$$

and $a_0^-, \ldots, a_N^-, a_0^+, \ldots, a_{N-1}^+$ are defined via

$$\frac{d\tilde{w}_N^-}{dx}(x) = \sum_{k=0}^{N} a_k^- p_k(x), \quad -\frac{d\tilde{w}_{N-1}^+}{dx}(x) = \sum_{k=0}^{N-1} a_k^+ p_k(x).$$

The approximations B_0^N and C_0^N of B_0 and C_0 are given by

$$B_0^N = \begin{bmatrix} 0_{(2N+3)\times 2} \\ I_2 \end{bmatrix}, \quad C_0^N = (B^N)^*.$$

The approximating state z^N is now obtained by orthogonal projection of z_0^N on the $(2N+4)$-dimensional linear subspace of $\mathbb{R}^{2N+5}$ of vectors satisfying

$$\int_0^1 \sum_{k=0}^{N} w_k^- p_k(x)dx + \int_0^1 \sum_{k=0}^{N-1} w_k^+ p_k(x)dx - \phi_0 - \phi_1 = 0.$$

Let this orthogonal projection be denoted by χ^N and the associated embedding from $\mathbb{R}^{2N+4}$ into $\mathbb{R}^{2N+5}$ by h^N. The approximating system $\Sigma(A^N, B^N, C^N)$ is obtained by restricting A_0^N, B_0^N, and C_0^N to this $(2N+4)$-dimensional space, i.e., $A^N = \chi^N A_0^N h^N$, $B^N = \chi^N B_0^N$, $C^N = C_0^N h^N$. The projection π^N and embedding i^N for the total approximation are now given as

$$i^N : \mathbb{R}^{2N+4} \to Z, \quad i^N = i_0^N h^N,$$
$$\pi^N : Z \to \mathbb{R}^{2N+4}, \quad \pi^N = \chi^N \pi_0^N.$$

Now $\pi^N i^N$ is an orthogonal projection in Z and $i^N \pi^N = I_{2N+4}$.

TABLE 6.1
Comparison of norms of the different approximations.

N	$\|X^N\|$	$\|X^N - \pi^N i^M X^M \pi^M i^N\|$
2	0.9802	6.31e-11
4	0.9802	1.81e-10
8	0.9802	6.25e-10
16	0.9802	2.32e-9
32	0.9802	1.57e-8
64	0.9802	0

In Ito and Propst [12] it was shown that for every $z \in D(A)$ there exists a sequence $z^N \in \mathbb{R}^{2N+4}$ such that

$$\|i^N z^N - z\| \to 0 \quad \text{and}$$
$$\|i^N A^N z^N - Az\| \to 0,$$

as $N \to \infty$. Similarly, we can prove that for all $z \in D(A^*)$ there exists a sequence $z^N \in \mathbb{R}^{2N+4}$ such that

$$\|i^N z^N - z\| \to 0 \quad \text{and}$$
$$\|i^N (A^N)^* z^N - A^* z\| \to 0.$$

It is easily shown that C^N and B^N converge strongly to C and B, respectively. As in Lemma 5.1 these results imply that $\Sigma(A^N, B^N, C^N)$ converges strongly to $\Sigma(A, B, B^*)$. Hence, we can conclude from Corollary 5.4 that if $\Sigma(A^N, B^N, C^N)$ is observable for all N, then the solution X of the infinite-dimensional LQ Riccati equation associated with $\Sigma(A, B, B^*)$ can be approximated by the sequence of solutions X^N of the matrix LQ Riccati equations associated with $\Sigma(A^N, B^N, C^N)$.

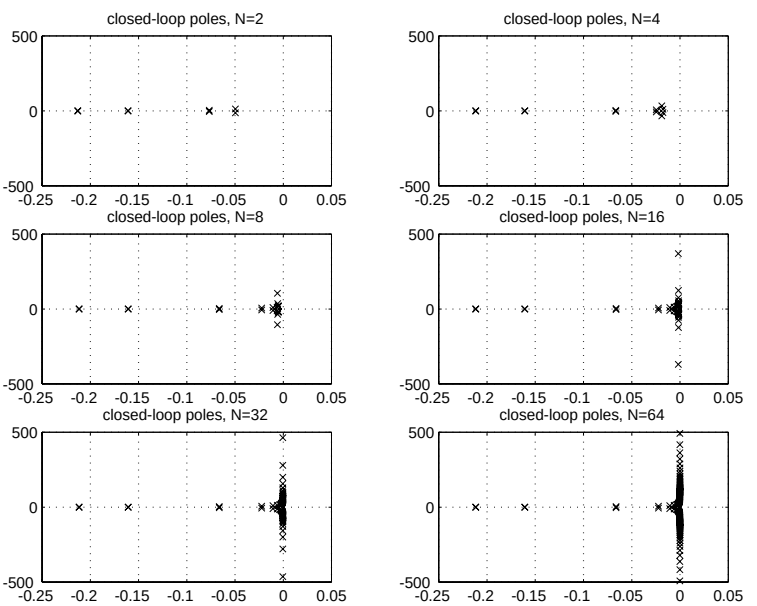


FIG. 6.1. *Closed-loop poles,* $N = 2, 4, 8, 16, 32, 64$.

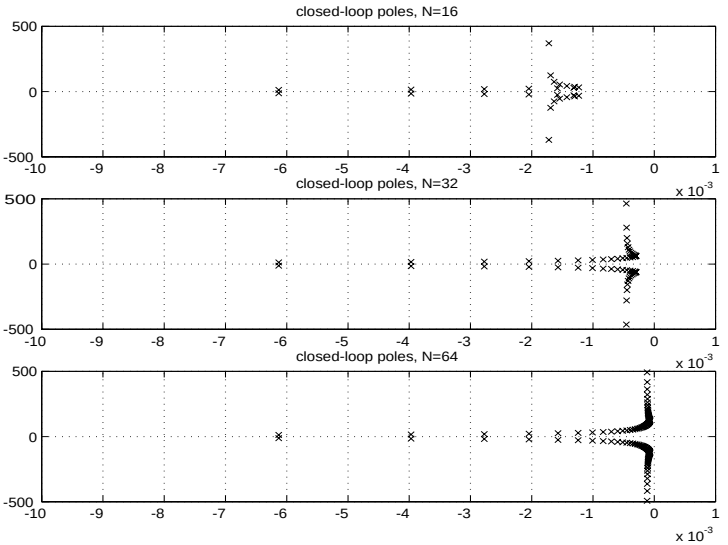


FIG. 6.2. *Closed-loop poles,* $N = 16, 32, 64$.

We implemented the approximation scheme in Matlab. The values for the parameters we used are

$$\rho_0 = \rho_1 = 1,$$
$$d_0 = d_1 = 0.01,$$
$$k_0 = k_1 = 1.$$

We computed the approximating systems for $N = 2, 4, 8, 16, 32, 64$. Then we computed the solutions to the Riccati equations corresponding to these approximations.

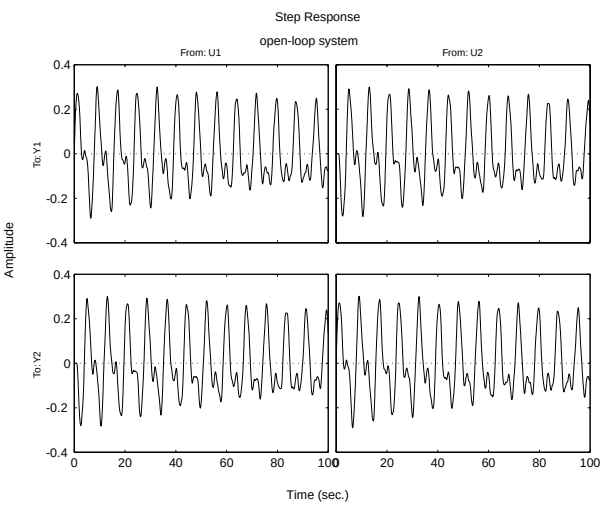


FIG. 6.3. *Open-loop step response,* $N = 16$.

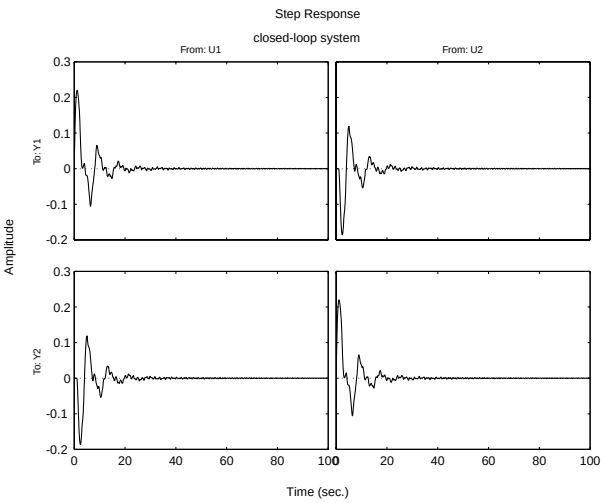


FIG. 6.4. *Closed-loop step response,* $N = 16$.

To illustrate the convergence of the solutions, we compare the solutions in a number of characteristics. First, we compute the matrix norms of the approximating solutions. Second, we restrict X^{64} to the spaces on which the lower-order approximations are defined and compute the norm of the difference: $\|X^N - \pi^N i^M X^M \pi^M i^N\|$. This norm measures to what extent the lower-order approximations matches the behavior of the higher-order one on the lower-dimensional space. These values are listed in Table 6.1, above.

Third, we give plots of the closed-loop poles. Finally, we computed the open-loop and closed-loop step responses for $N = 16$ and $N = 64$. Because these were indistinguishable for the two values of N used, we give only the plots for $N = 16$.

From these results, we observe the following. The norms of the approximations are very close to each other, and far larger than the norm of the difference between the low-order approximations and the restriction of X^{64} to the lower-order spaces.

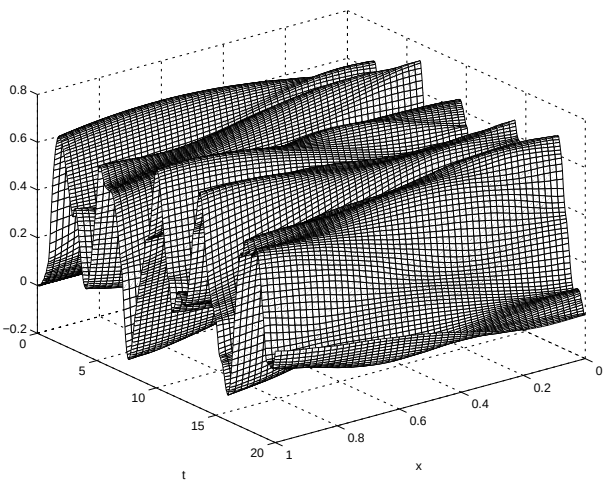


FIG. 6.5. *Open-loop step response,* $N = 16$.

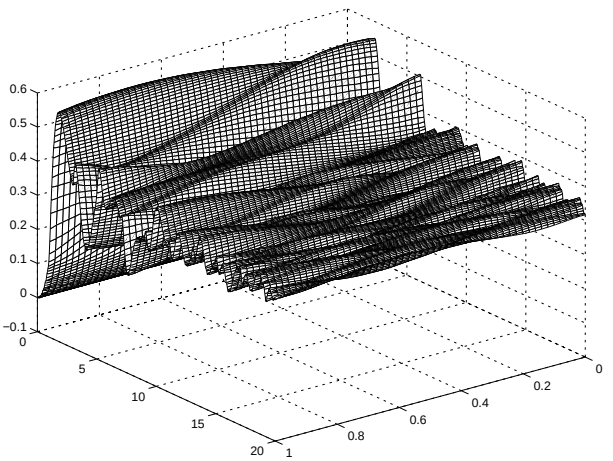


FIG. 6.6. *Closed-loop step response,* $N = 16$.

This indicates that the low-order approximations capture very well the behavior of the infinite-dimensional solution on the finite-dimensional spaces on which they are defined. The plots of the closed-loop poles in Figures 6.1 and 6.2 indicate that also the closed-loop system matrices converge as N tends to infinity. As expected, since the original system is not exponentially stabilizable, the poles are converging to the imaginary axis. The plots of the step responses are not very different for the different values of N. We plotted step responses for $N = 16$ in Figures 6.3 and 6.4. In Figures 6.5 and 6.6 we plotted the response of the velocity potential $\Psi(x,t)$ to a step input at $x = 0$, both for the open-loop and closed-loop system.

REFERENCES

[1] H. BANKS AND J. BURNS, *Hereditary control problems: Numerical methods based on averaging approximations*, SIAM J. Control Optimization, 16 (1978), pp. 169–208.

[2] J. BEALE, *Spectral properties of an acoustic boundary condition*, Indiana Univ. Math. J., 26 (1976), pp. 895–917.

[3] R. CURTAIN AND B. VAN KEULEN, *Robust control with respect to coprime factors of infinite-dimensional positive real systems*, IEEE Trans. Automat. Control, 37 (1992), pp. 868–871.

[4] R. CURTAIN AND J. OOSTVEEN, *Necessary and sufficient conditions for strong stability of distributed parameter systems*, Systems Control Lett., 37 (1999), pp. 11–18.

[5] R. CURTAIN AND H. ZWART, *An Introduction to Infinite-Dimensional Linear Systems Theory*, Springer-Verlag, New York, 1995.

[6] R. CURTAIN AND H. ZWART, *Riccati equations and normalized coprime factorizations for strongly stabilizable infinite-dimensional systems*, Systems Control Lett., 28 (1996), pp. 11–22.

[7] E. DAVIES, *One-Parameter Semigroups*, Academic Press, London, 1980.

[8] J. GIBSON, *A note on stabilization of infinite-dimensional linear oscillators by compact linear feedback*, SIAM J. Control Optim., 18 (1980), pp. 311–316.

[9] J. GIBSON, *Linear-quadratic optimal control of hereditary differential systems: Infinite-dimensional Riccati equations and numerical approximations*, SIAM J. Control Optim., 21 (1983), pp. 95–139.

[10] S. HANSEN AND G. WEISS, *New results on the operator Carleson measure criterion*, IMA J. Math. Control Inform., 14 (1997), pp. 3–32.

[11] K. ITO, *Strong convergence and convergence rates of approximating solutions for algebraic Riccati equations in Hilbert spaces*, in Distributed Parameter Systems, F. Kappel, K. Kunisch, and W. Schappacher, eds., Springer-Verlag, New York, 1987, pp. 151–166.

[12] K. Ito and G. Propst, *Legendre-tau-Padé approximations to the one-dimensional wave equation with boundary oscillators*, Numer. Funct. Anal. Optim., 90 (1997), pp. 57–70.
[13] S. Joshi, *Control of Large Flexible Space Structures*, Lecture Notes in Control Inform. Sci. 131, Springer-Verlag, Berlin, 1989.
[14] F. Kappel and D. Salamon, *An approximation theorem for the algebraic Riccati equation*, SIAM J. Control Optim., 28 (1990), pp. 1136–1147.
[15] T. Kato, *Perturbation Theory of Linear Operators*, Springer-Verlag, New York, 1966.
[16] J. Lasalle, *The Stability of Dynamical Systems*, Regional Conference Series in Applied Mathematics, SIAM, Philadelphia, 1976.
[17] J.-L. Lions, *Optimal Control of Systems Governed by Partial Differential Equations*, Springer-Verlag, Berlin, 1971.
[18] K. Mikkola, *On the Stable H_2 and H_∞ Infinite-dimensional Regular Problems and Their Algebraic Riccati Equations*, Tech. Report A383, Helsinki University of Technology, Institute of Mathematics, Helsinki, Finland, 1997.
[19] P. Morse and K. Ingard, *Theoretical Acoustics*, McGraw-Hill, New York, 1968.
[20] J. Oostveen and R. Curtain, *Riccati equations for strongly stabilizable bounded linear systems*, Automatica J. IFAC, 34 (1998), pp. 953–967.
[21] J. Oostveen and R. Curtain, *Robustly stabilizing controllers for dissipative infinite-dimensional systems with collocated actuators and sensors*, Automatica J. IFAC, 36 (2000), pp. 337–348.
[22] M. Slemrod, *A note on complete controllability and stabilizability for linear control systems in Hilbert space*, SIAM J. Control, 12 (1974), pp. 500–508.
[23] O. Staffans, *Quadratic optimal control of well-posed linear systems*, SIAM J. Control Optim., 37 (1998), pp. 131–164.
[24] G. Weiss, *Transfer functions of regular linear systems, part* 1: *Characterizations of regularity*, Trans. Amer. Math. Soc., 342 (1994), pp. 827–854.

SIAM J. CONTROL OPTIM.
Vol. 38, No. 6, pp. 1938–1984

SUPERLINEAR CONVERGENCE OF AFFINE-SCALING INTERIOR-POINT NEWTON METHODS FOR INFINITE-DIMENSIONAL NONLINEAR PROBLEMS WITH POINTWISE BOUNDS*

MICHAEL ULBRICH† AND STEFAN ULBRICH†

Abstract. We develop and analyze a superlinearly convergent affine-scaling interior-point Newton method for infinite-dimensional problems with pointwise bounds in L^p-space. The problem formulation is motivated by optimal control problems with L^p-controls and pointwise control constraints. The finite-dimensional convergence theory by Coleman and Li [*SIAM J. Optim.*, 6 (1996), pp. 418–445] makes essential use of the equivalence of norms and the exact identifiability of the active constraints close to an optimizer with strict complementarity. Since these features are not available in our infinite-dimensional framework, algorithmic changes are necessary to ensure fast local convergence. The main building block is a Newton-like iteration for an affine-scaling formulation of the KKT-condition. We demonstrate in an example that a stepsize rule to obtain an interior iterate may require very small stepsizes even arbitrarily close to a nondegenerate solution. Using a pointwise projection instead we prove superlinear convergence under a weak strict complementarity condition and convergence with Q-rate >1 under a slightly stronger condition if a smoothing step is available. We discuss how the algorithm can be embedded in the class of globally convergent trust-region interior-point methods recently developed by M. Heinkenschloss and the authors. Numerical results for the control of a heating process confirm our theoretical findings.

Key words. infinite-dimensional optimization, bound constraints, affine scaling, interior-point algorithms, superlinear convergence, trust-region methods, optimal control, nonlinear programming, optimality conditions

AMS subject classifications. 49K27, 49M15, 49M37, 65K05, 90C30, 90C48

PII. S0363012997325915

1. Introduction.

We introduce an affine-scaling interior-point Newton method for the solution of the infinite-dimensional nonlinear optimization problem

(P)
$$\begin{aligned} &\text{minimize} && f(u) \\ &\text{subject to} && u \in \mathcal{B} \stackrel{\text{def}}{=} \{u \in L^p : a(x) \le u(x) \le b(x) \text{ almost everywhere (a.e.) on } \Omega\} \end{aligned}$$

and study its local convergence behavior in detail. Here $\Omega \subset \mathbb{R}^n$ is a domain with positive and finite Lebesgue measure $0 < \mu(\Omega) < \infty$, and

$$L^t = L^t(\Omega), \quad 1 \le t \le \infty,$$

denotes the usual Banach space of (equivalence classes of) real-valued measurable functions for which the norm

$$\|u\|_t \stackrel{\text{def}}{=} \left(\int_\Omega |u(x)|^t \, dx \right)^{1/t} \quad (t < \infty), \quad \|u\|_\infty \stackrel{\text{def}}{=} \operatorname*{ess\,sup}_{x \in \Omega} |u(x)|$$

*Received by the editors August 11, 1997; accepted for publication (in revised form) December 3, 1999; published electronically August 3, 2000. The research of the first author was supported by Deutsche Forschungsgemeinschaft grant Ul157/1-1 and North Atlantic Treaty Organization grant CRG 960945. The research of the second author was supported by Deutsche Forschungsgemeinschaft grant Ul158/1-1 and North Atlantic Treaty Organization grant CRG 960945.

http://www.siam.org/journals/sicon/38-6/32591.html

†Lehrstuhl für Angewandte Mathematik und Mathematische Statistik, Technische Universität München, 80290 München, Germany (mulbrich@mathematik.tu-muenchen.de, sulbrich@mathematik.tu-muenchen.de).

is bounded. Let $2 \le p \le \infty$ and assume that the objective function $f : \mathcal{D} \longrightarrow \mathbb{R}$ is continuous on an open neighborhood $\mathcal{D} \subset L^p$ of $\mathcal{B}$. Additional requirements on f will be given below. The lower and upper bound functions $a, b \in L^\infty$ are assumed to have positive distance from each other, i.e.,

$$\operatorname*{ess\,inf}_{x\in\Omega} \ (b(x) - a(x)) > 0.$$

Then $\mathcal{B}$ has a nonempty L^∞-interior

$$\mathcal{B}^\circ \stackrel{\text{def}}{=} \bigcup_{\delta>0} \{u \in L^p \ : \ a(x) + \delta \le u(x) \le b(x) - \delta \ \text{ for almost all (a.a.) } \ x \in \Omega\}.$$

Problems of type (P) arise, for instance, when the black-box approach is applied to optimal control problems with bound-constrained L^p-control. See, e.g., the problems studied by Burger and Pogu [5], Kelley and Sachs [17], Sachs [24], and Tian and Dunn [25].

Our minimum assumptions on the objective function f follow.

ASSUMPTION.

(A1) *$f : \mathcal{D} \subset L^p \longrightarrow \mathbb{R}$ is twice continuously Fréchet differentiable with derivatives*

$$\begin{aligned} g \stackrel{\text{def}}{=} \nabla f &: \mathcal{D} \longrightarrow L^{p'}, \qquad && \frac{1}{p} + \frac{1}{p'} = 1. \\ \nabla^2 f &: \mathcal{D} \longrightarrow \mathcal{L}(L^p, L^{p'}), && \end{aligned}$$

Moreover, there is $C_g > 0$ such that $\|g(u)\|_\infty < C_g$ for all $u \in \mathcal{B}$.

Since $g(u)$ is an element of $L^{p'}$, the last requirement looks quite restrictive at first view. However, the boundedness of $\|g(u)\|_\infty$ is required only for $u \in \mathcal{B}$, which is a bounded subset of L^∞. This together with the fact that in many applications the gradient is defined via the solution of an adjoint differential equation can often be used to establish this assumption. For details we refer to section 11, where we apply our theory to the boundary control of a heating process.

The algorithm presented in this paper is based on the application of a Newton-like iteration to an affine-scaling formulation of the first-order necessary optimality conditions. For finite-dimensional problems this class of algorithms has been introduced and analyzed by Coleman and Li [6], [7]. Extensions to problems with additional equality constraints were studied in Dennis, Heinkenschloss, and Vicente [8], Heinkenschloss and Vicente [15], and Vicente [27], [28]. In all of the above papers except for [28], the affine-scaling Newton iteration is embedded in a trust-region interior-point algorithm to achieve global convergence. In a recent paper (Ulbrich, Ulbrich, and Heinkenschloss [26]) we extended the finite-dimensional global convergence theory of Coleman and Li [7] for trust-region interior-point algorithms to the infinite-dimensional problem class (P). The present paper continues these investigations and focuses on the local superlinear convergence of a closely related affine-scaling interior-point Newton method which plays the same important role in our setting as the ordinary Newton method does in the local analysis of trust-region algorithms for unconstrained optimization. Problem (P) is a special type of cone-constrained optimization problem in Banach space. For this very general class of problems Alt [2] developed a Lagrange–Newton-SQP method and proved quadratic convergence. A drawback of SQP-type methods lies in the fact that in each step a linearly cone-constrained quadratic problem or, equivalently, a linear generalized equation has to be solved. In our setting each SQP-subproblem would have the form (P) with the objective f replaced by a quadratic approximation. The solution of these problems is by no means trivial and

requires a multiple of the effort needed to perform a Newton-like step. Therefore, although SQP-methods are quadratically convergent, their efficiency crucially depends on the availability of fast solvers for the subproblems. The method proposed in this paper will be shown to converge with Q-order $1+\beta$, $0<\beta<1$, while, essentially, each iteration requires only the solution of a linear equation, which is usually much cheaper than solving an SQP-subproblem.

During the last fifteen years several attempts have been undertaken to develop algorithms for which each iteration is not much more expensive than an ordinary Newton step. One of these is the projected Newton method which was introduced by Bertsekas [3] for finite-dimensional bound-constrained problems. Kelley and Sachs [17] extended this method to problems of type (P) with special structure and proved local convergence with Q-rate $1+\beta$, $0<\beta<1$. The class of problems addressed in [17] is essentially the same as the one discussed in section 8 of this work. Although it is possible in the finite-dimensional case to prove quadratic convergence (see [3]), Kelley and Sachs could not establish this result in their infinite-dimensional setting. In this paper we develop local convergence results for infinite-dimensional affine-scaling interior-point Newton methods which are similar to those by Kelley and Sachs for projected Newton methods. Like Kelley and Sachs, we observe a gap between the achievable convergence rate in the finite- and infinite-dimensional settings. Our theory covers a more comprehensive problem class and requires weaker assumptions than that for projected Newton methods in [17]. The cost for one iteration of our algorithm is dominated by the solution of a linear equation and is therefore comparable to that of a projected Newton step.

The development of a local convergence theory for our infinite-dimensional setting turns out to be much more delicate than in the finite-dimensional case. First of all, strict complementarity, i.e., $g(\bar{u})(x)\neq 0$ for a.a. $x\in\Omega$ with $\bar{u}(x)\in\{a(x),b(x)\}$, at a local solution $\bar{u}\in\mathcal{B}$ of (P) does not guarantee that the absolute value of the gradient $g(\bar{u})$ is uniformly bounded away from zero on the active set. As a consequence, even for $u\in\mathcal{B}$ arbitrarily close to $\bar{u}$, the active set at $\bar{u}$ cannot be identified exactly by means of the information available at u. And, finally, since the L^t- and L^∞-norm, $1\le t<\infty$, are not equivalent, an iterate u_k may be very close to the solution $\bar{u}$ in L^t but still deviate substantially from $\bar{u}$ on a small set of nonzero measure. These are the main reasons why—in contrast to the finite-dimensional case—it seems not to be possible to achieve quadratic convergence in our general setting. They also make necessary several modifications of the original finite-dimensional algorithm investigated in [6] and [7] to establish superlinear convergence in the infinite-dimensional framework. Essentially, these modifications consist of the enforcement of strict feasibility by a modified projection instead of a stepsize rule and the introduction of a smoothing step to overcome the nonequivalence of norms. An example that proves the necessity of the first modification is given in Example 6.3.

The difficulties described and the necessity of modifications arise not only theoretically, but can also be observed in the finite-dimensional numerical practice. In Examples 6.3 and 6.5 we discuss in detail a scenario where our proof-driven modifications are shown to be necessary for mesh-independent convergence behavior. In this example, a simple problem of type (P) is investigated that meets all our assumptions. By discretizing problem (P) a finite-dimensional bound-constrained problem (PD) is obtained. To compute an approximate solution of (P) we apply a finite-dimensional analogue of our affine-scaling interior-point Newton method to the discretized problem (PD). The algorithm performs almost independently of the mesh size of the discretiza-

tion and converges locally superlinearly (very close to the solution even quadratically due to the finite-dimensionality of (PD)). This observed mesh-independent behavior can be viewed as the natural finite-dimensional counterpart to the availability of an infinite-dimensional convergence theory for this method. For comparison, we also apply to (PD) the affine-scaling interior-point Newton method without our modifications, which in finite dimensions is locally quadratically convergent [6], [7]. It turns out that the region of quadratic convergence is very small and, in addition, shrinks rapidly for increasingly fine meshes. Therefore, this original algorithm is not mesh-independent. This shows that the development of efficient algorithms for the solution of infinite-dimensional optimization problems also leads to improved finite-dimensional methods.

In the following we give a rough outline of the theory developed in this paper. As mentioned above, the heart of our algorithm is a Newton-like step applied to the affine-scaling formulation $d(u)g(u) = 0$ of the Karush–Kuhn–Tucker (KKT) conditions. Here $d(u) \in L^\infty$ denotes a suitably chosen weighting function, the affine-scaling function. In the Newton equation the generally nonexisting derivative of $u \longmapsto d(u)g(u)$ is replaced by an appropriate operator $G(u)$. If $u_k^s \in \mathcal{B}^\circ$ denotes the current (actually smoothed; see below) iterate, then the affine-scaling Newton step reads

$$G(u_k^s)(u_{k+1}^n - u_k^s) = -d(u_k^s)g(u_k^s).$$

The analysis of this iteration will be carried out under a regularity assumption on $G(u)$ and the following smoothness assumption.

ASSUMPTION.

(A2) *There are $2 \le q < r \le s \le \infty$, $s \ge p$, such that $g : \mathcal{B} \subset L^s \longrightarrow L^r$ is Lipschitz continuous with constant L_g and $g : \mathcal{B} \subset L^s \longrightarrow L^q$ is Lipschitz continuously Fréchet differentiable. We denote the Lipschitz constant of $\nabla g = \nabla^2 f$ by $L_{g'}$.*

We establish the estimate $\|u_{k+1}^n - \bar{u}\|_q = o(\|u_k^s - \bar{u}\|_s)$ if strict complementarity holds at the local solution $\bar{u}$ of (P), and $\|u_{k+1}^n - \bar{u}\|_q \le C\|u_k^s - \bar{u}\|_s^{1+\beta}$, $0 < \beta < 1$, if a slightly stronger strict complementarity condition is satisfied. This discrepancy of the norms is, among other things, caused by the fact that the complementarity can be arbitrarily weak on small sets. To overcome this difficulty we follow [17] and assume the availability of a smoothing step $u_k \in \mathcal{B}^\circ \longmapsto u_k^s = S_k^\circ(u_k) \in \mathcal{B}^\circ$ with $\|u_k^s - \bar{u}\|_s \le C_S\|u_k - \bar{u}\|_q$. Moreover, since u_{k+1}^n may lie outside of $\mathcal{B}^\circ$, we define a back-transport $u \longmapsto P[u_k^s](u) \in \mathcal{B}^\circ$ by an interior-point modification of the pointwise projection onto $\mathcal{B}$. We will see that a stepsize rule is inappropriate in our framework, although it yields quadratic convergence in the finite-dimensional case. We prove that the combination

$$u_k \rightsquigarrow u_k^s = S_k^\circ(u_k) \rightsquigarrow u_{k+1}^n \rightsquigarrow u_{k+1} = P[u_k^s](u_{k+1}^n)$$

generates sequences (u_k) and (u_k^s) that converge superlinearly to $\bar{u}$ in L^q and L^s, respectively. If the stronger strict complementarity condition holds we prove convergence with Q-rate $1 + \beta$. We apply our results to a class of problems with L^2-regularization for which a projected Newton method was analyzed in [17] and show that the assumptions therein imply ours. For this problem class a smoothing step can be derived from a fixed point formulation of the KKT-conditions. Moreover, we show that the second-order sufficiency condition of Dunn and Tian [9] implies our regularity assumption on G. Finally, we discuss how our algorithm can be embedded in the globally convergent class of trust-region interior-point methods recently introduced in

[26]. The resulting method is applied to the boundary control of a heating process which was already considered in [5], [20].

This paper is organized as follows. In section 2 we introduce some notation and put together several important estimates for L^p-spaces. Moreover, we state the first-order necessary optimality conditions for problem (P) in standard- and affine-scaling formulation. Our particular choice of the affine-scaling function and the basic affine-scaling Newton step are introduced in section 3. Here we discuss also why an iteration based on this step alone is, in general, neither well-defined nor convergent and sketch the idea of a smoothing step and a back-transport that take care of these problems. An outline of our algorithm and its convergence properties in a clearly arranged abstract setting is given in section 4. In section 5 we carry out a thorough analysis of the Newton-like step. In section 6 the affine-scaling interior-point Newton algorithm is formulated. Moreover, we introduce a back-transport based on a pointwise projection onto $\mathcal{B}$, explain why in our infinite-dimensional setting a stepsize-rule is not suitable for a back-transport, and address the smoothing step. Our convergence results are presented in section 7. In section 8 we apply our results to a class of L^2-regularized problems and show that our assumptions are weaker than those used in [17]. In section 9 we discuss the relationship between sufficient second-order conditions developed in [9] and the regularity assumptions we impose on the approximate derivative operator G. Section 10 addresses the question of how our algorithm can be used to accelerate the globally convergent class of trust-region interior-point algorithms recently proposed in [26]. Finally, we present numerical results for the boundary control of a heating process in section 11.

2. Preliminaries.

2.1. Notation. We write $B^c = \Omega \setminus B$ for the complement of a measurable set $B \subset \Omega$ and denote the characteristic function of B by χ_B, i.e., $\chi_B(x) = 1$ for $x \in B$, and $\chi_B(x) = 0$ otherwise. If $v : \Omega \longrightarrow \mathbb{R}$ is measurable, then we set $v_B \stackrel{\text{def}}{=} \chi_B v$. Moreover, we write $\|\cdot\|_{t,B}$ for $\|\chi_B \cdot\|_t$, $1 \le t \le \infty$.

$\mathcal{L}(Y, Z)$ is the space of bounded linear operators from the Banach space Y into the Banach space Z. The operator norm on $\mathcal{L}(L^{q_1}, L^{q_2})$ is denoted by $\|\cdot\|_{q_1,q_2}$. We write I for the identity operator $y \longmapsto y$. As representation of the dual space of L^t, $1 \le t < \infty$, we choose $L^{t'}$, $1/t + 1/t' = 1$, with the corresponding dual pairing $\langle v, w\rangle = \int_\Omega v(x)w(x)dx$, $v \in L^t$, $w \in L^{t'}$.

2.2. Some inequalities. For convenience, we recall a couple of well-known norm estimates for L^p-spaces.

LEMMA 2.1. *For all $1 \le q_1 \le q_2 \le \infty$ and $v \in L^{q_2}(\Omega)$ we have*

$$\|v\|_{q_1} \le m_{q_1,q_2}\|v\|_{q_2}$$

with $m_{q_1,q_2} = \mu(\Omega)^{\frac{1}{q_1}-\frac{1}{q_2}}$. Here $1/\infty$ has to be interpreted as zero.

Proof. See, e.g., [1, Thm. 2.8]. ∎

LEMMA 2.2 (interpolation inequality). *Given $1 \le q_1 \le q_2 \le \infty$ and $0 \le \theta \le 1$, let $1 \le q_0 \le \infty$ satisfy $1/q_0 = \theta/q_1 + (1-\theta)/q_2$. Then for all $v \in L^{q_2}$,*

$$\|v\|_{q_0} \le \|v\|_{q_1}^{\theta}\|v\|_{q_2}^{1-\theta}. \tag{1}$$

Proof. See [26, Lem. 5.2]. ∎

LEMMA 2.3. *Let $q_0 \in [1,\infty]$ and $q_1, q_1' \in [1,\infty]$ with $1/q_1 + 1/q_1' = 1$ be given. Then for all $u \in L^{q_0 q_1}$ and $v \in L^{q_0 q_1'}$ we have*

$$\|uv\|_{q_0} \le \|u\|_{q_0 q_1} \|v\|_{q_0 q_1'}.$$

Proof. In the nontrivial case $q_0 < \infty$, apply Hölder's inequality:

$$\|u\|_{q_0 q_1}^{q_0} \|v\|_{q_0 q_1'}^{q_0} = \||u|^{q_0}\|_{q_1} \||v|^{q_0}\|_{q_1'} \ge \||u|^{q_0}|v|^{q_0}\|_1 = \|uv\|_{q_0}^{q_0}. \qquad \square$$

LEMMA 2.4. *For $v \in L^q$, $1 \le q < \infty$, and all $\delta > 0$ holds*

$$\mu\left(\{x \in \Omega : \ |v(x)| \ge \delta\}\right) \le \delta^{-q} \|v\|_q^q.$$

Proof.

$$\|v\|_q^q = \||v|^q\|_1 \ge \|\chi_{\{|v| \ge \delta\}} |v|^q\|_1 \ge \mu(\{|v| \ge \delta\})\delta^q. \qquad \square$$

2.3. Necessary optimality conditions. The method is based on an affine-scaling formulation of the first-order necessary optimality conditions. A detailed derivation of these conditions can be found in [26]. Therein we also prove second-order necessary conditions which are not needed in our context.

THEOREM 2.5 (first-order necessary optimality conditions, KKT conditions). *Let $\bar u$ be a local minimizer of problem* (P) *and assume that f is differentiable at $\bar u$. In the case $p = \infty$ assume in addition that the gradient satisfies $g(\bar u) \in L^1$. Then*
(O1) $\bar u \in \mathcal{B}$,

$$\text{(O2)} \quad g(\bar u)(x) \begin{cases} = 0 & \text{for } x \in \Omega \text{ with } a(x) < \bar u(x) < b(x), \\ \ge 0 & \text{for } x \in \Omega \text{ with } \bar u(x) = a(x), \\ \le 0 & \text{for } x \in \Omega \text{ with } \bar u(x) = b(x) \end{cases} \qquad \text{a.e. on } \Omega,$$

are satisfied.

Proof. See [26, Thm. 3.1]. $\square$

The inequality (O2) can be converted into an equation by pointwise multiplication with an *affine-scaling* function $d(\bar u)$, where $d : \mathcal{B} \longrightarrow L^\infty$ satisfies

$$d(u)(x) \begin{cases} = 0 & \text{if } u(x) = a(x) \text{ and } g(u)(x) \ge 0, \\ = 0 & \text{if } u(x) = b(x) \text{ and } g(u)(x) \le 0, \\ > 0 & \text{else} \end{cases} \tag{2}$$

for a.a. $x \in \Omega$. For details we refer to [26]. The idea was first introduced by Coleman and Li in [7] for the finite-dimensional case.

LEMMA 2.6. *Let $f : \mathcal{D} \subset L^p \longrightarrow \mathbb{R}$ be differentiable and $\bar u \in \mathcal{B}$. In the case $p = \infty$ assume in addition that the gradient satisfies $g(u) \in L^1$, $u \in \mathcal{B}$. Then* (O2) *is equivalent to*

$$d(\bar u) g(\bar u) = 0 \tag{3}$$

for all d satisfying (2).

Proof. See [26, Lem. 3.2]. $\square$

3. A Newton-like step. As for all efficient methods, we aim to apply Newton's method to a suitable formulation of the optimality system. In our approach we take (3) which, according to Lemma 2.6, is equivalent to the first-order necessary condition (O2). We use the freedom provided by (2) to choose the affine-scaling function d in such a way that dg is as smooth as possible in a neighborhood of a KKT-point $\bar{u}$ of (P). A possible realization of (2) is the following function space analogue of the affine-scaling matrix of Coleman and Li [7]:

$$d_{\mathrm{I}}(u)(x) \stackrel{\text{def}}{=} \begin{cases} u(x) - a(x) & \text{if } g(u)(x) > 0 \text{ or} \\ & g(u)(x) = 0 \text{ and } u(x) \le (a(x) + b(x))/2, \\ b(x) - u(x) & \text{if } g(u)(x) < 0 \text{ or} \\ & g(u)(x) = 0 \text{ and } u(x) > (a(x) + b(x))/2. \end{cases} \tag{4}$$

In general, the mapping $u \in \mathcal{B} \subset L^\infty \longmapsto d_{\mathrm{I}}(u) \in L^\infty$ is discontinuous at a KKT-point $\bar{u}$ since $|d_{\mathrm{I}}(u) - d_{\mathrm{I}}(\bar{u})| = b - a - |u - \bar{u}|$ on $\{x \in \Omega : g(u)(x)g(\bar{u})(x) < 0\}$ (note that for all these x holds $g(\bar{u})(x) \neq 0$ and thus $d_{\mathrm{I}}(\bar{u})(x) = 0$, since $\bar{u}$ is a KKT-point). In comparison to our smoother choice of d given in (5) below, this would lead to additional complications in our theory. Our affine-scaling function d enjoys a nice Lipschitz continuity property (see Lemma 5.3), which is very convenient, especially for the investigations in sections 8 and 10. Nevertheless, our convergence theory in section 5 can be extended to the choice $d = d_{\mathrm{I}}$. One has to exploit the fact that the above-mentioned subset of Ω is small and that $d(u)g(u)$ is small on this set as well. We discuss this issue in Remark 5.15. Our affine-scaling function is defined as follows: Choose $\zeta \in (0, 1/2]$, $\kappa > 0$, and define

$$c : x \in \Omega \longmapsto \min\{\zeta(b(x) - a(x)), \kappa\}\ , \quad \nu \stackrel{\text{def}}{=} \operatorname*{ess\,inf}_{x \in \Omega}\ c(x).$$

Then our affine-scaling function is given by $d : \mathcal{B} \longrightarrow L^\infty$,

$$d(u)(x) = \begin{cases} \min\{|g(u)(x)|, c(x)\} & \text{if } -g(u)(x) > u(x) - a(x) \\ & \text{and } u(x) \le (a(x) + b(x))/2, \\ \min\{|g(u)(x)|, c(x)\} & \text{if } g(u)(x) > b(x) - u(x) \\ & \text{and } u(x) \ge (a(x) + b(x))/2, \\ \min\{u(x) - a(x), b(x) - u(x), c(x)\} & \text{else.} \end{cases} \tag{5}$$

Throughout the paper and without further notice we will frequently use the fact that

$$\|d(u)\|_\infty \le \kappa \quad \text{for all } u \in \mathcal{B}.$$

As we will see in Lemma 5.1, a suitable approximate derivative $G(u) \in \mathcal{L}(L^p, L^{p'})$ of $d(u)g(u)$ can be obtained by formally applying the product rule which yields

$$G(u) = d(u)\nabla^2 f(u) + d'(u)g(u)I \tag{6}$$

with $d' : \mathcal{B} \longrightarrow L^\infty$ suitably chosen. We recall that the only requirements on d' needed for the global convergence analysis in [26] are the conditions $d'(u)g(u) \ge 0$ and $\|d'(u)\|_\infty \le c_{d'}$ for all $u \in \mathcal{B}$. Our choice

$$d'(u) \stackrel{\text{def}}{=} \chi_{\{d(u) < c\}} \operatorname{sgn}(g(u)), \quad u \in \mathcal{B},$$

can be motivated as follows: Let $\bar{u}$ be a KKT-point and let u tend to $\bar{u}$ in L^p. Then the sets $\{g(u) > 0 \wedge d(u) = u - a \wedge g(\bar{u}) > 0\}$ tend to $\{g(\bar{u}) > 0\}$ in measure.

Analogously, the sets $\{g(u) < 0 \wedge d(u) = b - u \wedge g(\bar{u}) < 0\}$ tend to $\{g(\bar{u}) < 0\}$. On these sets, the choice $d'(u) = \operatorname{sgn}(g(u))$ is obtained by formal differentiation with respect to $u(x)$. Furthermore, $d'(u)g(u)$ tends to zero on the set $\{g(\bar{u}) = 0\}$ in $L^{p'}$ since $\|d'(u)\|_\infty$ is bounded. It turns out that the contribution of $d'(u)g(u)$ on this set is small enough for any uniformly bounded choice of $d'(u)$ to get a sufficiently good approximation $G(u)(u - \bar{u})$ of $d(u)g(u) - d(\bar{u})g(\bar{u})$ (cf. Lemma 5.1).

If u is an interior point of $\mathcal{B}$ with respect to the L^∞-norm, more precisely $u \in \mathcal{B}^\circ$, then the multiplication operator $d(u)I$ is an automorphism of L^t for all $1 \le t \le \infty$. Since our algorithm will rely on the bijectivity of $d(u_k)I$ at each iterate u_k, we require $u_k \in \mathcal{B}^\circ$ for all k. Given a current iterate $u^c \in \mathcal{B}^\circ$, we define a Newton-like step for the solution of the affine-scaling equation (3):

$$G(u^c)(u^n - u^c) = -d(u^c)g(u^c). \tag{7}$$

Let $\bar{u} \in \mathcal{B}$ be a KKT-point, i.e., $d(\bar{u})g(\bar{u}) = 0$. Then subtracting the trivial identity $G(u^c)(\bar{u} - \bar{u}) = -d(\bar{u})g(\bar{u})$ from (7) yields the equivalent equations

$$G(u^c)(u^n - \bar{u}) = R(u^c), \tag{8}$$

$$R(u) \stackrel{\text{def}}{=} d(\bar{u})g(\bar{u}) - d(u)g(u) - G(u)(\bar{u} - u). \tag{9}$$

For a classical analysis of the Newton-like iteration induced by (7) we would typically need that, for suitable q_1, q_2 and $u \in \mathcal{B}^\circ \subset L^{q_1}$ close to $\bar{u}$, the operator $G(u)$ is invertible in $\mathcal{L}(L^{q_1}, L^{q_2})$ and that $\|G(u)^{-1}R(u)\|_{q_1} = o(\|u - \bar{u}\|_{q_1})$. Moreover, for $u^c \in \mathcal{B}^\circ$ close to $\bar{u}$ the solution u^n of (7) is required to lie again in $\mathcal{B}^\circ$ to keep the iteration alive. $G(u)$ contains the multiplication operator $d'(u)g(u)I$ which for arbitrary $\varepsilon > 0$ is an automorphism of $L^{q_1}(\{x \in \Omega : |g(u)(x)| > \varepsilon\})$. Hence, $G(u) \in \mathcal{L}(L^{q_1}, L^{q_2})$ will, in general, hold only if $q_1 \ge q_2$. However, in Lemma 5.4 we will show that for $q_1 \ge q_2$ it is untenable to assume the uniform boundedness of $\|G(u)^{-1}\|_{q_2,q_1}$ in a neighborhood of $\bar{u}$. The following approach remedies the situation. With q and s as in (A2) we will introduce a multiplication operator $W(u) \in \mathcal{L}(L^q, L^q)$ such that the uniform boundedness of $(W(u)G(u))^{-1}$ in $\mathcal{L}(L^q, L^q)$ is a relatively weak requirement which is, e.g., implied by assumptions used for the analysis of a projected Newton method in [17]. Then, in Lemma 5.9 we will show that under suitable assumptions $\|W(u)R(u)\|_q = o(\|u - \bar{u}\|_s)$ holds. There seems to be no way to prove the more favorable estimate $\|W(u)R(u)\|_q = o(\|u - \bar{u}\|_q)$. Even the weaker estimate $\|R(u)\|_q = o(\|u - \bar{u}\|_q)$ requires at least the continuity of the gradient $g(u)$ from L^q to L^∞. For details see Lemma 5.1 and the proof of Lemma 5.5. Kelley and Sachs [17] overcame similar difficulties by introducing a smoothing step $u \in L^q \longmapsto u^s \in L^s$ with the property $\|u^s - \bar{u}\|_s \le \text{const}\|u - \bar{u}\|_q$. We take the same approach. Finally, it is very likely that the iteration eventually breaks down with an $u^n \notin \mathcal{B}^\circ$. Therefore, we must include a back-transport that takes u^n back into the interior of $\mathcal{B}$. This back-transport can be implemented as an interior-point modification of the pointwise projection $P(u) = \max\{a, \min\{b, u\}\}$ which satisfies $|P(u) - \bar{u}| \le |u - \bar{u}|$.

4. Outline of the algorithm in an abstract setting. The fundamental building blocks and convergence properties of the algorithm can be described most conveniently in the following abstract framework. Let X_0, X_1, and X_2 be Banach spaces, $\mathcal{K}^\circ \subset X_1$ be a convex nonempty set, and $X_1 \subset X_0$ be continuously embedded. Denote by $\mathcal{K}$ the closure of $\mathcal{K}^\circ$ in X_1. Given the mapping $E : \mathcal{K} \longrightarrow X_2$, we want to solve the equation

$$E(u) = 0, \quad u \in \mathcal{K}. \tag{10}$$

To this end, we define a Newton-like iteration based on the linear approximation $E(u+s) - E(u) \approx G(u)s$, $G : \mathcal{K}^\circ \longrightarrow \mathcal{L}(X_0, X_2)$. The kth Newton iteration is augmented by a *smoothing step* $u_k \longmapsto u_k^s = S_k^\circ(u_k) \in \mathcal{K}^\circ$ and a *back-transport* $P[v] : X_0 \longrightarrow \mathcal{K}^\circ$, $v \in \mathcal{K}^\circ$; see below. The index k indicates that the smoothing step S_k° can be chosen at runtime.

ALGORITHM 4.1 (abstract Newton iteration).

1. Choose $u_0 \in \mathcal{K}^\circ$.
2. For $k = 0, 1, 2, \ldots$
   - 2.1. If $E(u_k) = 0$, STOP.
   - 2.2. Select and perform a smoothing step: $u_k^s = S_k^\circ(u_k)$.
   - 2.3. Compute $u_{k+1}^n \in X_0$ from
   $$G(u_k^s)(u_{k+1}^n - u_k^s) = -E(u_k^s) \quad \text{(Newton-like step)}.$$
   - 2.4. Transport u_{k+1}^n back to $\mathcal{K}^\circ$: $u_{k+1} = P[u_k^s](u_{k+1}^n)$.

Let $\bar{u} \in \mathcal{K}$ be a solution to (10). Then we can rewrite the equation in step 2.3 as follows:

$$G(u_k^s)(u_{k+1}^n - \bar{u}) = E(\bar{u}) - E(u_k^s) - G(u_k^s)(\bar{u} - u_k^s) \stackrel{\text{def}}{=} R(u_k^s).$$

Algorithm 4.1 is locally superlinear convergent under the following general assumptions.

Abstract Assumptions. There are constants $\rho, C_S > 0$ such that the following hold:

1. If in the kth iteration of Algorithm 4.1 $\|u_k - \bar{u}\|_{X_0} < \rho$ holds, then in step 2.2 the smoothing step S_k° is chosen in such a way that
   $$\|S_k^\circ(u_k) - \bar{u}\|_{X_1} \le C_S \|u_k - \bar{u}\|_{X_0}.$$
2. There exists $C_P > 0$ and a monotone increasing function
   $$\delta_P : [0, C_S\rho) \longrightarrow [0, \infty), \quad \lim_{t \to 0^+} \delta_P(t) = 0, \tag{11}$$
   such that for all $u \in X_0$, $v \in \mathcal{K}^\circ$ with $\|v - \bar{u}\|_{X_1} < C_S\rho$ holds
   $$\|P[v](u) - \bar{u}\|_{X_0} \le C_P \|u - \bar{u}\|_{X_0} + \delta_P(\|v - \bar{u}\|_{X_1}) \|v - \bar{u}\|_{X_1}.$$
3. There are monotone increasing functions
   $$\gamma, \delta_R : [0, C_S\rho) \longrightarrow [0, \infty), \quad \lim_{t \to 0^+} \gamma(t)\delta_R(t) = 0, \tag{12}$$
   a Banach space X_3, and an operator $W : \mathcal{K}^\circ \longrightarrow \mathcal{L}(X_2, X_3)$ such that
   (a) for all $u \in \mathcal{K}^\circ$, $\|u - \bar{u}\|_{X_1} < C_S\rho$, and $r \in X_2$ there exists a unique $s \in X_1$ with
   $$G(u)s = r, \quad \|s\|_{X_0} \le \gamma(\|u - \bar{u}\|_{X_1}) \|W(u)r\|_{X_3};$$
   (b) for all $u \in \mathcal{K}^\circ$, $\|u - \bar{u}\|_{X_1} < C_S\rho$,
   $$\|W(u)R(u)\|_{X_3} \le \delta_R(\|u - \bar{u}\|_{X_1}) \|u - \bar{u}\|_{X_1}.$$

THEOREM 4.2. *Let $\bar{u} \in \mathcal{K}$ be a solution to* (10). *Assume that the above assumptions hold. Then there is $0 < \rho_0 \le \rho$ such that for all $u_0 \in \mathcal{K}^\circ$, $\|u_0 - \bar{u}\|_{X_0} < \rho_0$, Algorithm* 4.1 *is well-defined and either terminates with $u_k \in \mathcal{K}^\circ$ solving* (10) *or generates sequences $(u_k) \subset \mathcal{K}^\circ$ and $(u_k^s) \subset \mathcal{K}^\circ$ that converge superlinearly to $\bar{u}$ in X_0 and X_1, respectively.*

Proof. We introduce the abbreviations $\varepsilon_k \overset{\text{def}}{=} \|u_k - \bar{u}\|_{X_0}$ and $\varepsilon_k^s \overset{\text{def}}{=} \|u_k^s - \bar{u}\|_{X_1}$. Let $u_k \in \mathcal{K}^\circ$ satisfy $\varepsilon_k < \rho$. Then $\varepsilon_k^s < C_S\rho$ by Abstract Assumption 1, and thus, using the assumptions,

$$\begin{aligned}\varepsilon_{k+1} &= \|P[u_k^s](u_{k+1}^n) - \bar{u}\|_{X_0} \le C_P\|u_{k+1}^n - \bar{u}\|_{X_0} + \delta_P(\varepsilon_k^s)\varepsilon_k^s \\ &\le C_P\gamma(\varepsilon_k^s)\|W(u_k^s)R(u_k^s)\|_{X_3} + \delta_P(\varepsilon_k^s)\varepsilon_k^s \le (C_P\gamma(\varepsilon_k^s)\delta_R(\varepsilon_k^s) + \delta_P(\varepsilon_k^s))\varepsilon_k^s \\ &\le C_S(C_P\gamma(C_S\varepsilon_k)\delta_R(C_S\varepsilon_k) + \delta_P(C_S\varepsilon_k))\varepsilon_k.\end{aligned} \tag{13}$$

By (11) and (12), there is $0 < \rho_0 \le \rho$ such that

$$C_S(C_P\gamma(C_Sz)\delta_R(C_Sz) + \delta_P(C_Sz)) < 1 \quad \text{for all } 0 \le z < \rho_0.$$

Therefore, if $\varepsilon_0 < \rho_0 \le \rho$, we have $\varepsilon_k < \rho_0 \le \rho$ for all k. In particular, the algorithm is well-defined. Moreover,

$$\varepsilon_{k+1}^s \le C_S\varepsilon_{k+1} \le C_S(C_P\gamma(\varepsilon_k^s)\delta_R(\varepsilon_k^s) + \delta_P(\varepsilon_k^s))\varepsilon_k^s. \tag{14}$$

Now (13) yields superlinear convergence of (u_k) to $\bar{u}$ in X_0, and (14) yields superlinear convergence of (u_k^s) to $\bar{u}$ in X_1. ∎

Remark 4.3. It is easier to find a smoothing step $u_k \in \mathcal{K}^\circ \longmapsto S_k(u_k) \in X_1$ that satisfies all requirements in Abstract Assumption 1, except for the condition $S_k(u_k) \in \mathcal{K}^\circ$. If the operator $P[v]$ can be defined in such a way that, in addition to Abstract Assumption 2, for all $u \in X_1$ with $\|u - \bar{u}\|_{X_1} < C_S\rho$ and all $v \in \mathcal{K}^\circ$ with $\|v - \bar{u}\|_{X_0} < \rho$ holds

$$\|P[v](u) - \bar{u}\|_{X_1} \le \bar{C}_P\|u - \bar{u}\|_{X_1} + \bar{C}_P'\|v - \bar{u}\|_{X_0}, \tag{15}$$

then obviously $u_k \in \mathcal{K}^\circ \longmapsto S_k^\circ(u_k) \overset{\text{def}}{=} P[u_k](S_k(u_k))$ defines a smoothing step satisfying Abstract Assumption 1, with C_S replaced by $\bar{C}_PC_S + \bar{C}_P'$. For our problem (P) we will be able to define $P[v]$ in such a way that (15) holds; see Lemma 6.4. ∎

In our setting we have $\mathcal{K}^\circ = \mathcal{B}^\circ$ and, consequently, $\mathcal{K} = \mathcal{B}$. The mapping E is given by $u \longmapsto d(u)g(u)$. The crucial topics of our analysis consist in the proper choice of the spaces X_i, the weighting operator W, and the proof that under appropriate conditions the above abstract assumptions hold. A few remarks on the "nonstandard" building blocks of Algorithm 4.1 are in order. If there exists a projection $P : X_0 \longrightarrow \mathcal{K} \subset X_0$ onto $\mathcal{K}$ that is Lipschitz at $\bar{u}$, e.g., $P(u) = \min\{b, \max\{a, u\}\}$ for $X_0 = L^q$ and $\mathcal{K} = \mathcal{B}$, then the back-transport operator $P[v]$ can (and will) be implemented by an interior-point modification of P. More specifically, $P[v](u)$ will consist in the projection $P(u)$ of u onto $\mathcal{K}$ followed by a tiny step towards the point $v \in \mathcal{K}^\circ$ to achieve $P[v](u) \in \mathcal{K}^\circ$. The idea of a smoothing step was already used by Kelley and Sachs [17]. It is a tool to compensate the discrepancy of the X_0-norm on the left side and the stronger X_1-norm on the right side of the inequality

$$\|u_{k+1}^n - \bar{u}\|_{X_0} \le \gamma(\|u_k^s - \bar{u}\|_{X_1})\delta_R(\|u_k^s - \bar{u}\|_{X_1})\|u_k^s - \bar{u}\|_{X_1}$$

which is obtained by combining Abstract Assumptions 3(a) and (b).

5. Analysis of the Newton-like iteration. We return to the affine-scaling Newton iteration (8) and begin to verify the abstract assumptions of section 4. The following lemma states a pointwise estimate for the remainder term $R(u)$.

LEMMA 5.1. *Let* (A1) *hold. In addition, let* (O1) *and* (O2) *be satisfied at* $\bar{u}$. *Then for all* $u \in \mathcal{B}$ *the inequality*

$$|R(u)| \le d(u)|g(\bar{u}) - g(u) - \nabla^2 f(u)(\bar{u}-u)| + (|g(\bar{u})|d(u) + |g(u)||\bar{u}-u|) \tag{16}$$

holds on Ω *and, moreover,*

$$\begin{aligned} |R(u)| \le\ & d(u)|g(\bar{u}) - g(u) - \nabla^2 f(u)(\bar{u}-u)| \\ & + \min\{\max\{d(u), |g(u)|\}, |g(\bar{u}) - g(u)|\} \max\{|\bar{u}-u|, |g(\bar{u}) - g(u)|\} \end{aligned} \tag{17}$$

is satisfied on $J \stackrel{\text{def}}{=} \{x \in \Omega :\ |u(x) - \bar{u}(x)| < c(x)\}$.

Proof. Let $u \in \mathcal{B}$ be given and set $I = \{x \in \Omega :\ d(u)(x) < c(x)\}$. Then we get

$$\begin{aligned} & d(\bar{u})g(\bar{u}) - d(u)g(u) - G(u)(\bar{u}-u) \\ & \quad = d(\bar{u})g(\bar{u}) - d(u)g(u) - d(u)\nabla^2 f(u)(\bar{u}-u) - \chi_I |g(u)|(\bar{u}-u) \\ & \quad = d(u)(g(\bar{u}) - g(u) - \nabla^2 f(u)(\bar{u}-u)) + g(\bar{u})(d(\bar{u}) - d(u)) - \chi_I |g(u)|(\bar{u}-u). \end{aligned}$$

Since (O1) and (O2) are satisfied at $\bar{u}$ we have $d(\bar{u})g(\bar{u}) = 0$ by Lemma 2.6 and hence the first estimate is obvious. We complete the proof by verifying that for a.a. $x \in J$

$$\begin{aligned} R_1(u)(x) &\le \min\{\max\{d(u)(x), |g(u)(x)|\}, |g(\bar{u})(x) - g(u)(x)|\} \\ &\qquad \cdot \max\{|\bar{u}(x) - u(x)|, |g(\bar{u})(x) - g(u)(x)|\}, \\ R_1(u)(x) &\stackrel{\text{def}}{=} \big|g(\bar{u})(x)\big(d(\bar{u})(x) - d(u)(x)\big) - \chi_I(x)|g(u)(x)|\big(\bar{u}(x) - u(x)\big)\big|. \end{aligned} \tag{18}$$

We use again $d(\bar{u})g(\bar{u}) = 0$. On the subset of all $x \in J$ with $g(\bar{u})(x) = 0$ we get

$$R_1(u) = \chi_I |g(u)||\bar{u}-u| \le |g(u) - g(\bar{u})||\bar{u}-u|,$$

and (18) is obvious.

For all $x \in J$ with $g(\bar{u})(x) \ne 0$ we have $d(\bar{u})(x) = 0$. (O2) implies that only the cases $\bar{u}(x) = a(x)$ and $g(\bar{u})(x) > 0$ or $\bar{u}(x) = b(x)$ and $g(\bar{u})(x) < 0$ can occur.

We first look at $x \in J$ with $\bar{u}(x) = a(x)$ and $g(\bar{u})(x) > 0$. Since $\zeta \le 1/2$ and $x \in J$, we get $u(x) - a(x) < b(x) - u(x)$. Hence, we obtain (mind that $\bar{u}(x) = a(x)$)

$$d(u)(x) = \begin{cases} \min\{|g(u)(x)|, c(x)\} & \text{if } -g(u)(x) > u(x) - \bar{u}(x) \ge 0, \\ \min\{u(x) - \bar{u}(x), c(x)\} & \text{else.} \end{cases}$$

If $d(u)(x) = u(x) - \bar{u}(x) < c(x)$, then $x \in I$ and using $d(\bar{u})(x) = 0$, $g(\bar{u})(x) \ge 0$ we get for all these x

$$R_1(u) = \big||g(\bar{u})| - |g(u)|\big||\bar{u}-u| \le |g(\bar{u}) - g(u)||\bar{u}-u|.$$

If, in addition, $|g(\bar{u})(x) - g(u)(x)| \le d(u)(x)$, then (18) holds, for

$$R_1(u)(x) \le \min\{d(u)(x), |g(\bar{u})(x) - g(u)(x)|\}|\bar{u}(x) - u(x)|.$$

Otherwise, we have $|g(\bar{u})(x) - g(u)(x)| > d(u)(x) = u(x) - \bar{u}(x) = |\bar{u}(x) - u(x)|$, and therefore

$$R_1(u)(x) \le \min\{d(u)(x), |g(\bar{u})(x) - g(u)(x)|\}|g(\bar{u})(x) - g(u)(x)|,$$

which implies (18). If $d(u)(x) = |g(u)(x)| < c(x)$, then $x \in I$, $g(u)(x) \le 0 \le g(\bar{u})(x)$. Thus, we have for all such x that $\max\{|g(u)|, |g(\bar{u})|\} \le |g(\bar{u}) - g(u)|$ and

$$\begin{aligned} R_1(u) &= \big|-g(\bar{u})|g(u)| - |g(u)|(\bar{u} - u)\big| \\ &= |g(u)|\big||\bar{u} - u| - |g(\bar{u})|\big| \le |g(u)| \max\{|\bar{u} - u|, |g(\bar{u})|\} \\ &\le \min\{|g(u)|, |g(\bar{u}) - g(u)|\} \max\{|\bar{u} - u|, |g(\bar{u}) - g(u)|\}. \end{aligned}$$

It remains the case $x \in J \cap I^c$, i.e., $x \in J$ and $d(u)(x) = c(x)$. Here $u(x) - \bar{u}(x) \ge c(x) = d(u)(x)$ or $-g(u)(x) \ge c(x) = d(u)(x)$. Since $x \in J$, the first case cannot occur. Therefore, we have $g(u)(x) \le -d(u)(x) \le 0 \le g(\bar{u})(x)$ for all $x \in J \cap I^c$, and hence

$$R_1(u) = |g(\bar{u})d(u)| \le |g(\bar{u})||g(u)| \le \min\{|g(u)|, |g(\bar{u}) - g(u)|\}|g(\bar{u}) - g(u)|.$$

For $\bar{u}(x) = b(x)$ and $g(\bar{u})(x) < 0$ the same arguments can be used and the proof is complete. ∎

Let (O1) and (O2) hold for $\bar{u}$. We define the active set $\bar{A}$ and the inactive set $\bar{I}$:

$$\bar{A} = \{x \in \Omega : \bar{u}(x) \in \{a(x), b(x)\}\}, \quad \bar{I} = \bar{A}^c.$$

Furthermore, the usual strict complementarity condition shall hold at $\bar{u}$ (note that $|g(\bar{u})|$ is a Lagrange multiplier).

Assumption (strict complementarity condition).

(C) $g(\bar{u})(x) \ne 0$ *for a.a.* $x \in \bar{A}$.

In contrast to the finite-dimensional case the active set cannot, in general, be identified after a finite number of iterations under the strict complementarity condition (C), since the gradient may be arbitrarily small on the active set, especially near its boundary. But we shall use (C) to show that the residual set of "uncertainty" is small. We need the following continuity property of d.

Lemma 5.2. *Let the assumptions of Lemma* 2.6 *hold. In addition, let* (O1) *and* (O2) *be satisfied at* $\bar{u}$. *Then for all* $u \in \mathcal{B}$ *the inequality*

$$|d(u) - d(\bar{u})| \le \max\{|u - \bar{u}|, |g(u) - g(\bar{u})|\}$$

holds on $J \stackrel{\text{def}}{=} \{x \in \Omega : |u(x) - \bar{u}(x)| < (b(x) - a(x))/2\}$.

Proof. Let $x \in J$ be arbitrary. Since (O1) and (O2) hold at $\bar{u}$, the identity $d(\bar{u})g(\bar{u}) = 0$ is valid by Lemma 2.6. In addition, (O2) ensures that

$$d(\bar{u})(x) = \min\{\bar{u}(x) - a(x), b(x) - \bar{u}(x), c(x)\}.$$

By definition, we have

$$d(u)(x) = \min\{u(x) - a(x), b(x) - u(x), c(x)\} \quad \text{or} \tag{19}$$

$$d(u)(x) = \min\{|g(u)(x)|, c(x)\} \ge \min\{u(x) - a(x), b(x) - u(x), c(x)\}. \tag{20}$$

For all x from case (19) as well as all x with $d(u)(x) = \min\{|g(u)(x)|, c(x)\} \le d(\bar{u})(x)$ we get

$$|d(\bar{u}) - d(u)| \le |\min\{\bar{u} - a, b - \bar{u}, c\} - \min\{u - a, b - u, c\}| \le |u - \bar{u}|,$$

where we have used the inequality (cf. [26, Lem. 9.3])

$$|\min\{a_1, \dots, a_n\} - \min\{b_1, \dots, b_n\}| \le \max\{|a_1 - b_1|, \dots, |a_n - b_n|\}.$$

For all x with $d(u)(x) = \min\{|g(u)(x)|, c(x)\} > d(\bar{u})(x)$ we have

$$|d(u) - d(\bar{u})| \le d(u) \le |g(u)| \le |g(u) - g(\bar{u})|.$$

The last inequality is obvious if $g(\bar{u})(x) = 0$. For $g(\bar{u})(x) \neq 0$ it follows from the observation that $g(u)(x)$ and $g(\bar{u})(x)$ have different signs. In fact, by (O2) only the cases $\bar{u}(x) = a(x)$, $g(\bar{u})(x) > 0$ or $\bar{u}(x) = b(x)$, $g(\bar{u})(x) < 0$ can occur. If $\bar{u}(x) = a(x)$ and $g(\bar{u})(x) > 0$, then $u(x) - a(x) < b(x) - u(x)$ since $x \in J$. Hence, by the definition of $d(u)$, $d(u)(x) = \min\{|g(u)(x)|, c(x)\}$ is only possible for $g(u)(x) < 0$. Finally, if $\bar{u}(x) = b(x)$, $g(\bar{u})(x) < 0$, then $b(x) - u(x) < u(x) - a(x)$ and $d(u)(x) = \min\{|g(u)(x)|, c(x)\}$ requires $g(u)(x) > 0$. ∎

The pointwise estimates in Lemmas 5.1 and 5.2 can be converted into norm estimates by invoking Assumption (A2). As a consequence of Lemma 5.2 we get the Lipschitz continuity of d at $\bar{u}$.

LEMMA 5.3. *If* (O1), (O2) *hold at* $\bar{u}$ *and Assumptions* (A1) *and* (A2) *are satisfied, then for all* $u \in \mathcal{B}$

$$\|d(u) - d(\bar{u})\|_r \le \left(m_{r,s} + L_g + \frac{2\kappa m_{r,s}}{\nu}\right) \|u - \bar{u}\|_s \stackrel{\text{def}}{=} L_d \|u - \bar{u}\|_s$$

with $m_{r,s}$ *defined as in Lemma* 2.1.

Proof. On $B \stackrel{\text{def}}{=} \{x \in \Omega : |u(x) - \bar{u}(x)| < \nu/2\}$ Lemma 5.2 is applicable and yields with (A2) and Lemma 2.1

$$\|d(u) - d(\bar{u})\|_{r,B} \le \|\max\{|u - \bar{u}|, |g(u) - g(\bar{u})|\}\|_r \le (m_{r,s} + L_g)\|u - \bar{u}\|_s.$$

Since $|d(u)(x) - d(\bar{u})(x)| \le \kappa$, we get on B^c

$$\|d(u) - d(\bar{u})\|_{r,B^c} \le \|\kappa\|_{r,B^c} \le \left\|\kappa \frac{2|u - \bar{u}|}{\nu}\right\|_{r,B^c} \le \frac{2\kappa}{\nu}\|u - \bar{u}\|_r \le \frac{2\kappa m_{r,s}}{\nu}\|u - \bar{u}\|_s.$$

The triangle inequality completes the proof. ∎

In the finite-dimensional case the existence and uniform boundedness of $G(u)^{-1}$ in a neighborhood of $\bar{u}$ can be ensured if $\bar{u}$ satisfies sufficient second-order conditions with strict complementarity; see [7]. The following considerations show that the requirement of uniform boundedness of $G(u)^{-1}$ close to $\bar{u}$ is unacceptably strong in the infinite-dimensional setting. Since

$$g(\bar{u})(x) = 0 \quad \text{a.e. on } \bar{I}\,, \quad d(\bar{u})(x) = 0 \quad \text{a.e. on } \bar{A}, \tag{21}$$

and by (A2) and Lemma 5.3

$$\|d(u) - d(\bar{u})\|_r + \|g(u) - g(\bar{u})\|_r \le (L_d + L_g)\|u - \bar{u}\|_s,$$

the set

$$N_\varepsilon(u) \stackrel{\text{def}}{=} \{x \in \Omega : |g(u)(x)| + d(u)(x) \le \varepsilon\} \tag{22}$$

may have nonzero measure for arbitrarily small $\varepsilon > 0$ if $\|u - \bar{u}\|_s$ is small enough. Typically, an open neighborhood of a part of $\partial\bar{A}$ is contained in $N_\varepsilon(u)$.

Let $1 \le q_2 \le q_1 \le \infty$ and assume that $\|\nabla^2 f(u)\|_{q_1,q_2}$ is uniformly bounded on an L^s-neighborhood of $\bar{u}$. The following lemma shows that in the above scenario

$\|G(u)\|_{q_1,q_2}$ is uniformly bounded, but $\|G(u)^{-1}\|_{q_2,q_1}$ is not. This is caused by the fact that the operator

$$H(u) = W(u)G(u) \quad \text{with} \quad W(u) = \frac{1}{|g(u)| + d(u)} I \ , \quad u \in \mathcal{B}, \tag{23}$$

is still uniformly bounded in $\mathcal{L}(L^{q_1}, L^{q_2})$ although $\|W(u)\|_{q_2,q_2} \to \infty$ as $\|u - \bar{u}\|_s$ tends to zero. More precisely, we have the following lemma.

LEMMA 5.4. *Let $u \in \mathcal{B}^\circ$, $1 \le q_2 \le q_1 \le \infty$, and $\nabla^2 f(u) \in \mathcal{L}(L^{q_1}, L^{q_2})$. Then*

(i) $G(u) \in \mathcal{L}(L^{q_1}, L^{q_2})$, $\quad \|G(u)\|_{q_1,q_2} \le m_{q_2,q_1} \|g(u)\|_\infty + \kappa \|\nabla^2 f(u)\|_{q_1,q_2}$,

(ii) $H(u) \in \mathcal{L}(L^{q_1}, L^{q_2})$, $\quad \|H(u)\|_{q_1,q_2} \le m_{q_2,q_1} + \|\nabla^2 f(u)\|_{q_1,q_2}$,

(iii) *if $G(u)$ is invertible in $\mathcal{L}(L^{q_1}, L^{q_2})$, then*

$$\|G(u)^{-1}\|_{q_2,q_1} \ge \varepsilon^{-1} \|H(u)\|_{q_1,q_2}^{-1}$$

for all $\varepsilon > 0$ with $\mu(N_\varepsilon(u)) > 0$.

Here m_{q_2,q_1} is as in Lemma 2.1.

Proof. Assertion (i) follows immediately from the definition of $G(u)$. The estimate

$$\begin{aligned} \|H(u)v\|_{q_2} &\le \left\| \frac{\chi_{\{d(u)<c\}}|g(u)|}{|g(u)| + d(u)} v \right\|_{q_2} + \left\| \frac{d(u)}{|g(u)| + d(u)} \nabla^2 f(u) v \right\|_{q_2} \\ &\le m_{q_2,q_1} \|v\|_{q_1} + \|\nabla^2 f(u)\|_{q_1,q_2} \|v\|_{q_1} \end{aligned}$$

yields (ii). To prove (iii) let $G(u) \in \mathcal{L}(L^{q_1}, L^{q_2})$ be invertible and $\varepsilon > 0$ such that $\mu(N_\varepsilon(u)) > 0$. Then $\|w_\varepsilon\|_{q_2} > 0$ for $w_\varepsilon \stackrel{\text{def}}{=} \chi_{N_\varepsilon(u)}$, and, setting $v_\varepsilon = G(u)^{-1} w_\varepsilon$,

$$\|H(u)v_\varepsilon\|_{q_2} \le \|H(u)\|_{q_1,q_2} \|G(u)^{-1}\|_{q_2,q_1} \|w_\varepsilon\|_{q_2}.$$

On the other hand, the definition of $N_\varepsilon(u)$ yields

$$\|H(u)v_\varepsilon\|_{q_2} = \left\| \frac{w_\varepsilon}{|g(u)| + d(u)} \right\|_{q_2} \ge \frac{\|w_\varepsilon\|_{q_2}}{\varepsilon}.$$

Combining both estimates gives (iii). ∎

The identity

$$H(u) = \frac{\chi_{\{d(u)<c\}}|g(u)|}{|g(u)| + d(u)} I + \frac{d(u)}{|g(u)| + d(u)} \nabla^2 f(u)$$

shows that the operator $H(u)$ is "almost" a pointwise convex combination of the identity and the Hessian $\nabla^2 f(u)$. If (A2) and the strict complementarity condition (C) hold, then using (21) and Lemma 5.3, one can show with the same techniques as in the proof of Lemma 8.3 that

$$\frac{\chi_{\{d(u)<c\}}|g(u)|}{|g(u)| + d(u)} \xrightarrow{L^q} \chi_{\bar{A}} \quad \text{and} \quad \frac{d(u)}{|g(u)| + d(u)} \xrightarrow{L^q} \chi_{\bar{I}} \quad \text{a.e. as} \quad u \in \mathcal{B}^\circ \xrightarrow{L^s} \bar{u}.$$

Thus, they converge in all spaces L^t, $1 \le t < \infty$, by (A1) and the interpolation inequality of Lemma 2.2.

Hence, we impose the following assumption on $G(u)$ which, as we will see, is implied by the assumptions in the paper of Kelley and Sachs [17] on the projected Newton method (cf. Lemma 8.3) and in important cases by a sufficient second-order condition of Dunn and Tian [9] (see Theorem 9.5).

ASSUMPTION.

(A3) *There is* $0 < \rho_H \le 1$ *such that the operator* $H(u)$ *defined in* (23) *satisfies* $H(u) \in \mathcal{L}(L^q, L^q)$ *and is invertible for all* $u \in \mathcal{B}^\circ$, $\|u - \bar{u}\|_s < \rho_H$, *with uniformly bounded inverse, more precisely,* $\|H(u)^{-1}\|_{q,q} < C_H$.

We now return to the analysis of (8). Since for $u \in \mathcal{B}^\circ$ there is $\delta > 0$ with $d(u) > \delta$, the multiplication operator $W(u)$ defined in (23) is a linear continuous automorphism of L^t, $1 \le t \le \infty$. Applying $W(u^c)$ from the left to (8) yields the equivalent equation

$$H(u^c)(u^n - \bar{u}) = W(u^c)R(u^c). \tag{24}$$

Since $H(u^c) \in \mathcal{L}(L^q, L^q)$ is invertible by (A3) if $\|u^c - \bar{u}\|_s < \rho_H$, we derive an upper bound for the L^q-norm of the right-hand side, as follows.

LEMMA 5.5. *Let* (O1), (O2) *hold at* $\bar{u}$. *Moreover, let* (A1) *and* (A2) *be satisfied. Then for all* $u \in \mathcal{B}^\circ$ *holds*

$$\begin{aligned}\|W(u)R(u)\|_q \le L_{g'}\|u - \bar{u}\|_s^2 + (m_{r,s} + L_g)\|Q(u)\|_{\tilde{q}}\|u - \bar{u}\|_s \\ + \max\{\|g(\bar{u})\|_\infty, \|b - a\|_\infty\}\,\nu^{-\frac{s}{\tilde{q}}}\|u - \bar{u}\|_s^{\frac{s}{\tilde{q}}},\end{aligned} \tag{25}$$

where $\tilde{q} \overset{\text{def}}{=} \frac{qr}{r-q}$ $(= q$ *if* $r = \infty)$,

$$Q(u) = \frac{\min\{\max\{d(u), |g(u)|\}, |g(u) - g(\bar{u})|\}}{|g(u)| + d(u)}, \tag{26}$$

and the last term has to be interpreted as zero in the case $s = \infty$ *for* $\|u - \bar{u}\|_\infty < \nu$.

Proof. For $J \overset{\text{def}}{=} \{x \in \Omega : |u(x) - \bar{u}(x)| < \nu\}$ we may apply Lemma 5.1 and obtain with (26), (A2), and the mean value theorem

$$\begin{aligned}\|W(u)R(u)\|_q &\le \left\|\frac{d(u)}{|g(u)| + d(u)}\right\|_\infty \|g(\bar{u}) - g(u) - \nabla^2 f(u)(\bar{u} - u)\|_q \\ &\quad + \|Q(u)\max\{|u - \bar{u}|, |g(u) - g(\bar{u})|\}\|_{q,J} + \left\|\frac{|g(\bar{u})|d(u) + |g(u)||u - \bar{u}|}{|g(u)| + d(u)}\right\|_{q,J^c} \\ &\le \sup_{\tau\in[0,1]} \|\nabla^2 f(u + \tau(\bar{u} - u)) - \nabla^2 f(u)\|_{s,q}\|u - \bar{u}\|_s \\ &\quad + \|Q(u)\|_{\frac{qr}{r-q},J}\|\max\{|u - \bar{u}|, |g(u) - g(\bar{u})|\}\|_r \\ &\quad + \max\{\|g(\bar{u})\|_\infty, \|b - a\|_\infty\}\,\mu(J^c)^{1/q},\end{aligned} \tag{27}$$

where we have applied Lemma 2.3 with $q_0 = q$ and $q_1 = r/q$ in the last step. Now (A2) immediately yields the first two terms on the right-hand side of (25). To finish the proof, we first observe that $\mu(J^c) = 0$ for $\|u - \bar{u}\|_\infty < \nu$. Hence, we have (25) with the mentioned interpretation for $s = \infty$. If finally $s < \infty$, we have

$$\mu(J^c) = \|1\|_{s,J^c}^s \le \|(u - \bar{u})/\nu\|_{s,J^c}^s \le \nu^{-s}\|u - \bar{u}\|_s^s.$$

Using this in the last term of the above inequality, we get (25). ∎

It is important to notice that the term $Q(u)$ is crucial for our analysis since

$$|Q(u)(x)| = \left|\frac{\min\{\max\{d(u)(x), |g(u)(x)|\}, |g(u)(x) - g(\bar{u})(x)|\}}{|g(u)(x)| + d(u)(x)}\right| = O(1)$$

on $\{\max\{d(u), |g(u)|\} \le \text{const}|g(u) - g(\bar{u})|\}$. In contrast to the finite-dimensional case, these sets may have nonzero measure under any reasonable strict complementarity condition even if $\|u - \bar{u}\|_\infty$ is arbitrarily small. On the other hand, under

Assumption (A2) we get the following estimate on the complement of the set $N_\varepsilon(u)$ defined in (22):

$$|Q(u)(x)| \le \frac{|g(u)(x) - g(\bar{u})(x)|}{\varepsilon} \quad \text{for all } x \in N_\varepsilon(u)^c.$$

Remark 5.6. Since the estimate (25) is sharp and usually $\|Q(u)\|_\infty = O(1)$ for $u \in \mathcal{B}^\circ \xrightarrow{L^\infty} \bar{u}$, an estimate of the form

$$\|u^n - \bar{u}\|_\infty = o(\|u^c - \bar{u}\|_\infty) \quad (u^c \in \mathcal{B}^\circ \xrightarrow{L^\infty} \bar{u})$$

for the solution u^n of the affine-scaling Newton equation (7) does not, in general, hold even if (A2), (A3) are satisfied with $q = \infty$. □

The following lemma estimates the Lebesgue measure of the residual sets $N_\varepsilon(u)$.

LEMMA 5.7. *Let* (A1), (A2) *hold. If* $\bar{u}$ *satisfies* (O1), (O2), *and* (C), *then the following is true:*

(i) $\omega : [0,\infty) \longrightarrow [0,\infty)$, $\omega(\varepsilon) \stackrel{\text{def}}{=} \mu(N_\varepsilon(\bar{u}))$ *is monotone increasing and satisfies*

$$\lim_{\varepsilon \to 0+} \omega(\varepsilon) = \omega(0) = 0. \tag{28}$$

(ii) *For all* $u \in \mathcal{B}$ *holds*

$$\mu(N_\varepsilon(u)) \le \omega(2\varepsilon) + \varepsilon^{-r}(L_g + L_d)^r \|u - \bar{u}\|_s^r$$

with the obvious interpretation for $r = s = \infty$ *by setting* $\alpha^\infty = 0$ *for* $\alpha \in [0,1)$.

Proof. ω is nonnegative and increasing, since $N_{\tilde{\varepsilon}}(\bar{u}) \subset N_\varepsilon(\bar{u})$ for $0 < \tilde{\varepsilon} \le \varepsilon$. Hence, $\lim_{\varepsilon \to 0+} \omega(\varepsilon)$ exists and

$$\lim_{\varepsilon \to 0+} \omega(\varepsilon) = \mu\left(\bigcap_{\varepsilon > 0} N_\varepsilon(\bar{u})\right).$$

By (C) and the definition of d there is a set N of measure zero with

$$|g(\bar{u})(x)| + d(\bar{u})(x) > 0 \quad \text{for all } x \in N^c.$$

Hence, $N_0(\bar{u}) \subset N$ and thus $\omega(0) = \mu(N_0(\bar{u})) = 0$. Moreover, for all $x \in N^c$ there is $\varepsilon_0 > 0$ with $x \notin N_\varepsilon(\bar{u})$ for all $0 < \varepsilon < \varepsilon_0$. This shows

$$\bigcap_{\varepsilon > 0} N_\varepsilon(\bar{u}) \subset N,$$

which implies (28). To prove (ii), we use the triangle inequality and get

$$\begin{aligned} N_\varepsilon(u) &= \{|g(u)| + d(u) \le \varepsilon\} \\ &\subset \{|g(\bar{u})| + d(\bar{u}) \le \varepsilon + |g(u) - g(\bar{u})| + |d(u) - d(\bar{u})|\} \\ &\subset N_{2\varepsilon}(\bar{u}) \cup \{|g(u) - g(\bar{u})| + |d(u) - d(\bar{u})| \ge \varepsilon\}. \end{aligned}$$

In the case $r = \infty$, we have by (A2) and Lemma 5.3

$$\big\||g(u) - g(\bar{u})| + |d(u) - d(\bar{u})|\big\|_\infty \le (L_g + L_d)\|u - \bar{u}\|_\infty.$$

Hence, $N_\varepsilon(u) \subset N_{2\varepsilon}(\bar u)$ for $(L_g+L_d)\|u-\bar u\|_\infty < \varepsilon$, which is the obvious interpretation of (ii) for $r = s = \infty$. For $r < \infty$ we have by (A2) and Lemmas 2.4 and 5.3

$$\begin{aligned}\mu\left(\{|g(u)-g(\bar u)| + |d(u)-d(\bar u)| \ge \varepsilon\}\right) &\le \varepsilon^{-r}\big\||g(u)-g(\bar u)| + |d(u)-d(\bar u)|\big\|_r^r \\ &\le \varepsilon^{-r}(L_g+L_d)^r\|u-\bar u\|_s^r.\end{aligned}$$

This proves (ii). ∎

The following stronger strict complementarity condition will enable us to prove convergence with Q-rate > 1, since we get additional control on the growth of $\omega(\varepsilon)$.

Assumption (strong strict complementarity condition).

(CS) *There are* $\bar q > 0$, $C_C > 0$, *and* $\varepsilon_0 > 0$ *such that*

$$\omega(\varepsilon) = \mu(\{|g(\bar u)| + d(\bar u) \le \varepsilon\}) \le C_C\varepsilon^{\bar q} \quad \textit{for all} \;\; 0 < \varepsilon < \varepsilon_0.$$

Remark 5.8. It is easy to see that condition (CS) is satisfied if the following regularity assumptions on $\bar u$ and the active set $\bar A$ hold. They are relaxations of Assumption 2.4 in [17] as follows.

There is $c_0 > 0$ such that for all sufficiently small $\delta > 0$

$$\mu\left(\left\{x \in \Omega \,:\, \operatorname{dist}(x, \partial\bar A) \le \delta\right\}\right) \le c_0\delta$$

and for suitable $c_1 > 0$ the following growth estimates hold true:

$$\begin{aligned}|g(\bar u)(x)| &\ge c_1(\operatorname{dist}(x,\partial\bar A))^{1/\bar q} \quad &&\text{for all } x \in \bar A,\\ \min\{\bar u(x)-a(x), b(x)-\bar u(x)\} &\ge c_1(\operatorname{dist}(x,\partial\bar A))^{1/\bar q} \quad &&\text{for all } x \in \bar I = \bar A^c.\end{aligned}$$ ∎

The previous lemma enables us to estimate the norm of $Q(u)$. Together with Lemma 5.5 we get the following.

Lemma 5.9. *Let* (O1), (O2), *and* (C) *hold at* $\bar u$. *Assume that* (A1) *and* (A2) *are satisfied. Let* $\bar p \in (0,1)$ *and* $\rho \in (0,1]$ *be arbitrary such that* $(L_g+L_d)\rho^{1-\bar p} \le 1$. *Then there is* $C_{WR} > 0$ *depending only on* $\mu(\Omega)$, $\|b-a\|_\infty$, $\|g(\bar u)\|_\infty$, L_g, *and* $L_{g'}$ *but not on* q, r, s *such that for all* $u \in \mathcal{B}^\circ$, $\|u-\bar u\|_s < \rho$,

$$\|W(u)R(u)\|_q \le C_{WR}\Phi_{\bar p}(\|u-\bar u\|_s)\|u-\bar u\|_s\,, \tag{29}$$

$$\Phi_{\bar p}(z) = \omega(2z^{\bar p})^{1/\tilde q} + z^{(1-\bar p)\min\{1, r/\tilde q\}} + \left(\frac{z}{\nu}\right)^{\frac{s-q}{q}}, \tag{30}$$

where $\tilde q = qr/(r-q)$ *and* ω *is as in Lemma* 5.7.

Proof. Let $u \in \mathcal{B}^\circ$ be arbitrary with $\|u-\bar u\|_s < \rho$. According to Lemma 5.5 we have to estimate $\|Q(u)\|_{\tilde q}$ with $\tilde q = qr/(r-q)$ and Q given by (26). Let $\bar p \in (0,1)$ be arbitrary. We decompose Ω into the set

$$N(u) \overset{\text{def}}{=} N_{\|u-\bar u\|_s^{\bar p}}(u) = \left\{x \in \Omega \,:\, |g(u)(x)| + d(u)(x) \le \|u-\bar u\|_s^{\bar p}\right\}$$

and its complement $N(u)^c$. Assumption (A2) yields with the definition of $N(u)^c$

$$\|Q(u)\|_{r,N(u)^c} \le \left\|\frac{|g(u)-g(\bar u)|}{|g(u)|+d(u)}\right\|_{r,N(u)^c} \le \frac{\|g(u)-g(\bar u)\|_r}{\|u-\bar u\|_s^{\bar p}} \le L_g\|u-\bar u\|_s^{1-\bar p} \le 1. \tag{31}$$

If $\tilde q \le r$, i.e., $r \ge 2q$, one has

$$\|Q(u)\|_{\tilde q, N(u)^c} \le m_{\tilde q, r}\|Q(u)\|_{r,N(u)^c},$$

and for $\tilde{q} > r$, i.e., $q < r < 2q$, application of Lemma 2.2 with $q_0 = \tilde{q}$, $q_1 = r$, $q_2 = \infty$ yields by using $\|Q(u)\|_\infty \le 1$

$$\|Q(u)\|_{\tilde{q},N(u)^c} \le \|Q(u)\|_{r,N(u)^c}^{r/\tilde{q}}.$$

Combining this and (31) gives

$$\|Q(u)\|_{\tilde{q},N(u)^c} \le C_1 \|u-\bar{u}\|_s^{(1-\bar{p})\min\{1,r/\tilde{q}\}} \tag{32}$$

with $C_1 = L_g^{\min\{1,r/\tilde{q}\}} \max\{m_{\tilde{q},r}, 1\}$. Since $\|Q(u)\|_\infty \le 1$, we get, on the other hand, from Lemma 5.7 and Minkowski's inequality

$$\begin{aligned} \|Q(u)\|_{\tilde{q},N(u)} \le \mu(N(u))^{1/\tilde{q}} &\le \left(\omega(2\|u-\bar{u}\|_s^{\bar{p}}) + \left(\frac{(L_g+L_d)\|u-\bar{u}\|_s}{\|u-\bar{u}\|_s^{\bar{p}}}\right)^r\right)^{1/\tilde{q}} \\ &\le \omega(2\|u-\bar{u}\|_s^{\bar{p}})^{1/\tilde{q}} + (L_g+L_d)^{r/\tilde{q}} \|u-\bar{u}\|_s^{(1-\bar{p})r/\tilde{q}}. \end{aligned} \tag{33}$$

Combining (25), (32), (33), and $\|Q(u)\|_{\tilde{q}} \le \|Q(u)\|_{\tilde{q},N(u)^c} + \|Q(u)\|_{\tilde{q},N(u)}$ gives (29). Since $m_{q_1,q_2} \le \max\{1, \mu(\Omega)\}$, it is easy to see that C_{WR} depends only on the quantities listed above. ∎

Our first main result is the following.

THEOREM 5.10. *Let* (O1), (O2), *and* (C) *hold at* $\bar{u}$. *If Assumptions* (A1), (A2), *and* (A3) *are satisfied, then for all* $u^c \in \mathcal{B}^\circ$ *with* $\|u^c - \bar{u}\|_s < \rho_H$ (7) *has a unique solution* $u^n \in L^q$. *In addition, for every* $\bar{p} \in (0,1)$ *and* $0 < \rho \le \rho_H$ *satisfying* $(L_g + L_d)\rho^{1-\bar{p}} \le 1$ *there is* $C > 0$ *depending only on* $\mu(\Omega), \|b-a\|_\infty, \|g(\bar{u})\|_\infty, L_g, L_{g'}$, *and* C_H, *but not on* q, r, s *such that for all* $u^c \in \mathcal{B}^\circ$ *with* $\|u^c - \bar{u}\|_s < \rho$

$$\|u^n - \bar{u}\|_q \le C\Phi_{\bar{p}}(\|u^c - \bar{u}\|_s)\|u^c - \bar{u}\|_s \tag{34}$$

with $\Phi_{\bar{p}}$ *given by* (30).

Proof. For $u^c \in \mathcal{B}^\circ$, $\|u^c - \bar{u}\|_s < \rho_H$, the unique solvability of (7) is obvious by the assumptions. In addition now let $\|u^c - \bar{u}\|_s < \rho$ hold. Since $\bar{u}$ satisfies (O1), (O2), the equations (7) and (24) are equivalent. By the choice of $\rho > 0$ we may apply (A3) to obtain

$$\|u^n - \bar{u}\|_q \le \|H(u^c)^{-1}\|_{q,q} \|W(u^c)R(u^c)\|_q \le C_H \|W(u^c)R(u^c)\|_q.$$

Lemma 5.9 completes the proof with $C = C_H C_{WR}$. ∎

For the important case $r = s = \infty$ the proof of Theorem 5.10 can be obtained without the careful analysis of residual sets in Lemmas 5.3, 5.7, and 5.5 since these sets have measure zero for $\|u - \bar{u}\|_\infty$ small. We have the following corollary.

COROLLARY 5.11. *Under the additional assumptions* $r = s = \infty$ *and* $\rho \le \nu$, *Theorem* 5.10 *holds with* (34) *replaced by*

$$\|u^n - \bar{u}\|_q \le C\ (\omega(2\|u^c - \bar{u}\|_\infty^{\bar{p}})^{1/q} + \|u^c - \bar{u}\|_\infty^{1-\bar{p}})\|u^c - \bar{u}\|_\infty. \tag{35}$$

In the very likely case that in addition to the assumptions of Theorem 5.10, the strong strict complementarity condition (CS) holds, we get the following even stronger result.

COROLLARY 5.12. *In addition to the assumptions of Theorem* 5.10 *let condition* (CS) *hold. Then with the choice* $\bar{p} = \min\{r/(r+\bar{q}), \tilde{q}/(\tilde{q}+\bar{q})\}$ *the estimate* (34) *implies that for all* $u^c \in \mathcal{B}^\circ$ *with* $\|u^c - \bar{u}\|_s < \rho$

$$\|u^n - \bar{u}\|_q \le \bar{C}\left(\|u^c - \bar{u}\|_s^{\frac{\bar{q}}{\tilde{q}+\max\{1,\tilde{q}/r\}\bar{q}}} + \left(\frac{\|u^c - \bar{u}\|_s}{\nu}\right)^{\frac{s-q}{q}}\right)\|u^c - \bar{u}\|_s,$$

where $\tilde{q} = qr/(r-q)$ *and* $\bar{C}$ *depends on* $\mu(\Omega), \|b-a\|_\infty, \|g(\bar{u})\|_\infty, L_g, L_{g'}, C_H, C_C, \bar{q}$, *and* ε_0, *but not on* q, r, s. *For* $\rho \le (\varepsilon_0/2)^{1/\bar{p}}$ *the constant* $\bar{C}$ *can be chosen independently of* C_C, ε_0, *and* $\bar{q}$.

Proof. From (CS) we have $\omega(\varepsilon) \le C_C \varepsilon^{\bar{q}}$ for a fixed $\bar{q} > 0$ and all $\varepsilon \in]0, \varepsilon_0[$. Obviously, if we choose $C_C \ge \mu(\Omega)\varepsilon_0^{-\bar{q}}$ and remember $\omega(0) = 0$, the bound for $\omega(\varepsilon)$ holds for all $\varepsilon \ge 0$. We determine the optimal choice of $\bar{p}$ in (34) from

$$\frac{\bar{p}\bar{q}}{\tilde{q}} = (1-\bar{p})\min\{1, r/\tilde{q}\}.$$

If $r \le \tilde{q}$, this gives $\bar{p} = \frac{r}{r+\bar{q}} \le \frac{\tilde{q}}{\tilde{q}+\bar{q}}$ and the common exponent $\frac{\bar{p}\bar{q}}{\tilde{q}} = \frac{\bar{q}}{\tilde{q}+\bar{q}(\tilde{q}/r)}$. If $r > \tilde{q}$, we get $\bar{p} = \frac{\tilde{q}}{\tilde{q}+\bar{q}} < \frac{r}{r+\bar{q}}$ and $\frac{\bar{p}\bar{q}}{\tilde{q}} = \frac{\bar{q}}{\tilde{q}+\bar{q}}$. ∎

Remark 5.13. It is possible to prove an even higher convergence speed by splitting Ω in the proof of Lemma 5.9 not only in $N_{\|u-\bar{u}\|_s^{\bar{p}}}(u)$ and its complement but in $N^0(u), N^1(u)\setminus N^0(u), \dots, N^l(u)\setminus N^{l-1}(u), N^l(u)^c$, where $N^k(u) \stackrel{\text{def}}{=} N_{\|u-\bar{u}\|_s^{\bar{p}_k}}(u)$ and $1 > \bar{p}_0 > \bar{p}_1 > \cdots > \bar{p}_l > 0$. Now the $\bar{p}_k$ can be chosen in such a way that the smallest exponent is maximized. In favor of the clarity of the presentation we have not applied this more sophisticated technique. ∎

For $r = s = \infty$ we state the more handy result below.

COROLLARY 5.14. *If in Corollary* 5.11 *the condition* (C) *is replaced by the strong strict complementarity condition* (CS), *then for all* $u^c \in \mathcal{B}^\circ$ *with* $\|u^c - \bar{u}\|_\infty < \rho$

$$\|u^n - \bar{u}\|_q \le \bar{C}\|u^c - \bar{u}\|_\infty^{1+\bar{q}/(q+\bar{q})}.$$

Remark 5.15. The analysis can also be carried out for the Coleman–Li affine-scaling function d_{I} with $d_{\mathrm{I}}' \stackrel{\text{def}}{=} \operatorname{sgn}(g)$. Hereby one can proceed as follows:

(1) The estimate (16) holds and a simplified version of (17) can be established on $J = \{g(u)g(\bar{u}) > 0 \vee g(\bar{u}) = 0\}$.

(2) An analogue of Lemma 5.5 can be established if J is chosen as in (1). The last term in the second line of (27) can be estimated similar to the second one by

$$\|1\|_{\frac{qr}{r-q}, J^c}\||g(u) - g(\bar{u})| + |u-\bar{u}|\|_r \le (L_g + m_{r,s})\mu(J^c)^{\frac{r-q}{qr}}\|u-\bar{u}\|_s.$$

(3) Part (i) of Lemma 5.7 remains true and instead of (ii) one can prove

$$\mu(N_\varepsilon(u)\cap J) \le \omega(2\varepsilon) + \varepsilon^{-r}(L_g + m_{r,s})^r\|u-\bar{u}\|_s^r,$$

since $N_\varepsilon(u)\cap J \subset \{|g(u)| + d(\bar{u}) - |u-\bar{u}| \le \varepsilon\}$. Moreover, for all $\varepsilon > 0$ $J^c \subset N_\varepsilon(\bar{u}) \cap \{|g(u) - g(\bar{u})| > \varepsilon\}$ holds and thus $\mu(J^c) \le \omega(\varepsilon) + \varepsilon^{-r}L_g^r\|u-\bar{u}\|_s^r$, which shows that $\mu(J^c) \to 0$ as $u \to \bar{u}$ in L^s if the strict complementarity (C) holds.

(4) Now, one can proceed along the lines of Lemma 5.9. ∎

Assume for a moment that the iteration $u_0 \in \mathcal{B}^\circ$, $G(u_k)(u_{k+1}-u_k) = -d(u_k)g(u_k)$, is well-defined, i.e., $(u_k) \subset \mathcal{B}^\circ$ in particular. We have already observed that the sequence (u_k) may fail to converge superlinearly in L^∞ even if (A2), (A3) hold with $q = \infty$. As pointed out in [17] and [18] the same is true for directly applied projected Newton methods because the active set cannot be identified on a residual set of nonzero measure. In these papers a smoothing step is used to achieve fast L^∞-convergence. Theorem 5.10 will enable us to add such a modification if an appropriate smoothing step is available. The same problems arise also in the case $s < \infty$, since a result of the form (34) requires $s > q$.

Moreover, as we will see in Example 6.3, the case $u_{k+1} \notin \mathcal{B}^\circ$ occurs very likely for some k. Hence, a back-transport into $\mathcal{B}^\circ$ is necessary. Therefore, we will use the following ingredients to design a superlinearly convergent algorithm (cf. the outline in section 4).

SMOOTHING STEP: $u_k \in \mathcal{B}^\circ \longmapsto u_k^s = S_k^\circ(u_k) \in \mathcal{B}^\circ$ with $\|u_k^s - \bar{u}\|_s \le C_S \|u_k - \bar{u}\|_q$.

NEWTON STEP: $u_{k+1}^n \in L^q$ solves $G(u_k^s)(u_{k+1}^n - u_k^s) = -d(u_k^s)g(u_k^s)$.

BACK-TRANSPORT: $u_{k+1}^n \in L^q \longmapsto u_{k+1} = P[u_k^s](u_{k+1}^n) \in \mathcal{B}^\circ$
with $\|u_{k+1} - \bar{u}\|_q \le C_P \|u_{k+1}^n - \bar{u}\|_q + C_P' \|u_k^s - \bar{u}\|_s^2$.

Here C_S, C_P, and C_P' are positive constants.

5.1. An affine-scaling Newton algorithm. Provided that smoothing step and back-transport with the above properties are available, the previous considerations and the abstract convergence theory in section 4 suggest the following algorithm.

ALGORITHM 5.16 (affine-scaling interior-point Newton algorithm).

1. Choose $u_0 \in \mathcal{B}^\circ$.
2. For $k = 0, 1, 2, \ldots$
   - 2.1. If $d(u_k)g(u_k) = 0$, STOP.
   - 2.2. Select and perform a smoothing step: $u_k^s = S_k^\circ(u_k)$.
   - 2.3. Compute $u_{k+1}^n \in L^q$ from
   $$G(u_k^s)(u_{k+1}^n - u_k^s) = -d(u_k^s)g(u_k^s) \quad \text{(affine-scaling Newton step)}.$$
   - 2.4. Transport u_{k+1}^n back to $\mathcal{B}^\circ$: $u_{k+1} = P[u_k^s](u_{k+1}^n)$.

6. Back-transport and smoothing-step.

6.1. The back-transport. Since the solution u_{k+1}^n of the affine-scaling Newton equation in step 2.3 is not necessarily an interior point of $\mathcal{B}$, a back-transport into $\mathcal{B}^\circ$ is needed. In [7] a stepsize rule is used for this purpose. A reflection technique was proposed in [4] and [6]. We will see that in our function space, setting very small stepsizes σ_k may be necessary to achieve $u_k^s + \sigma_k(u_{k+1}^n - u_k^s) \in \mathcal{B}^\circ$. Thus, a stepsize rule fails to provide superlinear convergence; cf. Example 6.3. Therefore, we will propose and analyze a projection technique which is also an attractive alternative to reflection techniques in the finite-dimensional case.

6.1.1. Back-transport by projection. Since $\bar{u} \in \mathcal{B}$, the pointwise projection $P(u)$ of u onto $\mathcal{B}$ with $P : L^1 \longrightarrow \mathcal{B}$ defined by

$$P(u) = \max\{a, \min\{b, u\}\} \tag{36}$$

obviously satisfies $|P(u) - v| \leq |u - v|$ on Ω for all $v \in \mathcal{B}$. Hence,

$$\|P(u) - v\|_t \leq \|u - v\|_t \tag{37}$$

for all $t \in [1, \infty]$, $v \in \mathcal{B}$, and $u \in L^t$.

As mentioned earlier, an interior-point modification $P[v]$, $v \in \mathcal{B}^\circ$, of P can be used to obtain a back-transport satisfying the required property

$$\|P[v](u) - \bar{u}\|_q \leq C_P \|u - \bar{u}\|_q + C'_P \|v - \bar{u}\|_s^2. \tag{38}$$

In fact, for $\xi \in (0, 1)$, typically $\xi > 0.9$, and $v \in \mathcal{B}^\circ$ choose

$$P[v] : L^q \longrightarrow \mathcal{B}^\circ, \quad P[v](u) = v + \max\{\xi, 1 - \|P(u) - v\|_q\}(P(u) - v). \tag{39}$$

Then obviously

$$P[v](u) - P(u) = \min\{1 - \xi, \|P(u) - v\|_q\}(v - P(u)),$$

and hence

$$\begin{aligned} &\|P[v](u) - P(u)\|_q \leq \|P(u) - v\|_q^2, \\ &\|P[v](u) - P(u)\|_t \leq \|b - a\|_t \|P(u) - v\|_q \ , \quad 1 \leq t \leq \infty. \end{aligned} \tag{40}$$

Using this, we can derive (38).

LEMMA 6.1. *Let P and $P[v]$, $v \in \mathcal{B}^\circ$, be defined according to* (36) *and* (39). *Then condition* (38) *holds with $C_P = (2\|b - a\|_q + 1)$, $C'_P = 2m_{q,s}^2$.*

Proof. Let $v \in \mathcal{B}^\circ$ and $u \in L^q$. Using the properties of $P[v]$ yields

$$\begin{aligned} \|P[v](u) - \bar{u}\|_q &\leq \|P[v](u) - P(u)\|_q + \|P(u) - \bar{u}\|_q \leq \|P(u) - v\|_q^2 + \|u - \bar{u}\|_q \\ &\leq 2\ (\|P(u) - \bar{u}\|_q^2 + \|v - \bar{u}\|_q^2) + \|u - \bar{u}\|_q \\ &\leq (2\|b - a\|_q + 1)\|u - \bar{u}\|_q + 2\|v - \bar{u}\|_q^2 \\ &\leq (2\|b - a\|_q + 1)\|u - \bar{u}\|_q + 2m_{q,s}^2 \|v - \bar{u}\|_s^2. \end{aligned}$$

∎

6.1.2. Projection vs. stepsize rule for back-transport. The following arguments and Example 6.3 below show that even if (A2), (A3) hold for $q = \infty$ and $\|u_k^s - \bar{u}\|_\infty$ is arbitrarily small, stepsizes $\sigma_k \leq \varepsilon \ll 1$ may be necessary to ensure $u_k^s + \sigma_k(u_{k+1}^n - u_k^s) \in \mathcal{B}^\circ$. Let $u_k^s \in \mathcal{B}^\circ$ be arbitrary. From step 2.3 in Algorithm 5.16 we deduce for x with $d(u_k^s)(x) < c(x)$ and $g(u_k^s)(x) \neq 0$

$$u_k^s(x) - u_{k+1}^n(x) = \left(\operatorname{sgn}(g(u_k^s)(x)) + \frac{(\nabla^2 f(u_k^s)(u_{k+1}^n - u_k^s))(x)}{|g(u_k^s)(x)|} \right) d(u_k^s)(x). \tag{41}$$

If we look at those $x \in \Omega$ where in addition $\bar{u}(x) = a(x)$ and $|g(u_k^s)(x)|$ is small, say, $|g(u_k^s)(x)| \leq u_k^s(x) - a(x)$, then

$$d(u_k^s)(x) = u_k^s(x) - a(x).$$

Thus, we need stepsize $\sigma_k \leq \varepsilon$ if

$$\left(\nabla^2 f(u_k^s)(u_{k+1}^n - u_k^s)\right)(x) \geq \left(1 + \varepsilon^{-1}\right) |g(u_k^s)(x)|,$$

since (41), the equality for $d(u_k^s)(x)$, and the last inequality imply

$$u_k^s(x) - u_{k+1}^n(x) \geq \varepsilon^{-1}(u_k^s(x) - a(x)).$$

But even for $\|u_k^s - \bar{u}\|_\infty$ arbitrarily small the set

$$\left\{x \in \Omega : \ \bar{u}(x) = a(x), \ \left(\nabla^2 f(u_k^s)(u_{k+1}^n - u_k^s)\right)(x) \geq \left(1 + \varepsilon^{-1}\right)|g(u_k^s)(x)|\right\}$$

may have nonzero measure, because $|g(u_k^s)|$ is typically very small on a neighborhood of $\partial\bar{A}$.

Since superlinear convergence can be guaranteed only if the sequence of stepsizes converges to one, a stepsize rule for the Newton-like step is unsuitable for the infinite-dimensional case although it was proven to give quadratic convergence in the finite-dimensional case (see [7]).

Remark 6.2. In the finite-dimensional case one can easily show by using a componentwise version of (41) that $\sigma_k = 1 - O(\|u_{k+1} - u_k\|)$ if second-order sufficiency conditions with strict complementarity hold at $\bar{u}$. See [6], [7]. □

The following example illustrates that in the infinite-dimensional case the above scenario of small stepsizes can really occur.

Example 6.3. We consider problem (P) with quadratic objective function

$$f : u \in L^2([0,1]) \longmapsto \frac{1}{2}\|u\|_2^2 - \frac{1}{4}\left(\int_0^1 u(x)\,dx\right)^2$$

and feasible set $\mathcal{B} \stackrel{\text{def}}{=} \{u \in L^2([0,1]) : \ a(x) \stackrel{\text{def}}{=} x - \frac{1}{2} \leq u(x) \leq 10 \stackrel{\text{def}}{=} b(x) \text{ a.e.}\}$. f is smooth with

$$g(u) = u - \frac{1}{2}\int_0^1 u(x)\,dx\ , \quad \nabla^2 f(u)\,v = g(v),$$

and strictly convex, since by Jensen's inequality $(v, \nabla^2 f(u)v)_2 \geq \frac{1}{2}\|v\|_2^2$ for all $v \in L^2$. The unique global minimum of f on $\mathcal{B}$ is given by $\bar{u}(x) = \max\{y, a(x)\}$ with $y = 3/2 - \sqrt{2}$, because f is strictly convex, $g(\bar{u}) = \bar{u} - y = 0$ on the inactive set $\bar{I} = [0, \hat{x})$, $\hat{x} = y + 1/2$, and $g(\bar{u}) = \bar{u} - y \geq 0$ on the active set $\bar{A} = [\hat{x}, 1] = \{\bar{u} = a\}$. It is easy to check that (A1)–(A3), (C), and (CS) hold for $p = q = 2$, $r = s = \infty$. For $0 < \varepsilon < 1$ the function $u_\varepsilon \in \mathcal{B}^\circ$, $u_\varepsilon(x) \stackrel{\text{def}}{=} \bar{u}(x) + \varepsilon|x - \hat{x}| + \varepsilon^2/10$, is strictly feasible with $\|u_\varepsilon - \bar{u}\|_\infty = \hat{x}\varepsilon + \varepsilon^2/10 < \varepsilon$. Moreover, the gradient $g(u_\varepsilon)$ is negative in a neighborhood of the boundary point $\hat{x}$ of $\bar{A}$ which leads to the above scenario of small stepsizes:

$$g(u_\varepsilon)(\hat{x}) = \frac{\varepsilon}{20}\,(-45 + 30\sqrt{2} + \varepsilon) < -\frac{\varepsilon}{20}, \quad 0 < \varepsilon < 1.$$

Now we analyze what happens if we take u_ε as starting point for an affine-scaling Newton step s_ε, i.e., $G(u_\varepsilon)s_\varepsilon = -d(u_\varepsilon)g(u_\varepsilon)$, or, in detail,

$$s_\varepsilon - \frac{1}{2}\frac{d(u_\varepsilon)}{d'(u_\varepsilon)g(u_\varepsilon) + d(u_\varepsilon)}\int_0^1 1 \cdot s_\varepsilon(x)\,dx = -\frac{d(u_\varepsilon)g(u_\varepsilon)}{d'(u_\varepsilon)g(u_\varepsilon) + d(u_\varepsilon)}. \tag{42}$$

Since the operator $(d'(u_\varepsilon)g(u_\varepsilon) + d(u_\varepsilon))^{-1}G(u_\varepsilon)$ on the left (which coincides with $H(u_\varepsilon)$ for ε small enough) is a "rank-one modification" of the identity, its inverse

TABLE 1

ε	$\sigma_{\max}$	$\frac{\|(u_\varepsilon+s_\varepsilon)-P(u_\varepsilon+s_\varepsilon)\|_2}{\|s_\varepsilon\|_2}$
1.0E−2	1.77E−2	4.88E−3
1.0E−3	1.78E−3	1.41E−3
1.0E−4	1.78E−4	4.43E−4
1.0E−10	1.78E−10	4.42E−6
1.0E−14	1.78E−14	4.42E−8

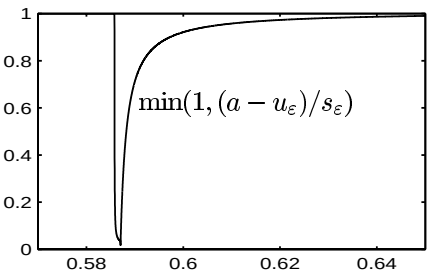


FIG. 1.

can be explicitly determined by applying the Sherman–Morrison–Woodbury lemma in $L^2([0,1])$. It is possible to derive an explicit closed formula for the function s_ε. Table 1 shows the maximum stepsize $\sigma_{\max} \stackrel{\text{def}}{=} \max\{\sigma \in [0,1] : u_\varepsilon + \sigma s_\varepsilon \in \mathcal{B}\}$ for $c \stackrel{\text{def}}{=} 2$, and the relative L^2-norm of the part of s_ε that would be cut off by a pointwise projection. All values were obtained using the closed formula for s_ε. Figure 1 depicts for $\varepsilon = 1/100$ a plot of the pointwise allowed maximum stepsize. We see that the severe restriction on the stepsize stems from points in a small neighborhood of $\hat{x} \approx 0.5857$. This neighborhood contains a zero of g_ε located in the active set which leads exactly to the scenario discussed above. Outside of the neighborhood depicted in Figure 1 the pointwise allowed maximum stepsize is ≥ 0.99.

Thus, u_ε is a sequence tending to $\bar{u}$ in L^∞ as $\varepsilon \to 0^+$ for which the corresponding maximum stepsizes $\sigma_{\max}$ tend to zero whereas the relative L^2-difference between the affine-scaling Newton step s_ε and the projected step $P(u_\varepsilon + s_\varepsilon) - u_\varepsilon$ approaches zero. □

The previous example shows that a stepsize rule may lead to arbitrarily small stepsizes for iterates that are arbitrarily close to the solution $\bar{u}$. Thus, it has to be replaced by another back-transport to achieve fast local convergence. On the other hand, the example motivates the back-transport by projection introduced in section 6.1.1, since it leads only to a tiny change of the step in the L^2-norm.

After the introduction of a smoothing step in the next section, this example will be continued in Example 6.5. There we will illustrate (among other things) that the use of a stepsize rule may lead to almost a stagnation of the iteration whereas the proposed projection technique in conjunction with a smoothing step yields fast convergence.

6.2. The smoothing step. We have already observed that a smoothing step is necessary because the strongest available estimate after one iteration of (7) is (34) with $q < s$. In the further analysis we assume that in each iteration k of Algorithm

5.16 a—possibly noninterior—smoothing step

$$u_k \in \mathcal{B} \longmapsto S_k(u_k) \in L^s \tag{43}$$

is available. Let $\bar{u} \in \mathcal{B}$ satisfy (O1) and (O2). Analogously to the Abstract Assumption 1 in section 4 we make the following assumption.

ASSUMPTION (smoothing property).

(S) *There are $\rho_S > 0$ and $L_S > 0$ such that the smoothing steps S_k defined in* (43) *possess the following property:*

$$\|S_k(u_k) - \bar{u}\|_s \le L_S \|u_k - \bar{u}\|_q \quad \text{for all } k \text{ with } \|u_k - \bar{u}\|_q < \rho_S.$$

This assumption allows us to choose $S_k(u_k) = u_k$ if $u_k \in \{u : \|u - \bar{u}\|_s \le L_S \|u - \bar{u}\|_q\}$. As already outlined in Remark 4.3, the smoothing step

$$u_k \longmapsto S_k^\circ(u_k) \stackrel{\text{def}}{=} P[u_k](S_k(u_k)) \tag{44}$$

is an interior-point modification of S_k as follows.

LEMMA 6.4. *Let $\bar{u}$ satisfy* (O1), (O2) *and let* (S) *hold. If $P[v]$ is defined by* (39), *then S_k° as defined in* (44) *satisfies $S_k^\circ(u_k) \in \mathcal{B}^\circ$. Moreover,*

$$\|S_k^\circ(u_k) - \bar{u}\|_s \le C_S \|u_k - \bar{u}\|_q$$

holds for all k with $\|u_k - \bar{u}\|_q < \rho_S$, where $C_S = (m_{q,s}\|b - a\|_s + 1)L_S + \|b - a\|_s$.

Proof. $P[u_k](S_k(u_k)) \in \mathcal{B}^\circ$ does obviously hold for all $k \ge 0$, since $u_k \in \mathcal{B}^\circ$. Now let $k \ge 0$ be arbitrary with $\|u_k - \bar{u}\|_q < \rho_S$. Using the properties (37), (40) of P and $P[u_k]$, we get

$$\begin{aligned}
\|P[u_k](S_k(u_k)) - \bar{u}\|_s &\le \|P[u_k](S_k(u_k)) - P(S_k(u_k))\|_s + \|P(S_k(u_k)) - \bar{u}\|_s \\
&\le \|b - a\|_s \|P(S_k(u_k)) - u_k\|_q + \|S_k(u_k) - \bar{u}\|_s \\
&\le \|b - a\|_s (\|P(S_k(u_k)) - \bar{u}\|_q + \|u_k - \bar{u}\|_q) + \|S_k(u_k) - \bar{u}\|_s \\
&\le (m_{q,s}\|b - a\|_s + 1)\|S_k(u_k) - \bar{u}\|_s + \|b - a\|_s \|u_k - \bar{u}\|_q \\
&\le C_S \|u_k - \bar{u}\|_q,
\end{aligned}$$

where we have used (S) in the last step. ∎

We will show in section 8 how a smoothing step can be constructed for a class of regularized problems by using a fixed point formulation of the KKT-conditions (O1), (O2).

In the following example we will demonstrate that our proof-driven modifications of the basic interior-point Newton method—to replace the stepsize rule by a projection based procedure and to augment the iteration with a smoothing step—are actually necessary to achieve mesh-independence after discretization.

Example 6.5. We return to Example 6.3 to compare the performance of three variants of the affine-scaling interior-point Newton method:

(I) Algorithm 5.16.

(II) Algorithm 5.16 with 2.4 replaced by a stepsize rule:

2.4′ Transport u_{k+1}^n back to $\mathcal{B}^\circ$ by the following stepsize rule:

$$\begin{aligned}
s_k^n &= u_{k+1}^n - u_k^s, \quad \sigma_{k,\max} = \max\{\sigma \in [0,1] : u_k^s + \sigma s_k^n \in \mathcal{B}\}, \\
u_{k+1} &= u_k^s + \max\{\xi, 1 - \|s_k^n\|_2\}\, \sigma_{k,\max} s_k^n, \quad \xi \text{ as in (39)}.
\end{aligned}$$

(III) Algorithm 5.16 without smoothing step 2.2.

In Algorithms (I) and (II) we apply smoothing to u_k only if the L^2- and L^∞-norm of $u_k - u^s_{k-1}$ differ too much:

$$S^\circ_k(u_k) \stackrel{\text{def}}{=} \begin{cases} P[u_k](u_k - g(u_k)) & \text{if } k \geq 1 \text{ and } \|u_k - u^s_{k-1}\|_\infty \geq 3\|u_k - u^s_{k-1}\|_2, \\ u_k & \text{else.} \end{cases}$$

The interior-point modification of the projected gradient step is indeed a smoothing step. This follows from Lemma 6.4 and the discussion in section 8. For the numerical realization of the methods we have discretized the problem by approximating $L^2([0,1])$ with piecewise linear functions on a uniform grid with $N+1$ points for $N = 200, 2000$, and 200000, respectively. To check the decrease properties of the new iterates, we use the fact that u^n_{k+1} solves the affine-scaling Newton equation in step 2.3 if and only if u^n_{k+1} is a stationary point of the quadratic function $\psi[u^s_k](u)$ defined by (62); cf. section 10. This function is used as quadratic model in the interior-point trust-region methods recently analyzed in [26]. Since $\psi[u^s_k]$ is strictly convex in our context, it attains its global minimum at s^n_k. We start the iterations with $u_0 = u_\varepsilon$, $\varepsilon = 0.5$, and $\xi = 0.999995$. We begin with a comparison of Algorithms (I) and (II) for $N = 200$.

TABLE 2

	Alg. (I) (Projection)			Alg. (II) (Stepsize rule)			
k	$\|u_{k+1} - \bar{u}\|_2$	$\|u^s_{k+1} - \bar{u}\|_\infty$	r_k †	$\|u_{k+1} - \bar{u}\|_2$	$\|u^s_{k+1} - \bar{u}\|_\infty$	r_k †	$\sigma_{k,\max}$
0	4.2275E−2	8.1290E−2	0.9999	7.6898E−2	1.4443E−1	0.9288	0.7332
1	4.2356E−3	8.5289E−3	0.9998	7.6053E−2	1.4287E−1	0.0255	0.0128
2	1.8431E−4	1.5081E−3	1.0000	7.4395E−2	1.3981E−1	0.0502	0.0254
3	4.3224E−5	3.1222E−6*	1.0000	7.1203E−2	1.3391E−1	0.0973	0.0499
4	6.3244E−10	8.7489E−9	1.0000	6.5276E−2	1.2295E−1	0.1833	0.0963

† $r_k = \psi[u^s_k](u_{k+1})/\psi[u^s_k](u^n_{k+1})$. * Smoothing occurred, i.e., $u^s_{k+1} \neq u_{k+1}$.

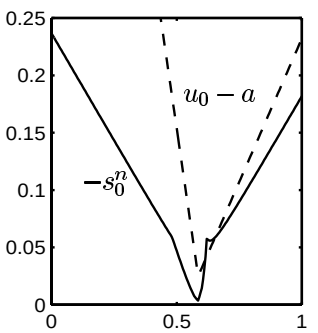


FIG. 2

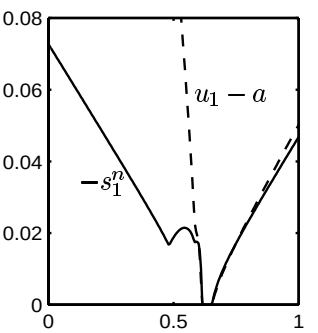


FIG. 3

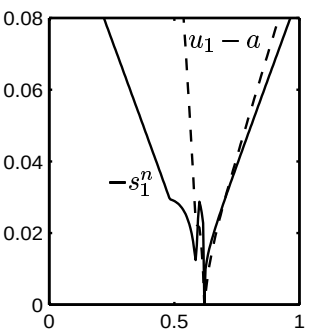


FIG. 4

The distances $\|u_{k+1} - \bar{u}\|_2$, $\|u^s_{k+1} - \bar{u}\|_\infty$ from the solution $\bar{u}$ of the discrete problem and the decrease ratio $\psi[u^s_k](u_{k+1})/\psi[u^s_k](u^n_{k+1})$ are shown in Table 2. For method (II) we have also added $\sigma_{k,\max}$. Figure 2 depicts $-s^n_0$ and the distance to the lower bound $u_0 - a$ (dashed). Figures 3 and 4 show the same quantities, i.e., $-s^n_1$ and $u_1 - a$, after one step of Algorithms (I) and (II), respectively. We see that the stepsize rule in (II) leads to an iterate u_1 and a new search direction s^n_1 that requires a very small stepsize of 0.0128 yielding almost no progress. The reason is depicted in Figure 4:

s_1^n has a small peak on the set where the distance to the lower bound is small. On the other hand, the part of $s_k^n = u_{k+1}^n - u_k^s$ that is cut off by a projection is very small. Hence, the projection leads to a nearly optimal decrease of $\psi[u_k^s]$ in every step. Figures 5 and 6 show the first iterates for both iterations. While our Algorithm (I) converges in 5 steps to high accuracy, method (II) needs 28 iterations to enter the region of quadratic convergence which exists according to the finite-dimensional theory. Then it converges in two more steps to high accuracy. Using the starting points of Example 6.3 for sufficiently fine discretizations leads to approximately the same small stepsizes as in Example 6.3. Thus, Algorithm (II) is not mesh-independent and its region of quadratic convergence shrinks with the grid-size. Since for the starting point u_ε, $\varepsilon = 0.5$, also in this example the performance of Algorithm (II) is worse for finer grids—it needs 270 iterations for $N = 2000$—we compare only the mesh-dependence of Algorithms (I) and (III). Table 3 contains the distances $\|u_{k+1} - \bar{u}\|_2$, $\|u_{k+1}^s - \bar{u}\|_\infty$ from the solution $\bar{u}$ of the discrete problems for $N = 2000$ and $N = 200000$. As indicated by our results we observe mesh-independent superlinear convergence of Algorithm (I). On the other hand, we see that Algorithm (III) is *not* mesh-independent and that the region of fast local convergence obviously shrinks. □

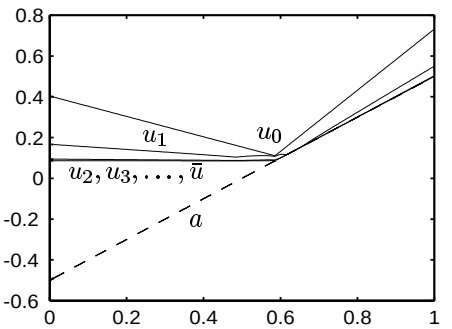


FIG. 5

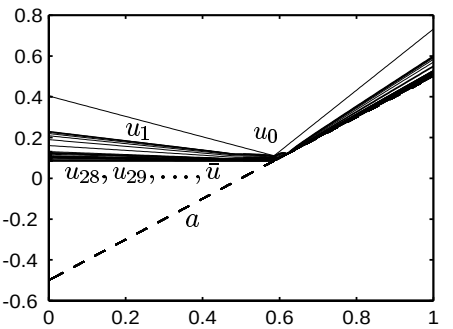


FIG. 6

TABLE 3

	$N = 2000$				$N = 200000$			
	Alg. (I)		Alg. (III)		Alg. (I)		Alg. (III)	
k	$\|e_{k+1}\|_2$ †	$\|e_{k+1}^s\|_\infty$ †	$\|e_{k+1}\|_2$	$\|e_{k+1}\|_\infty$	$\|e_{k+1}\|_2$	$\|e_{k+1}^s\|_\infty$	$\|e_{k+1}\|_2$	$\|e_{k+1}\|_\infty$
0	4.2E−02	8.1E−02	4.2E−02	8.1E−02	4.2E−02	8.1E−02	4.2E−02	8.1E−02
1	4.2E−03	8.5E−03	4.2E−03	8.5E−03	4.2E−03	8.5E−03	4.2E−03	8.5E−03
2	1.8E−04	1.6E−03	1.8E−04	1.6E−03	1.8E−04	1.6E−03	1.8E−04	1.6E−03
3	3.8E−05	2.9E−06*	3.8E−05	7.2E−04	3.8E−05	2.9E−06*	3.8E−05	8.1E−04
4	5.2E−10	2.0E−08	1.2E−05	3.1E−04	2.8E−09	9.2E−07	1.2E−05	4.1E−04
5	3.2E−14	4.4E−16*	3.5E−06	1.2E−04	5.9E−10	1.2E−12*	4.0E−06	2.0E−04
6	1.2E−16	1.5E−16	8.2E−07	3.0E−05	5.0E−14	6.5E−14	1.4E−06	1.0E−04
7			8.0E−08	3.3E−06			4.8E−07	5.0E−05
8			1.1E−09	5.0E−08			1.7E−07	2.5E−05
9			2.7E−13	1.2E−11			6.0E−08	1.2E−05
10			2.0E−16	2.6E−16			2.1E−08	5.6E−06
11							7.3E−09	2.5E−06
12							2.4E−09	9.7E−07
13							6.3E−10	2.8E−07
14							8.8E−11	3.9E−08
15							2.3E−12	1.0E−09
16							1.3E−14	7.3E−13

† $e_{k+1} = u_{k+1} - \bar{u}$, $e_{k+1}^s = u_{k+1}^s - \bar{u}$. * Smoothing occurred, i.e., $u_{k+1}^s \neq u_{k+1}$.

The above example shows that both of our algorithmic modifications, the projection as back-transport and the smoothing step, are necessary to obtain mesh-independent superlinear convergence. This demonstrates that the finite-dimensional convergence theory for Algorithm (II), which guarantees local quadratic convergence, is not sufficient to ensure mesh-independent behavior.

7. The convergence result. In the following we will always work with the smoothing steps

$$u_k \in \mathcal{B}^\circ \longmapsto S_k^\circ(u_k) \stackrel{\text{def}}{=} P[u_k](S_k(u_k)) \in \mathcal{B}^\circ, \quad S_k \text{ as in (43)}.$$

We will now prove that Algorithm 5.16 converges superlinearly (respectively, with Q-order $1 + \bar{q}/(\tilde{q} + \max\{1, \tilde{q}/r\}\,\bar{q})$) in L^s to $\bar{u}$ if $\bar{u}$ satisfies the first-order necessary conditions with strict complementarity (C) (respectively, (CS)) as well as (A3), a smoothing step exists, and $\|u_0 - \bar{u}\|_q$ is small enough. More precisely, we have the following theorem.

THEOREM 7.1. *Let $\bar{u}$ satisfy* (O1), (O2), *and* (C). *If* (A1)–(A3) *and* (S) *hold, then for $\bar{p} \in (0,1)$ there is $\rho > 0$ such that for all $u_0 \in \mathcal{B}^\circ$ with $\|u_0 - \bar{u}\|_q < \rho$ Algorithm* 5.16 *is well-defined and produces iterates with*

$$\|u_{k+1} - \bar{u}\|_q \le \bar{C}_1 \Phi_{\bar{p}}(C_S \|u_{k+1} - \bar{u}\|_q) \|u_{k+1} - \bar{u}\|_q, \tag{45}$$

$$\|u_{k+1}^s - \bar{u}\|_s \le \bar{C}_2 \Phi_{\bar{p}}(\|u_{k+1}^s - \bar{u}\|_s) \|u_{k+1}^s - \bar{u}\|_s, \tag{46}$$

where $\bar{C}_1, \bar{C}_2 > 0$ depend on $\mu(\Omega), \|b - a\|_\infty, \|g(\bar{u})\|_\infty, L_g, L_{g'}, C_H, L_S$, but not on q, r, s and $\Phi_{\bar{p}}$ is given by (30), *i.e.,*

$$\Phi_{\bar{p}}(z) = \omega(2z^{\bar{p}})^{1/\tilde{q}} + z^{(1-\bar{p})\min\{1, \frac{r}{\tilde{q}}\}} + \left(\frac{z}{\nu}\right)^{\frac{s-q}{q}}.$$

In the case $r = s = \infty$ the function $\Phi_{\bar{p}}$ simplifies to $\Phi_{\bar{p}}(z) = \omega(2z^{\bar{p}})^{1/q} + z^{1-\bar{p}}$.

Proof. Choose $0 < \rho \le \rho_S$. We will reduce ρ as the proof proceeds. From Lemma 6.4 we deduce

$$\|u_k^s - \bar{u}\|_s \le C_S \|u_k - \bar{u}\|_q. \tag{47}$$

By choosing $\rho > 0$ appropriately we can apply Theorem 5.10 with ρ replaced by $C_S\rho$ and $u^c = u_k^s$. We obtain that for all $u_k \in \mathcal{B}^\circ$, $\|u_k - \bar{u}\|_q < \rho$,

$$\|u_{k+1}^n - \bar{u}\|_q \le C\Phi_{\bar{p}}(\|u_k^s - \bar{u}\|_s)\|u_k^s - \bar{u}\|_s \le CC_S\Phi_{\bar{p}}(C_S\|u_k - \bar{u}\|_q)\|u_k - \bar{u}\|_q, \tag{48}$$

where $\Phi_{\bar{p}}$ is given by (30). Lemma 6.1 yields

$$\begin{aligned} \|u_{k+1} - \bar{u}\|_q = \|P[u_k^s](u_{k+1}^n) - \bar{u}\|_q &\le C_P\|u_{k+1}^n - \bar{u}\|_q + C_P'\|u_k^s - \bar{u}\|_s^2 \\ &\le C_S\ (CC_P\Phi_{\bar{p}}(C_S\|u_k - \bar{u}\|_q) + C_SC_P'\|u_k - \bar{u}\|_q)\ \|u_k - \bar{u}\|_q. \end{aligned} \tag{49}$$

This proves (45), since the $\Phi_{\bar{p}}$-term is of lowest order. By the properties of ω (see Lemma 5.7), $\Phi_{\bar{p}}(z)$ tends to zero as $z \to 0$. Hence, possibly after a further reduction of ρ, the algorithm is well-defined, since $u_0 \in \mathcal{B}^\circ$, $\|u_0 - \bar{u}\|_q < \rho$ implies $u_k \in \mathcal{B}^\circ$,

$\|u_k - \bar{u}\|_q < \rho$ for all k. Now (46) is obtained by combining (47) with k replaced by $k+1$, and the first inequalities in (48) and (49):

$$\begin{aligned}\|u_{k+1}^s - \bar{u}\|_s &\le C_S \|u_{k+1} - \bar{u}\|_q \\ &\le C_S (C_P \|u_{k+1}^n - \bar{u}\|_q + C_P' \|u_k^s - \bar{u}\|_s^2) \\ &\le C_S \left(C C_P \Phi_{\bar{p}}(\|u_k^s - \bar{u}\|_s) + C_P' \|u_k^s - \bar{u}\|_s\right) \|u_k^s - \bar{u}\|_s. \qquad \square\end{aligned}$$

If, in addition, (CS) holds, we get convergence with Q-order > 1 as follows.

COROLLARY 7.2. *In addition to the assumptions of Theorem* 7.1 *let condition* (CS) *hold at* $\bar{u}$. *Then with the choice* $\bar{p} = \min\{r/(r+\bar{q}), \tilde{q}/(\tilde{q}+\bar{q})\}$ *Theorem* 7.1 *yields*

$$\begin{aligned}\|u_{k+1} - \bar{u}\|_q &\le \bar{C}_1 \Phi_{CS}(C_S \|u_k - \bar{u}\|_q) \|u_k - \bar{u}\|_q, \\ \|u_{k+1}^s - \bar{u}\|_s &\le \bar{C}_2 \Phi_{CS}(\|u_k^s - \bar{u}\|_s) \|u_k^s - \bar{u}\|_s\end{aligned}$$

with $\bar{C}_1, \bar{C}_2 > 0$ *as in Theorem* 7.1, $\tilde{q} = \frac{qr}{r-q}$, *and*

$$\Phi_{CS}(z) = z^{\frac{\bar{q}}{\bar{q}+\max\{1,\tilde{q}/r\}\bar{q}}} + \left(\frac{z}{\nu}\right)^{\frac{s-q}{q}}.$$

In the case $r = s = \infty$ *the function* Φ_{CS} *assumes the simple form* $\Phi_{CS}(z) = z^{\frac{\bar{q}}{\bar{q}+\bar{q}}}$.

Proof. This follows immediately from Theorem 7.1 and Corollary 5.12. $\square$

8. Application to a class of regularized problems. In this section we apply our convergence theory to the following class of regularized problems which contains the one considered in the analysis of projected Newton methods by Kelley and Sachs [17]: We investigate problem (P) with the L^2-regularized objective function

$$f : u \in \mathcal{D} \subset L^p \longmapsto k(u) + \frac{1}{2}\|\sqrt{\alpha}(u - u^0)\|_2^2,$$

where $\alpha, u^0 \in L^\infty$ and $k : \mathcal{D} \longmapsto \mathbb{R}$ such that (A1) holds. The gradient is given by

$$g(u) = \alpha u - \alpha u^0 + \nabla k(u) \stackrel{\text{def}}{=} \alpha u + K(u).$$

We make the following assumption.

ASSUMPTION.

(A2′) *$g(u) = \alpha u + K(u)$ with $\alpha \in L^\infty$, $\alpha(x) \ge \alpha_0 > 0$ for a.a. $x \in \Omega$. Furthermore, there are $2 \le q < s \le \infty$ such that $g : \mathcal{B} \subset L^s \longrightarrow L^q$ is Lipschitz continuously Fréchet differentiable and K has the following smoothing property:*

$$K : \mathcal{B} \subset L^q \longrightarrow L^s$$

is Lipschitz continuous with Lipschitz constant L_K.

Obviously, (A2′) implies the Lipschitz continuity of $g : \mathcal{B} \subset L^s \longrightarrow L^s$ with Lipschitz constant $L_g = \|\alpha\|_\infty + m_{q,s} L_K$. Hence, (A2′) implies (A2) with $r = s$.

To perform a smoothing step we use a technique proposed in [17]. The following fixed point formulation of the optimality conditions (O1), (O2) is essential.

LEMMA 8.1. *Let* (A1) *hold. Then* (O1), (O2) *are satisfied at* $\bar{u}$ *if and only if*

$$\bar{u} = P(\bar{u} - \sigma g(\bar{u})), \tag{50}$$

where $\sigma \in L^\infty$, $\sigma > 0$ *a.e., is arbitrary.*

If in addition (A2′) *holds, then* $\bar{u}$ *satisfies* (O1), (O2) *if and only if*

$$\bar{u} = P(-\alpha^{-1}K(\bar{u})) \ (= P(\bar{u} - \alpha^{-1}g(\bar{u}))).$$

Furthermore, for all $u, v \in \mathcal{B}$ *the following holds true:*

$$\|P(-\alpha^{-1}K(u)) - P(-\alpha^{-1}K(v))\|_s \le \|\alpha^{-1}(K(u) - K(v))\|_s \le \frac{L_K}{\alpha_0}\|u - v\|_q, \tag{51}$$

i.e., the step $u_k \in \mathcal{B} \longmapsto S_k(u_k) \stackrel{\text{def}}{=} P(-\alpha^{-1}K(u_k))$ *has the smoothing property* (S) *for all* $u_k \in \mathcal{B}$.

Proof. Let $\bar{u} \in \mathcal{D}$ be arbitrary. Then (50) is satisfied if and only if (O1) holds (since the right-hand side of (50) is in $\mathcal{B}$) and

$$\sigma(x)g(\bar{u})(x) \begin{cases} \ge 0 & \text{if } \bar{u}(x) = a(x), \\ \le 0 & \text{if } \bar{u}(x) = b(x), \\ = 0 & \text{else} \end{cases} \qquad \text{a.e. on } \Omega.$$

Since $\sigma(x) > 0$ for a.a. $x \in \Omega$, this is nothing else but (O2). If in addition (A2′) holds, then $\bar{u} - \sigma g(\bar{u}) = (1 - \sigma\alpha)\bar{u} - \sigma K(\bar{u})$ and the choice $\sigma = \alpha^{-1}$ establishes the second assertion. (51) is easily obtained by using the smoothing property in (A2′) and the fact that $\|P(v) - P(w)\|_s \le \|v - w\|_s$ for all $v, w \in L^s$, since P is a pointwise projection. ∎

Remark 8.2. The smoothing step is a scaled projected gradient step obtained by making a scaled gradient step $-\alpha^{-1}g(u_k)$ and projecting the result pointwise onto $\mathcal{B}$. Moreover, $P(-\alpha^{-1}K(u_k)) - u_k$ is a descent direction for f at $u_k \in \mathcal{B}$ (cf. [13]): Since P is also the projection onto $\mathcal{B}$ in the scaled Hilbert space $(L^2, (\alpha \cdot \dots, \dots)_2)$ we get by using well-known properties of projections on closed convex sets in Hilbert space

$$0 \ge \big(\alpha(u_k - P(u_k - \alpha^{-1}g(u_k))), u_k - \alpha^{-1}g(u_k) - P(u_k - \alpha^{-1}g(u_k))\big)_2$$

and hence (note that we use the L^2 inner product as dual pairing)

$$\big\langle P(u_k - \alpha^{-1}g(u_k)) - u_k, g(u_k)\big\rangle \le -\big\|\sqrt{\alpha}(P(u_k - \alpha^{-1}g(u_k)) - u_k)\big\|_2^2. \qquad ∎$$

The preceding lemma shows that the convergence results of the previous section hold for the considered class of regularized problems if $u_k \longmapsto S_k(u_k) = P(-\alpha^{-1}K(u_k))$ is used as smoothing step.

We have already mentioned that (CS) is weaker than the corresponding assumption in [17] for the analysis of the projected Newton method. To allow a further comparison with the results in [17] we will show that (A3) is implied by Assumptions 2.1 and 2.3 in [17] which are stronger than (A2′) and the requirement that

$$\tilde{H}(u) \stackrel{\text{def}}{=} I + \alpha^{-1}\chi_{\bar{I}}K'(u)\chi_{\bar{I}}, \quad \bar{I} = \Omega \setminus \bar{A}, \tag{52}$$

has an inverse for all $u \in \mathcal{B}$, $\|u - \bar{u}\|_s < \tilde{\rho}$ with $\|\tilde{H}(u)^{-1}\|_{q,q} \le C_{\tilde{H}}$ (in [17] only $s = \infty$ is considered).

We use the following analogue of Assumption 2.2 in [17] which implies the local Lipschitz continuity of $K : \mathcal{B} \subset L^q \longrightarrow L^s$ in $\bar{u}$.

Assumption.

(A4) *There is* $\rho_K > 0$ *such that for all* $u \in \mathcal{B}$ *with* $\|u - \bar{u}\|_s < \rho_K$

$$\|K'(u)\|_{q,s} \le C_{K'}.$$

Here $K'(u) \in \mathcal{L}(L^p, L^{p'})$ *denotes the Fréchet derivative of* K *at* u.

The following lemma shows that (A3) is satisfied if (A3) holds for $\tilde{H}(u)$ defined in (52) instead of $H(u)$. Hence, (A3) is implied by Assumptions 2.2 and 2.3 in [17] for the choice $s = \infty$.

LEMMA 8.3. *Let* (O1), (O2), *and* (C) *hold at* $\bar{u}$, $\bar{A}$ *denote the active set, and* $\bar{I} = \bar{A}^c$. *If Assumptions* (A1), (A2′), *and* (A4) *are satisfied, then the following is true: If there is* $\bar{\rho} > 0$ *such that for all* $u \in \mathcal{B}$, $\|u - \bar{u}\|_s < \bar{\rho}$,

$$\bar{H}(u) \stackrel{\text{def}}{=} I + \alpha^{-1}\chi_{\bar{I}}K'(u) : L^q \longrightarrow L^q$$

is invertible with $\|\bar{H}(u)^{-1}\|_{q,q} \le C_{\bar{H}}$, *then* (A3) *holds for* $\rho_H > 0$ *sufficiently small.*

The lemma remains true if $\bar{H}(u)$ *is replaced by* $\tilde{H}(u)$ *defined in* (52) *since the uniformly bounded invertibility of* $\tilde{H}(u)$ *implies that of* $\bar{H}(u)$:

$$\bar{H}(u)^{-1} = \tilde{H}(u)^{-1} \cdot \left(I - \alpha^{-1}\chi_{\bar{I}}K'(u)\chi_{\bar{A}}I\right). \tag{53}$$

Proof. We note that $H(u)$ in Assumption (A3) can be equivalently replaced by

$$\hat{H}(u) \stackrel{\text{def}}{=} \frac{\chi_{\{d(u)<c\}}|g(u)|}{\chi_{\{d(u)<c\}}|g(u)| + \alpha d(u)}I + \frac{d(u)}{\chi_{\{d(u)<c\}}|g(u)| + \alpha d(u)}\nabla^2 f(u)$$

if (A1) and (A2′) are satisfied. This follows from the identity

$$\hat{H}(u) = \frac{|g(u)| + d(u)}{\chi_{\{d(u)<c\}}|g(u)| + \alpha d(u)}H(u)$$

and the fact that the first factor is continuously invertible, since

$$\frac{|g(u)| + d(u)}{\chi_{\{d(u)<c\}}|g(u)| + \alpha d(u)} \in \begin{cases} [\min\{\|\alpha\|_\infty^{-1}, 1\}, \max\{\alpha_0^{-1}, 1\}] & \text{on } \{d(u) < c\}, \\ [\|\alpha\|_\infty^{-1}, \alpha_0^{-1}(1 + \nu^{-1}C_g)] & \text{on } \{d(u) \ge c\}. \end{cases}$$

In particular, there exists a constant $C_{\hat{H}H}$ with

$$\|H(u)^{-1}\|_{q,q} \le C_{\hat{H}H}\|\hat{H}(u)^{-1}\|_{q,q}.$$

According to a standard result of operator theory we can establish (A3) with $C_H = 2C_{\hat{H}H}C_{\bar{H}}$ by finding $\rho_H > 0$ such that $\|\bar{H}(u) - \hat{H}(u)\|_{q,q} \le 1/(2C_{\bar{H}})$ for all $u \in \mathcal{B}^\circ$, $\|u - \bar{u}\|_s < \rho_H$. To this end, let $\rho \le \min\{1, \bar{\rho}, \rho_K\}$ and $u \in \mathcal{B}^\circ$, $\|u - \bar{u}\|_s < \rho$, be arbitrary. We will adjust ρ as the proof proceeds. We observe that

$$\bar{H}(u) - \hat{H}(u) = \left(\frac{\chi_{\bar{I}}}{\alpha} - \frac{d(u)}{\chi_{\{d(u)<c\}}|g(u)| + \alpha d(u)}\right)K'(u), \tag{54}$$

apply Lemma 2.3 with $q_0 = q$, $q_1 = s/q$, $q_1' = s/(s-q)$, use (A4), and obtain

$$\begin{aligned} \|\bar{H}(u) - \hat{H}(u)\|_{q,q} &\le \left\|\frac{\chi_{\bar{I}}}{\alpha} - \frac{d(u)}{\chi_{\{d(u)<c\}}|g(u)| + \alpha d(u)}\right\|_{\frac{qs}{s-q}} \|K'(u)\|_{q,s} \\ &\le \left\|\frac{\chi_{\bar{I}}}{\alpha} - \frac{d(u)}{\chi_{\{d(u)<c\}}|g(u)| + \alpha d(u)}\right\|_{\frac{qs}{s-q}} C_{K'}. \end{aligned}$$

We notice that $q \le \hat{q} \stackrel{\text{def}}{=} qs/(s-q) < \infty$ ($\hat{q} = q$ if $s = \infty$) and split Ω to estimate the first factor in the last expression. For

$$B(u) \stackrel{\text{def}}{=} \left\{x \in \Omega : \chi_{\{d(u)<c\}}(x)|g(u)(x)| + \alpha(x)d(u)(x) \le \sqrt{\rho}\right\}$$

and $\rho < \nu^2\alpha_0^2$ we have $B(u) \subset \{d(u) < c\}$, and thus with $\alpha_1 = \min\{\alpha_0, 1\}$

$$B(u) \subset \{x \in \Omega : \alpha_1|g(u)(x)| + \alpha_1 d(u)(x) \le \sqrt{\rho}\} = N_{\sqrt{\rho}/\alpha_1}(u).$$

Denote the complement of $B(u)$ by $B^c(u)$. The key observation is that the parenthesis in (54) is small on $B^c(u)$ and the measure of the residual set $B(u)$ is small as well. We get by Lemma 5.7 and Minkowski's inequality

$$\begin{aligned}\left\|\frac{\chi_{\bar I}}{\alpha} - \frac{d(u)}{\chi_{\{d(u)<c\}}|g(u)| + \alpha d(u)}\right\|_{\hat q, B(u)} &\le \frac{1}{\alpha_0}\mu(B(u))^{1/\hat q} \le \frac{1}{\alpha_0}\mu(N_{\sqrt{\rho}/\alpha_1}(u))^{1/\hat q}\\ &\le \frac{1}{\alpha_0}\left(\omega\left(2\frac{\sqrt{\rho}}{\alpha_1}\right) + ((L_g + L_d)\alpha_1\sqrt{\rho})^s\right)^{1/\hat q}\\ &\le \frac{1}{\alpha_0}\left(\omega\left(2\frac{\sqrt{\rho}}{\alpha_1}\right)^{1/\hat q} + ((L_g + L_d)\alpha_1\sqrt{\rho})^{s/\hat q}\right).\end{aligned}$$

Moreover, we obtain as in the proof of Theorem 5.10

$$\left\|\frac{\chi_{\bar I}}{\alpha} - \frac{d(u)}{\chi_{\{d(u)<c\}}|g(u)| + \alpha d(u)}\right\|_{\hat q, B^c(u)} \le C_1 \left\|\frac{\chi_{\bar I}}{\alpha} - \frac{d(u)}{\chi_{\{d(u)<c\}}|g(u)| + \alpha d(u)}\right\|_{s, B^c(u)}^{\min\{1, s/\hat q\}}$$

by applying Lemma 2.2 in the case $s < \hat q$ (note that the function under the norm is nonnegative and pointwise bounded by $1/\alpha_0$). We can choose $C_1 = m_{\hat q, s}$ if $\hat q \le s$ and $C_1 = \alpha_0^{-1+s/\hat q}$ if $s < \hat q$. Since $g(\bar u) = 0$ a.e. on $\bar I$ by (O2) we get with (A2)

$$\begin{aligned}\left\|\frac{\chi_{\bar I}}{\alpha} - \frac{d(u)}{\chi_{\{d(u)<c\}}|g(u)| + \alpha d(u)}\right\|_{s, B^c(u)\cap\bar I} &= \left\|\frac{\chi_{\{d(u)<c\}}|g(u) - g(\bar u)|}{\alpha\left(\chi_{\{d(u)<c\}}|g(u)| + \alpha d(u)\right)}\right\|_{s, B^c(u)\cap\bar I}\\ &\le \left\|\frac{g(u) - g(\bar u)}{\alpha_0\sqrt{\rho}}\right\|_s \le \frac{L_g}{\alpha_0}\sqrt{\rho}.\end{aligned}$$

The fact that $d(\bar u) = 0$ a.e. on $\bar A$ yields together with Lemma 5.3

$$\begin{aligned}\left\|\frac{\chi_{\bar I}}{\alpha} - \frac{d(u)}{\chi_{\{d(u)<c\}}|g(u)| + \alpha d(u)}\right\|_{s, B^c(u)\setminus\bar I} &= \left\|\frac{d(u) - d(\bar u)}{\chi_{\{d(u)<c\}}|g(u)| + \alpha d(u)}\right\|_{s, B^c(u)\setminus\bar I}\\ &\le \left\|\frac{d(u) - d(\bar u)}{\sqrt{\rho}}\right\|_s \le L_d\sqrt{\rho}.\end{aligned}$$

Hence, there are constants $C_1, C_2 > 0$ such that

$$\|\bar H(u) - \hat H(u)\|_{q,q} \le \left(C_1\omega\left(2\frac{\sqrt{\rho}}{\alpha_1}\right)^{1/\hat q} + C_2\sqrt{\rho}^{\min\{1, \frac{s}{\hat q}\}}\right) C_{K'}.$$

Due to Lemma 5.7, after a possible reduction of $\rho > 0$ the right-hand side is $\le 1/(2C_{\bar H})$ and the choice $\rho_H = \rho$ completes the first part of the proof.

Now assume that the assumptions hold for $\tilde H(u)$ instead of $\bar H(u)$. We only have to verify the explicit formula (53) for $\bar H(u)^{-1}$. For $v \in L^q$ we look at the equation

$$v = \bar H(u)h = \tilde H(u)h + \alpha^{-1}\chi_{\bar I}K'(u)\chi_{\bar A}h. \tag{55}$$

Premultiplication by $\chi_{\bar{A}}$ shows $h_{\bar{A}} = v_{\bar{A}}$, and hence

$$h = \tilde{H}(u)^{-1}\left(v - \alpha^{-1}\chi_{\bar{I}}K'(u)v_{\bar{A}}\right).$$

Therefore, the operator given in (53) is a left inverse of $\bar{H}(u)$. It is also a right inverse; to see this we note that $\chi_{\bar{A}}\tilde{H}(u) = \chi_{\bar{A}}I$, hence $\chi_{\bar{A}}\tilde{H}(u)^{-1} = \chi_{\bar{A}}I$, and, consequently,

$$(\tilde{H}(u) + \alpha^{-1}\chi_{\bar{I}}K'(u)\chi_{\bar{A}}I)\tilde{H}(u)^{-1}(I - \alpha^{-1}\chi_{\bar{I}}K'(u)\chi_{\bar{A}}I) = I,$$

where we have used the fact that $\chi_{\bar{A}}\alpha^{-1}\chi_{\bar{I}} = 0$. □

Remark 8.4. The results of Lemma 8.3 remain true if α also depends on u. □

9. Second-order sufficient conditions. We will now study how Algorithm 5.16 behaves in the neighborhood of a point $\bar{u}$ satisfying the second-order sufficient condition given by Dunn and Tian in [9]. We will show that it implies (A3) in the case $q = 2$ under the additional assumptions of the previous section and also for $q > 2$ if the range of $\tilde{H}(u)$ is dense in L^q. In section 10 we will use this sufficiency condition to show that the developed affine-scaling Newton method produces acceptable steps for the trust-region globalization considered in [26] if the iterates u_k are close enough to $\bar{u}$. In our notation, the second-order sufficiency conditions by Dunn and Tian [9] read as follows.

ASSUMPTION (second-order sufficient conditions by Dunn and Tian).
(OS) *Condition* (A1) *holds and there are* $t \in [1,\infty]$, $c_r > 0$ *such that*

$$|\langle v, \nabla^2 f(u)w\rangle| \le c_r\|v\|_2\|w\|_2 \quad \textit{for all } u \in \mathcal{B}, v, w \in L^\infty, \tag{56}$$

$$\lim_{\substack{u\in\mathcal{B}\\ \|u-\bar{u}\|_t\to 0}} \sup_{\substack{w\in L^\infty\\ \|w\|_2=1}} \langle w, (\nabla^2 f(u) - \nabla^2 f(\bar{u}))w\rangle = 0. \tag{57}$$

Moreover, (O1), (O2) *are satisfied at* $\bar{u}$ *and there are sets* $A \subset \bar{A}$, $I = A^c$ *and constants* $c_1, c_2 > 0$ *with*

$$g(\bar{u}) \ge c_1 \quad \textit{on } A, \quad \langle \chi_I w, \nabla^2 f(\bar{u})\chi_I w\rangle \ge c_2\|\chi_I w\|_2^2 \quad \textit{for all } w \in L^\infty. \tag{58}$$

Remark 9.1. Dunn and Tian [9] deduce (58) from the second-order necessary conditions (see also [26]), if in addition a pointwise strict complementarity condition similar to (C), the standard L^2 coercivity analogue of the necessary conditions (i.e., the second condition in (58) only for $\bar{I} = \bar{A}^c$ instead of I), and certain structure/continuity conditions hold that are typically satisfied by Bolza ODE optimal control problems. In such cases, the gap between the sufficient condition (OS) and the necessary conditions is therefore narrower than it might appear. Extensions of these results to the case of general pointwise affine constraints are established in [10], [11]. In particular, the application of the theory in [10] to nonnegativity constraints improves the sufficiency result in [9]. □

Remark 9.2. Condition (OS) is weaker (stronger) than the sufficient second-order condition of Maurer in [23] if $\|\cdot\|_2$ (respectively, $\|\cdot\|_1$) is chosen as the weak norm. Since we prefer a result of the form $f(u) - f(\bar{u}) \ge C\|u - \bar{u}\|_l^2$ for $l = 2$ rather than for $l = 1$, condition (OS) better meets our requirements. Moreover, it is obvious that in view of Lemma 2.2 the requirement $t \in [1,\infty)$ could be equivalently replaced by $t \in \{2,\infty\}$ since the relative topology of L^t on $\mathcal{B}$ is the same for all $t \in [1,\infty)$. □

9.1. L^∞-optimality. The following theorem shows that (OS) implies the L^∞-optimality of $\bar u$ for (P) (cf. [9]).

THEOREM 9.3. *Let the formal second-order sufficiency condition* (OS) *hold. Then $\bar u$ is a strict L^∞-optimizer for* (P)*, more precisely: There are $\rho > 0$ and $C > 0$ such that*

$$u \in \mathcal{B},\ \|u-\bar u\|_\infty < \rho \Longrightarrow f(u) - f(\bar u) \geq C\|u-\bar u\|_2^2.$$

Proof. The proof is a variant of the one given for Lemma 1 in [9]. Let $u \in \mathcal{B}$. With $v = u - \bar u$ we get from (OS)

$$\begin{aligned} f(u) - f(\bar u) &= \langle v, g(\bar u)\rangle + \frac{1}{2}\left(\langle v_I, \nabla^2 f(u) v_I\rangle + \langle v_A, \nabla^2 f(u)(v_A + 2v_I)\rangle\right) + o(\|v\|_2^2) \\ &\geq c_1\|v_A\|_1 + \frac{c_2}{2}\|v_I\|_2^2 - \frac{c_r}{2}\left(\|v_A\|_2^2 + 2\|v_A\|_2\|v_I\|_2\right) + o(\|v\|_2^2) \\ &\geq c_1\frac{\|v_A\|_2^2}{\|v_A\|_\infty} + \frac{c_2}{2}\|v_I\|_2^2 - \frac{c_r}{2}\left(1 + \frac{2c_r}{c_2}\right)\|v_A\|_2^2 - \frac{c_2}{4}\|v_I\|_2^2 + o(\|v\|_2^2) \\ &\geq \left(\frac{c_1}{\|v_A\|_\infty} - \frac{c_r(2c_r + c_2)}{2c_2}\right)\|v_A\|_2^2 + \frac{c_2}{4}\|v_I\|_2^2 + o(\|v\|_2^2), \end{aligned}$$

where we have used $2\alpha\beta \leq c\alpha^2 + \beta^2/c$ with $c = c_2/(2c_r)$. Note that $o(\|v\|_2^2)$ is meant for $\|v\|_\infty \to 0$. We see that the assertion follows for all $u \in \mathcal{B}$ with $\|u-\bar u\|_\infty < \rho$ if $\rho > 0$ is small enough. ∎

9.2. L^2-optimality. We make now additional assumptions on the structure of the second derivative which are met by the class of regularized problems considered in the previous section and are similar to those in [25].

ASSUMPTION.

(A5) $\nabla^2 f(u) = \beta(u)I + K'(u)$, *where $\beta : \mathcal{B} \subset L^2 \longrightarrow L^\infty$ is continuous and K' satisfies* (A4) *for suitable $2 \leq q < s \leq \infty$.*

We have the following variant of Theorem 4 in [9].

THEOREM 9.4. *Let the formal second-order sufficiency condition* (OS) *with $t < \infty$ (i.e., also for $t = 2$) and* (A5) *hold. If, in addition, $\beta(\bar u)(x) \geq \beta_0 > 0$ a.e. on Ω, then $\bar u$ is a strict L^2-optimizer (and hence L^t-optimizer, $t \in [1,\infty]$) for* (P) *in the following sense: There are $\rho > 0$ and $C > 0$ such that*

$$u \in \mathcal{B},\ \|u-\bar u\|_2 < \rho \Longrightarrow f(u) - f(\bar u) \geq C\|u-\bar u\|_2^2.$$

Proof. We compute as in the proof of Theorem 9.3

$$\begin{aligned} f(u) - f(\bar u) &\geq c_1\|v_A\|_1 + \frac{c_2}{2}\|v_I\|_2^2 + \frac{\beta_0}{2}\|v_A\|_2^2 + \frac{1}{2}\langle v_A, K'(u)(v_A + 2v_I)\rangle + o(\|v\|_2^2) \\ &\geq c_1\|v_A\|_1 + \frac{c_2}{2}\|v_I\|_2^2 + \frac{\beta_0}{2}\|v_A\|_2^2 - \|K'(u)\|_{q,s}\|v_A\|_{s'}\|v\|_q + o(\|v\|_2^2), \end{aligned}$$

where $v = u - \bar u$ and $1/s + 1/s' = 1$. Then $1 \leq s' < 2 \leq q < s$ and $1/q + 1/s' \stackrel{\text{def}}{=} 1 + \delta > 1$ because of $s > q$. Now by Lemma 2.2

$$\|v_A\|_{s'} \leq \|v_A\|_1^{1/s'}\|v_A\|_\infty^{1-1/s'}, \quad \|v\|_q \leq \|v\|_2^{2/q}\|v\|_\infty^{1-2/q}.$$

Since $\|v_A\|_\infty \leq \|v\|_\infty \leq \|b-a\|_\infty$, we find $C_1 > 0$ with

$$\|v_A\|_{s'}\|v\|_q \leq C_1\|v_A\|_1^{1/s'}\|v\|_2^{2/q} \leq C_1\left(\frac{1}{p_1'}\|v_A\|_1^{p_1'/s'} + \frac{1}{p_1}\|v\|_2^{2p_1/q}\right)$$

for all $p_1, p_1' \in (1,\infty)$, $1/p_1 + 1/p_1' = 1$, according to Young's inequality. We choose $p_1 = q(1+\varepsilon)$ and get

$$\frac{1}{p_1'} = 1 - \frac{1}{q(1+\varepsilon)} = \frac{1}{s'} + \frac{1}{q} - \frac{1}{q(1+\varepsilon)} - \delta < \frac{1}{s'(1+\varepsilon')}$$

for $\varepsilon, \varepsilon' > 0$ small enough. Hence, for $\|v\|_2$ small

$$\|K'(u)\|_{q,s}\|v_A\|_{s'}\|v\|_q \le C_K C_1(\|v_A\|_1^{1+\varepsilon'} + \|v_A\|_2^{2+2\varepsilon}),$$

which completes the proof. □

We shall now study in which cases condition (A3) is implied by the formal second-order sufficiency condition (OS). We will thereby restrict ourselves to problems which satisfy the structural assumptions of section 8.

THEOREM 9.5. *Let* (OS), (A2′) *with* $t < \infty$, *and* (A4) *hold. Then the following is true:*

(1) *If* $q = 2$, *then* (A3) *is satisfied.*

(2) *For* $q > 2$ *there are* $\rho > 0$ *and* $C_{\tilde H} > 0$ *such that for all* $u \in \mathcal{B}^\circ$, $\|u - \bar u\|_s < \rho$, *the operator* $\tilde H(u)$ *in* (52) *has the properties*

(i) $\tilde H(u) \in \mathcal{L}(L^q, L^q)$ *and* $\|\tilde H(u)v\|_q \ge C_{\tilde H}\|v\|_q$ *for all* $v \in L^q$;

(ii) *the range of* $\tilde H(u) : L^q \longrightarrow L^q$ *is closed in* L^q.

Hence, if $K' : \mathcal{B} \subset L^s \longrightarrow \mathcal{L}(L^q, L^q)$ *is continuous at* $\bar u$ *and if the range of* $\tilde H(\bar u)$ *is dense in* L^q, *then* (A3) *is satisfied.*

Proof. Let $0 < \rho \le \rho_K$ and $u \in \mathcal{B}$, $\|u - \bar u\|_s < \rho$. We will adjust ρ in what follows. By (A4) and the definition of $\tilde H(u)$ we have $\tilde H(u) \in \mathcal{L}(L^q, L^q)$ and

$$\langle \alpha v, \tilde H(u)w\rangle \le (m_{2,q}^2\|\alpha\|_\infty + m_{q',s}C_{K'})\|v\|_q\|w\|_q \quad \text{for all } v, w \in L^q. \tag{59}$$

From $t < \infty$, Lemmas 2.1 and 2.2 we deduce that $u \in \mathcal{B}$, $\|u - \bar u\|_s \to 0$ implies $\|u - \bar u\|_t \to 0$. Using $\alpha(\tilde H(u) - \tilde H(\bar u)) = \chi_{\bar I}(\nabla^2 f(u) - \nabla^2 f(\bar u))\chi_{\bar I}$ we obtain from (A2′) and (57) by a density argument in L^q

$$\lim_{\substack{u\in\mathcal{B}\\ \|u-\bar u\|_s\to 0}} \sup_{\substack{w\in L^q\\ \|w\|_2=1}} \langle w, (\tilde H(u) - \tilde H(\bar u))w\rangle = 0. \tag{60}$$

For arbitrary $w \in L^q$ we have with (OS)

$$\begin{aligned}
\langle \alpha w, \tilde H(\bar u)w\rangle &= \langle \alpha w, w\rangle + \langle \alpha w_{\bar I}, \alpha^{-1}K'(\bar u)w_{\bar I}\rangle\\
&= \langle \alpha w_{\bar A}, w_{\bar A}\rangle + \langle \alpha w_{\bar I}, w_{\bar I}\rangle + \langle w_{\bar I}, K'(\bar u)w_{\bar I}\rangle\\
&= \langle \alpha w_{\bar A}, w_{\bar A}\rangle + \langle w_{\bar I}, \nabla^2 f(\bar u)w_{\bar I}\rangle\\
&\ge \alpha_0\|w_{\bar A}\|_2^2 + c_2\|w_{\bar I}\|_2^2 \ge \min\{\alpha_0, c_2\}\,\|w\|_2^2 \stackrel{\text{def}}{=} C_1\|w\|_2^2.
\end{aligned}$$

Together with (60) this shows that for sufficiently small $\rho > 0$ we have

$$\langle \alpha w, \tilde H(u)w\rangle \ge \frac{C_1}{2}\|w\|_2^2 \quad \text{for all } w \in L^q. \tag{61}$$

Hence, in the case $q = 2$ the symmetric operator $\alpha\tilde H(u) \in \mathcal{L}(L^2, L^2)$ is bounded by (59) and positive by (61). We therefore may apply the Lax–Milgram theorem (which in our symmetric case is an immediate consequence of Riesz's representation theorem), yielding that $\tilde H(u)$ is continuously invertible in $\mathcal{L}(L^2, L^2)$ with $\|\tilde H(u)^{-1}\|_{2,2} \le 2\|\alpha\|_\infty/C_1$. By Lemma 8.3 this implies (A3) for sufficiently small $\rho_H > 0$.

Next we assume $q > 2$ and establish the second part of (2i). Let $w \in L^q$ be arbitrary. For $c > 0$ which will be adjusted later we consider two cases: If $\|w\|_2 \geq c\|w\|_q$, then by (61)

$$\frac{cC_1}{2}\|w\|_2\|w\|_q \leq \frac{C_1}{2}\|w\|_2^2 \leq \langle \alpha w, \tilde{H}(u)w\rangle \leq \|\alpha\|_\infty m_{q',2}\|w\|_2\|\tilde{H}(u)w\|_q,$$

from which

$$\|\tilde{H}(u)w\|_q \geq \frac{cC_1}{2\|\alpha\|_\infty m_{q',2}}\|w\|_q.$$

In the case $\|w\|_2 < c\|w\|_q$ we compute

$$\|\tilde{H}(u)w\|_q \geq \|w\|_q - \frac{1}{\alpha_0}\|\chi_{\bar{I}}K'(u)w_{\bar{I}}\|_q \geq \|w\|_q - \frac{1}{\alpha_0}\|K'(u)w_{\bar{I}}\|_q.$$

Applying Lemma 2.2 we obtain with (A4) and suitable $\theta \in (0,1)$

$$\|K'(u)w_{\bar{I}}\|_q \leq \|K'(u)w_{\bar{I}}\|_2^\theta \|K'(u)w_{\bar{I}}\|_s^{1-\theta} \leq \|K'(u)w_{\bar{I}}\|_2^\theta (C_{K'}\|w\|_q)^{1-\theta}.$$

Obviously, (59) also holds with the left side replaced by $|\langle v, \nabla^2 f(u)w\rangle|$. Therefore, (56) implies by a density argument together with (A2′) and (A4) that

$$\|K'(u)v\|_2 = \|\nabla^2 f(u)v - \alpha v\|_2 \leq (c_r + \|\alpha\|_\infty)\|v\|_2 \stackrel{\text{def}}{=} C_r\|v\|_2 \quad \text{for all } v \in L^q.$$

This gives

$$\|K'(u)w_{\bar{I}}\|_q \leq (C_r\|w\|_2)^\theta (C_{K'}\|w\|_q)^{1-\theta} \leq (C_r c)^\theta C_{K'}^{1-\theta}\|w\|_q,$$

and choosing $c > 0$ small enough we achieve

$$\|\tilde{H}(u)w\|_q \geq \frac{1}{2}\|w\|_q,$$

as long as $\|w\|_2 < c\|w\|_q$. Since $c > 0$ can be adjusted independently of w, (i) is shown.

To prove (ii), let $(w_k) \subset L^q$ be arbitrary. Then

$$\begin{aligned}
&\tilde{H}(u)w_k \stackrel{L^q}{\to} v \in L^q \quad (k \to \infty)\\
&\stackrel{\text{(i)}}{\Longrightarrow} \|w_k - w_l\|_q \leq C_{\tilde{H}}^{-1}\|\tilde{H}(u)w_k - \tilde{H}(u)w_l\|_q \to 0 \quad (k,l \to \infty)\\
&\Longrightarrow w_k \stackrel{L^q}{\to} w \in L^q \quad (k \to \infty) \stackrel{\text{(i)}}{\Longrightarrow} \tilde{H}(u)w_k \stackrel{L^q}{\to} \tilde{H}(u)w \quad (k \to \infty).
\end{aligned}$$

If the range of $\tilde{H}(\bar{u}) : L^q \longrightarrow L^q$ is dense, then $\tilde{H}(\bar{u})$ is injective by (i) and surjective by (ii). Thus, it has a continuous inverse by the open mapping theorem and (i) shows $\|\tilde{H}(\bar{u})^{-1}\|_{q,q} \leq C_{\tilde{H}}^{-1}$. If, in addition, $K' : \mathcal{B} \subset L^s \longrightarrow \mathcal{L}(L^q, L^q)$ is continuous at $\bar{u}$, then for $\|u - \bar{u}\|_s$ small enough $\tilde{H}(u)^{-1} \in \mathcal{L}(L^q, L^q)$ exists and (A3) is satisfied for sufficiently small $\rho_H > 0$. ∎

The previous result shows that—at least in the case $q = 2$—the application of Algorithm 5.16 to the class of problems considered in section 8 leads to superlinear convergence in a neighborhood of a point $\bar{u}$ satisfying (OS). This is especially important since formal sufficiency conditions of type (OS) are the usual starting point for proving that a rapidly convergent local method meets the trial step requirements of a globally convergent algorithm in a neighborhood of a local optimizer. Hence, it is important that the local convergence theory can be established under a sufficiency condition that is as weak as possible.

10. Trust-region globalization. The aim of this section is to show that near a local optimizer satisfying (OS) Algorithm 5.16 produces admissible trial steps for the globally convergent affine-scaling interior-point trust-region algorithm that we proposed and analyzed in [26]. The trust-region globalization extends ideas of Coleman and Li in [7] and uses the fact that $u^n \in L^q$ solves (7) for given $u^c \in \mathcal{B}^\circ$ if and only if u^n is a stationary point of the quadratic function $\psi[u^c] : L^q \longrightarrow \mathbb{R}$,

$$\psi[u^c](u) \stackrel{\text{def}}{=} \langle u - u^c, g(u^c)\rangle + \frac{1}{2}\langle u - u^c, M(u^c)(u - u^c)\rangle, \tag{62}$$
$$\text{where} \quad M(u) \stackrel{\text{def}}{=} d(u)^{-1}G(u) = \chi_{\{d(u)<c\}}|g(u)|d(u)^{-1}I + \nabla^2 f(u).$$

Here and in the following we use the standard notation s for the trial steps although it collides with the norm index occurring in (A2). There is no danger of ambiguity. As shown in [26], a globally convergent algorithm can be obtained as follows: Denote by $u_k \in \mathcal{B}^\circ$ the current iterate. We compute a trial step s_k as an approximate solution to the trust-region subproblem

$$\text{minimize } \psi[u_k](u_k + s) \text{ subject to } u_k + s \in \mathcal{B},\ \|s\|_q \le \Delta_k. \tag{63}$$

This trial step is required to satisfy the following.

Fraction of Cauchy decrease condition.

(D) $u_k + s_k \in \mathcal{B}^\circ$, $\|s_k\|_q \le \beta_0\Delta_k$, *and* $\psi[u_k](u_k + s_k) \le \beta\psi[u_k]^c$, *where*

$$\psi[u_k]^c \stackrel{\text{def}}{=} \beta \min\{\psi[u_k](u_k + s)s = -\tau d_k^\vartheta g_k,\ \tau \ge 0,\ u_k + s \in \mathcal{B},\ \|s\|_q \le \Delta_k\}$$

with fixed constants $\beta_0 > 0$, $0 < \beta < 1$, $\vartheta \ge 1$. The step s_k is accepted, i.e., $u_{k+1} = u_k + s_k$, if $r_k > \eta_1$, where $0 < \eta_1 < 1$ is fixed and the decrease ratio $r_k = r(u_k, s_k)$ is given by

$$r(u_k, s_k) \stackrel{\text{def}}{=} \frac{f(u_k + s_k) - f(u_k)}{\langle s_k, g(u_k)\rangle + \langle s_k, \nabla^2 f(u_k)s_k\rangle/2}.$$

Otherwise, i.e., if $r_k \le \eta_1$, the step is rejected: $u_{k+1} = u_k$. For our presentation it is convenient to use an update rule for the trust-region radius Δ_k that is slightly different from the one given in [26]. However, it is not hard to verify that all the convergence results stated therein remain valid. In our update rule we fix $0 < \eta_1 < \eta_2 < \eta_3 < 1$, $0 < \beta_0\gamma_0 \le \gamma_1 < 1 < \gamma_2 \le \gamma_3$, $\Delta_{\min} > 0$, and choose

$$\Delta^+ \in \begin{cases} [\gamma_0\|s_k\|_q, \gamma_1\Delta_k] & \text{if } r_k \le \eta_1, \\ [\gamma_1\Delta_k, \Delta_k] & \text{if } \eta_1 < r_k < \eta_2, \\ [\Delta_k, \gamma_2\Delta_k] & \text{if } \eta_2 < r_k < \eta_3, \\ [\gamma_2\Delta_k, \gamma_3\Delta_k] & \text{else,} \end{cases} \qquad \Delta_{k+1} := \begin{cases} \Delta^+ & \text{if } r_k \le \eta_1, \\ \max\{\Delta_{\min}, \Delta^+\} & \text{else.} \end{cases}$$

For a detailed formulation of the algorithm and its convergence properties we refer to [26]. The theory developed therein (adapted to our update rule) states that under Assumption (A1) each accumulation point of the sequence (u_k) satisfies the first-order necessary optimality conditions (O1), (O2), and, moreover, the second-order necessary condition [26, Thm. 3.3, (O3)] if (D) is replaced by a fraction of optimal decrease condition.

Keeping in mind that trust-region methods for unconstrained problems inherit their local convergence behavior from Newton's method, it is natural to try to accelerate the above trust-region method by means of Algorithm 5.16. We combine both methods as follows.

ALGORITHM 10.1 (trust-region interior-point Newton method).

1. Choose $\Delta_k \geq \Delta_{\min}$, $u_0 \in \mathcal{B}^\circ$.
2. For $k = 0, 1, 2, \ldots$
   - 2.1. If $d(u_k)g(u_k) = 0$, STOP.
   - 2.2. Select and perform a smoothing step: $u_k^s = S_k^\circ(u_k)$.
   - 2.3. Compute a trial step by Algorithm 5.16: $s_k = \min\{1, \Delta_k/\|s_k^N\|_q\}s_k^N$, where

$$s_k^N = u_{k+1}^N - u_k^s,\ u_{k+1}^N = P[u_k^s](u_{k+1}^n),\ u_{k+1}^n = u_k^s - G(u_k^s)^{-1}d(u_k^s)g(u_k^s).$$

     If s_k satisfies (D) for $\psi[u_k^s]$ then goto step 2.5.
   - 2.4. Compute a trial step that satisfies (D) for $\psi[u_k^s]$, e.g., by a descent method that starts with a line search along $-d(u_k^s)^\vartheta g(u_k^s)$.
   - 2.5. Compute the decrease ratio $r_k = r(u_k^s, s_k)$ and the new trust-region radius Δ_{k+1}. If $r_k > \eta_1$ then set $u_{k+1} = u_k^s + s_k$. Otherwise set $u_{k+1} = u_k^s$ and go to step 2.2.

Now suppose that one of the accumulation points $\bar{u} \in \mathcal{B}$ of (u_k) satisfies the second-order sufficiency condition (OS). The question is: Does this globally convergent method eventually turn into Algorithm 5.16 and thus inherit its superlinear convergence?

It is beyond the scope of this paper to answer this question in full generality, since this would require us to analyze the effect of the smoothing step on the global convergence behavior of the trust-region algorithm. We try to find a reasonable compromise by developing results that are rigorously applicable whenever the smoothing steps do not affect the global convergence. This is certainly the case if the smoothing steps decrease the objective function f; see Remark 8.2 in this context. Moreover, we require the following.

ASSUMPTION.

(A6) *$\bar{u} \in \mathcal{B}$ is an accumulation point of (u_k) at which* (OS) *holds. Moreover, condition* (A2) *is satisfied with $r = s = \infty$.*

As a first result we show that the quadratic model $\psi[u_k^s]$ has a unique minimizer if $\|u_k^s - \bar{u}\|_\infty$ is sufficiently small. To show this, we first prove the following lemma.

LEMMA 10.2. *Let* (A6) *hold. Then there are $\rho > 0$, $C_M > 0$ such that*

$$\langle v, M(u)v\rangle \geq C_M\|v\|_2^2 \quad \textit{for all } v \in L^q$$

for all $u \in \mathcal{B}^\circ$, $\|u - \bar{u}\|_\infty < \rho$.

Proof. We know that (O1), (O2) hold at $\bar{u}$. Let I and A be defined as in (OS). Since $|g(\bar{u})| \geq c_1$ a.e. on A, we get by (O2) that $d(\bar{u}) = 0$ a.e. on A and hence for sufficiently small $\rho > 0$ and all $u \in \mathcal{B}^\circ$, $\|u - \bar{u}\|_\infty < \rho$

$$|g(u)| \geq c_1/2 \quad \text{and} \quad c > d(u) \leq L_d\rho \quad \text{a.e. on } A$$

by (A2) and Lemma 5.3, respectively. (A2) yields with a density argument in L^q that (56)–(58) hold also for L^∞ replaced by L^q and thus, possibly after reducing ρ, we have

$$\langle v_I, \nabla^2 f(u)v_I\rangle \geq \frac{c_2}{2}\|v_I\|_2^2 \ \text{ for all } v \in L^q$$

as long as $u \in \mathcal{B}^\circ$, $\|u - \bar{u}\|_\infty < \rho$. Hence, for all $v \in L^q$

$$\begin{aligned}\langle v, M(u)v\rangle &= \left\langle v_A, \frac{|g(u)|}{d(u)} v_A \right\rangle + \left\langle v_I, \frac{\chi_{\{d(u)<c\}}|g(u)|}{d(u)} v_I \right\rangle \\ &\quad + \langle v_I, \nabla^2 f(u) v_I\rangle + \langle v_A, \nabla^2 f(u)(2v_I + v_A)\rangle \\ &\geq \frac{c_1}{2L_d\rho}\|v_A\|_2^2 + \frac{c_2}{2}\|v_I\|_2^2 - c_r\|v_A\|_2^2 - 2c_r\|v_A\|_2\|v_I\|_2.\end{aligned}$$

With the standard estimate

$$2c_r\|v_A\|_2\|v_I\|_2 \leq \frac{c_2}{4}\|v_I\|_2^2 + \frac{4c_r^2}{c_2}\|v_A\|_2^2$$

we arrive at

$$\langle v, M(u)v\rangle \geq \left(\frac{c_1}{2L_d\rho} - c_r - \frac{4c_r^2}{c_2}\right)\|v_A\|_2^2 + \frac{c_2}{4}\|v_I\|_2^2.$$

Now for $\rho > 0$ sufficiently small the assertion follows. ∎

THEOREM 10.3. *Let* (A3) *and* (A6) *hold and* (u_k), (u_k^s), (u_k^n), (u_k^N) *be generated by Algorithm* 10.1. *If* ρ *is sufficiently small and* $\|u_k^s - \bar{u}\|_\infty < \rho$, *then* $u_{k+1}^n \in L^q$ *is a global minimizer of* $\psi[u_k^s]$ *and*

$$\psi[u_k^s](u_{k+1}^n) = -\frac{1}{2}\langle u_{k+1}^n - u_k^s, M(u_k^s)(u_{k+1}^n - u_k^s)\rangle \leq -\frac{C_M}{2}(\|s^e\|_2^2 + \|s^i\|_2^2), \tag{64}$$

$$\psi[u_k^s](P(u_{k+1}^n)) \leq \psi[u_k^s](u_{k+1}^n) - \frac{C_M}{2}\|s^e\|_2^2 + (c_r + \nu^{-1}L_g\rho)\|s^e\|_2\|s^i\|_2 \tag{65}$$

with $s^e = u_{k+1}^n - P(u_{k+1}^n)$, $s^i = P(u_{k+1}^n) - u_k^s$. *Moreover,*

$$\psi[u_k^s](u_{k+1}^N) \leq \max\{\xi, 1 - \|s^i\|_q\}\psi[u_k^s](P(u_{k+1}^n)), \tag{66}$$

and hence

$$\frac{\psi[u_k^s](u_{k+1}^N)}{\psi[u_k^s](u_{k+1}^n)} \geq \max\{\xi, 1 - \|s^i\|_q\}\left(1 - O\left(\frac{\|s^e\|_2}{\|s^i\|_2}\right)\right). \tag{67}$$

Proof. Setting $s = u_{k+1}^n - u_k^s$ we have $s = s^e + s^i$ and $s^e s^i \geq 0$ a.e. on Ω. We use the abbreviations $d_k = d(u_k^s)$, $g_k = g(u_k^s)$, $M_k = M(u_k^s)$. According to step (3) in Algorithm 5.16, we have $M_k s = -g_k$. Hence, u_{k+1}^n is a stationary point of $\psi[u_k^s]$ and, therefore, its global minimum by Lemma 10.2. Moreover,

$$\psi[u_k^s](u_{k+1}^n) = -\frac{1}{2}\langle s, M_k s\rangle \leq -\frac{C_M}{2}\|s^e + s^i\|_2^2 \leq -\frac{C_M}{2}(\|s^e\|_2^2 + \|s^i\|_2^2).$$

To prove the second inequality we observe that for all $x \in \Omega$ with $s(x)g_k(x) < 0$ and $d_k(x) < c(x)$ we have

$$s^e(x) = 0 \quad \text{or} \quad |s^i(x)| \geq d_k(x).$$

In fact, if $s^e(x) \neq 0$, then either $s^i(x) = b(x) - u_k^s(x) > 0$ or $s^i(x) = a(x) - u_k^s(x) < 0$. In the first case we have $g_k(x) < 0$ and thus $d_k(x) = b(x) - u_k^s(x) = s^i(x)$ or $d_k(x) < c(x) < b(x) - u_k^s(x) = s^i(x)$. The second case enforces $g_k(x) > 0$ which

implies $d_k(x) = u_k^s(x) - a(x) = |s^i(x)|$ or $d_k(x) < c(x) < u_k^s(x) - a(x) = |s^i(x)|$. Hence, we get with $N = \{x \in \Omega : s(x)g_k(x) < 0\}$ and $J = \{x \in \Omega : d_k(x) < c(x)\}$

$$\left\langle s_N^e, \frac{|g_k|_J}{d_k} s^i \right\rangle \geq -\langle s_{N\cap J}^e, g_k\rangle.$$

Using this, we obtain

$$\begin{aligned}
\psi[u_k^s](u_{k+1}^n) &= \langle s^e + s^i, g_k\rangle + \frac{1}{2}\left\langle s^e + s^i, M_k(s^e + s^i)\right\rangle \\
&= \psi[u_k^s](P(u_{k+1}^n)) + \langle s_{N^c}^e, g_k\rangle + \langle s_{N\setminus J}^e, g_k\rangle + \langle s_{N\cap J}^e, g_k\rangle + \left\langle s_N^e, \frac{|g_k|_J}{d_k} s^i \right\rangle \\
&\quad + \left\langle s_{N^c}^e, \frac{|g_k|_J}{d_k} s^i \right\rangle + \frac{1}{2}\langle s^e, M_k s^e\rangle + \langle s^e, \nabla^2 f(u_k^s) s^i\rangle \\
&\geq \psi[u_k^s](P(u_{k+1}^n)) + \langle s_{N\setminus J}^e, g_k\rangle + \frac{1}{2}\langle s^e, M_k s^e\rangle + \langle s^e, \nabla^2 f(u_k^s) s^i\rangle.
\end{aligned}$$

We still need an estimate for $\langle s_{N\setminus J}^e, g_k\rangle$. To this end, we use the inclusion $N \setminus J \subset J^c$. Since $|d_k(x) - d(\bar u)(x)| \leq L_d\rho$ and $d_k(x) \geq c(x) \geq \nu$ on J^c we have $d(\bar u)(x) > 0$ a.e. on J^c for $\rho > 0$ small enough. (O2) yields $g(\bar u)(x) = 0$ on J^c and thus

$$|\langle s_{N\setminus J}^e, g_k\rangle| \leq \|s_e\|_{1,J^c}\|g_k - g(\bar u)\|_\infty \leq L_g\rho\mu(J^c)^{1/2}\|s_e\|_2.$$

Furthermore, $|g_k(x)| \leq L_g\rho < \nu$ on J_c for small ρ, which, since $d_k(x) \geq \nu$, requires

$$\nu \leq d_k(x) \leq \min\{b(x) - u_k^s(x), u_k^s(x) - a(x)\} \leq |s^i(x)|.$$

Hence, by Lemma 2.4

$$\mu(J^c) \leq \mu\{x \in \Omega : |s^i(x)| \geq \nu\} \leq \nu^{-2}\|s^i\|_2^2.$$

We conclude $|\langle s_{N\setminus J}^e, g_k\rangle| \leq \nu^{-1}L_g\rho\|s_e\|_2\|s^i\|_2$. Therefore, (65) holds. Now

$$u_{k+1}^N - u_k^s = P[u_k^s](u_{k+1}^n) - u_k^s = \max\{\xi, 1 - \|s^i\|_q\}(P(u_{k+1}^n) - u_k^s) \stackrel{\text{def}}{=} \tau(P(u_{k+1}^n) - u_k^s).$$

This implies (66), for

$$\begin{aligned}
\psi[u_k^s](P[u_k^s](u_{k+1}^n)) &= \tau\langle s^i, g_k\rangle + \frac{\tau^2}{2}\langle s^i, M_k s^i\rangle \\
&\leq \tau\left(\langle s^i, g_k\rangle + \frac{1}{2}\langle s^i, M_k s^i\rangle\right) = \tau\psi[u_k^s](P(u_{k+1}^n)),
\end{aligned}$$

where the inequality follows from $0 \leq \tau < 1$ and $\langle s^i, M_k s^i\rangle \geq 0$; see Lemma 10.2. Now (64)–(66) and a straightforward calculation give (67). ∎

Let the assumptions of Theorem 5.10 hold. Using (34) we have for $\|u_k^s - \bar u\|_\infty$ small enough

$$\begin{aligned}
\|u_{k+1}^n - u_k^s\|_q &\leq \|u_{k+1}^n - \bar u\|_q + \|u_k^s - \bar u\|_q \leq (C\Phi_{\bar p}(\|u_k^s - \bar u\|_\infty) + m_{q,\infty})\|u_k^s - \bar u\|_\infty \\
&\leq C_\Delta\|u_k^s - \bar u\|_\infty
\end{aligned}$$

with C_Δ appropriately chosen. Now

$$\|u_{k+1}^N - u_k^s\|_q = \|P[u_k^s](u_{k+1}^n) - u_k^s\|_q \leq \|P(u_{k+1}^n) - u_k^s\|_q \leq C_\Delta\|u_k^s - \bar u\|_\infty. \tag{68}$$

Hence, for $\|u_k^s - \bar{u}\|_\infty$ small enough we have (cf. (D))

$$u_{k+1}^N \in \mathcal{B}^\circ, \quad \|u_{k+1}^N - u_k^s\|_q \le \beta_0 \Delta_{\min}.$$

Using this in Theorem 10.3 we can show that if $\|s^e\|_2/\|s^i\|_2$ eventually remains small enough, then Algorithm 10.1 turns into the superlinearly convergent Algorithm 5.16. In particular, this happens if no smoothing steps are required.

THEOREM 10.4. *Let the assumptions of Theorem* 10.3 *as well as* (C) *and* (S) *hold. Then there are $\rho > 0$, $\varepsilon > 0$ such that if step $k-1$ was accepted and $\|u_k - \bar{u}\|_q < \rho$, then $s_k = s_k^N$ and step k is accepted whenever*

$$\frac{\|u_{k+1}^n - P(u_{k+1}^n)\|_2}{\|P(u_{k+1}^n) - u_k^s\|_2} < \varepsilon \tag{69}$$

holds. If there is $C_1 > 0$ with $\|u_k^s - \bar{u}\|_\infty \le C_1 \|u_k^s - \bar{u}\|_q$, then (69) *is automatically satisfied for $\|u_k - \bar{u}\|_q$ small enough.*

Proof. Assume that step $k-1$ was accepted and $\|u_k - \bar{u}\|_q < \rho$ with $\rho > 0$ sufficiently small. We use s^e and s^i as defined in Theorem 10.3. Since $\|u_k^s - \bar{u}\|_\infty \le C_S \|u_k - \bar{u}\|_q$ by (S), we get with (68)

$$\|u_{k+1}^N - u_k^s\|_q \le \|s^i\|_q \le C_\Delta C_S \|u_k - \bar{u}\|_q \le C_\Delta C_S \rho.$$

Hence, $u_{k+1}^N = u_k^s + s_{k+1}^N \in \mathcal{B}^\circ$ with $\|s_k^N\|_q \le \Delta_{\min} \le \Delta_k$ for ρ small enough. Choose $0 < \tilde{\varepsilon} < 1$ such that $(1-\tilde{\varepsilon})^2 > \beta$ with β given in (D). Possibly after reducing ρ we achieve $\|s^i\|_q \le \tilde{\varepsilon}$. For $0 < \varepsilon < 1$ sufficiently small we have by (67) and (69)

$$\frac{\psi[u_k^s](u_{k+1})}{\psi[u_k^s](u_{k+1}^n)} \ge (1-\tilde{\varepsilon})^2 > \beta.$$

Since u_{k+1}^n is the global minimizer of $\psi[u_k^s]$ by Theorem 10.3, $s_k = s_k^N$ obviously satisfies (D) for $\psi[u_k^s]$.

Now assume $\|u_k^s - \bar{u}\|_\infty \le C_1 \|u_k^s - \bar{u}\|_q$. Then Lemma 2.2 yields with $\theta = 2/q$

$$\|u_k^s - \bar{u}\|_q \le \|u_k^s - \bar{u}\|_2^\theta \|u_k^s - \bar{u}\|_\infty^{1-\theta} \le C_1^{1-\theta} \|u_k^s - \bar{u}\|_2^\theta \|u_k^s - \bar{u}\|_q^{1-\theta}$$

and thus

$$\|u_k^s - \bar{u}\|_2 \ge C_1^{1-\frac{1}{\theta}} \|u_k^s - \bar{u}\|_q \overset{\text{def}}{=} C_3 \|u_k^s - \bar{u}\|_q.$$

To show (69) for ρ small enough we use $\|u_{k+1}^n - P(u_{k+1}^n)\|_2 \le \|u_{k+1}^n - \bar{u}\|_2$ and get

$$\begin{aligned} \|P(u_{k+1}^n) - u_k^s\|_2 &\ge \|u_k^s - \bar{u}\|_2 - \|P(u_{k+1}^n) - u_{k+1}^n\|_2 - \|u_{k+1}^n - \bar{u}\|_2 \\ &\ge \|u_k^s - \bar{u}\|_2 - 2\|u_{k+1}^n - \bar{u}\|_2 \ge \frac{C_3}{C_1} \|u_k^s - \bar{u}\|_\infty - 2\|u_{k+1}^n - \bar{u}\|_2. \end{aligned}$$

Moreover, Theorem 5.10 yields

$$\|u_{k+1}^n - \bar{u}\|_2 \le m_{2,q} \|u_{k+1}^n - \bar{u}\|_q \le m_{2,q} C \Phi_{\bar{p}}(\|u_k^s - \bar{u}\|_\infty) \|u_k^s - \bar{u}\|_\infty.$$

Hence, for $\|u_k^s - \bar{u}\|_\infty$ sufficiently small we get

$$\frac{\|u_{k+1}^n - P(u_{k+1}^n)\|_2}{\|P(u_{k+1}^n) - u_k^s\|_2} \le \left(\frac{C_3 \|u_k^s - \bar{u}\|_\infty}{C_1 \|u_{k+1}^n - \bar{u}\|_2} - 2 \right)^{-1} \le \left(\frac{C_3}{C_1 m_{2,q} C \Phi_{\bar{p}}(\|u_k^s - \bar{u}\|_\infty)} - 2 \right)^{-1}$$

and the last term is $< \varepsilon$ for small ρ, since $\Phi_{\bar{p}}(\|u_k^s - \bar{u}\|_\infty) \le \Phi_{\bar{p}}(C_s \rho)$ tends to zero as $\rho \to 0$. ∎

11. Application to a control problem. In this section we present numerical results for the application of Algorithm 5.16 to a boundary control problem governed by a nonlinear heat equation which is a simplified model for the heating of a probe in a kiln. Let $Q \stackrel{\text{def}}{=} (0,1)$ denote the spatial domain with $x = 0$ at the boundary and $x = 1$ at the inside of the probe. The temperature $y(x,t)$, $(x,t) \in Q \times (0,T) \stackrel{\text{def}}{=} Q_T$ of the probe satisfies the nonlinear heat equation

$$\begin{aligned} \tau(y)y_t - (\kappa(y)y_x)_x &= h && \text{on } Q_T, \\ \kappa(y(0,t))y_x(0,t) &= \zeta(y(0,t) - u(t)), && t \in (0,T), \\ \kappa(y(1,t))y_x(1,t) &= 0, && t \in (0,T), \\ y(x,0) &= y_0(x), && x \in Q, \end{aligned} \tag{70}$$

where $y_0 : Q \longrightarrow \mathbb{R}$ is the initial temperature, $\tau, \kappa : \mathbb{R} \longrightarrow \mathbb{R}$ denote the specific heat capacity and the heat conduction, respectively, $h : Q_T \longrightarrow \mathbb{R}$ is a source term, $\zeta \in \mathbb{R}$ a given scalar, and $u : (0,T) \longrightarrow \mathbb{R}$ the control. For consistency with our notations let $\Omega \stackrel{\text{def}}{=} (0,T)$.

The control u shall be determined in such a way that the temperature $y(1,t)$ inside the probe follows a given temperature profile $y_d(t)$. Since it is well known that this nonlinear inverse heat conduction problem is ill-posed, we add a regularization in the control space and choose as objective function

$$J(y,u) = \frac{1}{2}\int_0^T \left((y(1,t) - y_d(t))^2 + \alpha\, u(t)^2\right) dt$$

with $y_d \in L^\infty((0,T))$. The problem was considered in [5]. We define the space

$$W(0,T) \stackrel{\text{def}}{=} \left\{ y \in L^2(0,T;H^1(Q)) \,:\, y_t \in L^2(0,T;H^1(Q)') \right\}$$

with norm $\|y\|_{W(0,T)} \stackrel{\text{def}}{=} \|y\|_{L^2(0,T;H^1)} + \|y_t\|_{L^2(0,T;(H^1)')}$. It is well known that $W(0,T)$ is a Hilbert space and that the embedding $W(0,T) \hookrightarrow C(0,T;L^2(Q))$ is continuous. Under the assumption that $\kappa, \tau \in C(\mathbb{R})$ with

$$0 < \kappa_1 \le \kappa(s) \le \kappa_2, \quad 0 < \tau_1 \le \tau(s) \le \tau_2 \quad \text{for all } s \in \mathbb{R}$$

it is shown in [5] that for all $h \in L^2(0,T;L^2(Q))$, $y_0 \in L^2(Q)$, and $u \in L^2(\Omega)$ there exists a solution $y \in W(0,T)$ of the state equation (70) which satisfies the stability estimate

$$\|y\|_{W(0,T)} \le C\left(\|h\|_{L^2(0,T;L^2)} + \|u\|_2 + \|y_0\|_2\right). \tag{71}$$

Uniqueness is proven under the additional assumption $y_x \in L^\infty(0,T;L^r(Q))$, $r > 2$, and $\kappa, \tau \in C^1(\mathbb{R})$. Furthermore, it was shown that for $\alpha > 0$ there exists an optimal solution $\bar{u} \in L^2(\Omega)$ of the control problem

$$\text{minimize } J(y,u) \quad \text{subject to} \quad y \in W(0,T),\ u \in L^2(\Omega) \text{ satisfy (70).} \tag{72}$$

With the lower and upper bounds $a, b \in L^\infty(\Omega)$, $b - a \ge \nu > 0$, we introduce the additional box constraints

$$u \in \mathcal{B} \stackrel{\text{def}}{=} \left\{ u \in L^2(\Omega) \,:\, a \le u \le b \right\}. \tag{73}$$

Since $\mathcal{B}$ is a closed bounded convex subset of $L^2(\Omega)$, exactly the same arguments as in [5] can be used to prove the existence of an optimal control $\bar{u} \in \mathcal{B}$ for $\alpha \ge 0$.

Assuming that for $u \in \mathcal{B}$ the solution $y = y(u)$ to (70) is unique, we can define the reduced objective function $f(u) \stackrel{\text{def}}{=} J(y(u), u)$ for which (72), (73) is equivalent to (P). This way of eliminating y is called the black-box approach.

For the rest of this paragraph assume that κ and τ are constant. Then the existence and stability result (71) is well known. As will be shown now, it implies that the affine linear mapping $u \in L^2(\Omega) \longmapsto y(1,.) \in L^q(\Omega)$ is (even completely) continuous for $2 \le q < 4$ and therefore smooth. Thus, the objective $f : L^2(\Omega) \longrightarrow \mathbb{R}$ is a well-defined and smooth quadratic function. The case $2 < q < 4$ will be used below to derive a regularity result for the gradient. In the proof of the assertion on $u \longmapsto y(1,.)$ we use the symbol "$\hookrightarrow$" for continuous and "$\hookrightarrow\hookrightarrow$" for compact embeddings. (71) shows the continuity of $u \in L^2(\Omega) \longmapsto y \in W(0,T)$. To complete the argument we show that $y \in W(0,T) \longmapsto y(1,.) \in L^q(\Omega)$ is compact for $1 \le q < 4$. Let $1/2 < \theta < \Theta < 1$. Since $H^1(Q) \hookrightarrow\hookrightarrow H^\Theta(Q) \hookrightarrow H^1(Q)'$, we have by a Lions lemma that $W(0,T) \hookrightarrow\hookrightarrow L^2(0,T;H^\Theta(Q))$ (see [21, Thm. 5.1]). Moreover, from the interpolation result $H^\theta(Q) = [L^2(Q), H^\Theta(Q)]_{\theta/\Theta}$ it can be deduced that

$$\|\cdot\|_{L^{2\Theta/\theta}(0,T;H^\theta)} \le C \, \|\cdot\|^{1-\theta/\Theta}_{L^\infty(0,T;L^2)} \|\cdot\|^{\theta/\Theta}_{L^2(0,T;H^\Theta)}.$$

Hence, $W(0,T) \hookrightarrow L^\infty(0,T;L^2(Q))$ and $W(0,T) \hookrightarrow\hookrightarrow L^2(0,T;H^\Theta(Q))$ yield the compact embedding $W(0,T) \hookrightarrow\hookrightarrow L^{2\Theta/\theta}(0,T;H^\theta(Q))$. Finally, since $H^\theta(Q) \hookrightarrow C([0,1])$, we conclude that $y \in W(0,T) \longmapsto y(1,.) \in L^{2\Theta/\theta}(\Omega)$ is compact. Now (71) shows the complete continuity of $u \in L^2(\Omega) \longmapsto y(1,.) \in L^q(\Omega)$ for $1 \le q < 4$.

In particular, the regularization in J is necessary, since (72) is ill-posed for $\alpha = 0$.

By standard results (see [22]) the gradient representation g of $u \longmapsto J(y(u), u)$ with respect to the inner product on $L^2(\Omega)$ is given by $g(u) = \alpha u + K(u)$, where $K(u) = \zeta p(0,.)$ and the *adjoint state* p satisfies

$$\begin{aligned} \tau p_t + \kappa p_{xx} &= 0, && \text{on } Q_T, \\ \kappa p_x(0,t) &= \zeta p(0,t), && t \in (0,T), \\ \kappa p_x(1,t) &= y(1,t) - y_d(t), && t \in (0,T), \\ p(x,T) &= 0, && x \in Q \end{aligned} \tag{74}$$

in the weak sense. Using Green's function, p is given by an integral equation of Volterra type with weakly singular kernel from which one can deduce that (74) defines a completely continuous affine linear mapping $y(1,.) \in L^q(\Omega) \longmapsto p \in C(Q_T)$ for all $q > 2$ (see, e.g., [24]). Combining this with the previous considerations we obtain the complete continuity of the affine linear mapping $u \in L^2(\Omega) \longmapsto K(u) = \zeta p(0,.) \in C(\Omega)$. Hence, the Fréchet-derivative K' of $K : L^2(\Omega) \longrightarrow C(\Omega)$ exists and is given by the compact linear operator $K'(u) : v \in L^2(\Omega) \longmapsto K(v) - K(0) \in C(\Omega)$. We conclude that the assumptions (A1), (A2′), and (A4) are satisfied for $q = 2$, $s = r = \infty$ and the results of section 8 can be applied.

While similar results can be shown for nonlinear boundary conditions (cf. [24], [17], [19]), a differentiability result for the nonlinear problem (72) seems not to be available. Since (72) is of importance in applications, e.g., the sterilization of canned food, we nevertheless present numerical results for the nonlinear problems and content ourselves with the complete justification of our assumptions for the case of constant κ and τ.

11.1. Discretization. As in [14], [20] we use the discretization of (72) proposed in [5]. For the space discretization we approximate $H^1(Q)$ in the variational formulation of (70) by the space $V_{\Delta x}$ of continuous functions that are piecewise linear on

the intervals $[i\Delta x, (i+1)\Delta x]$, $\Delta x \stackrel{\text{def}}{=} 1/N_x$, $i = 0, \dots, N_x - 1$. Since the time differentiation in the variational form of (70) is linear with respect to the transformed state $\phi(y) \stackrel{\text{def}}{=} \int_0^y \tau(\xi)\, d\xi$, a discontinuous Galerkin method with respect to ϕ is used where $L^2(0,T;H^1(Q))$ is approximated by the space Y_Δ of $V_{\Delta x}$-valued functions that are piecewise constant on $(k\Delta t, (k+1)\Delta t]$, $\Delta t \stackrel{\text{def}}{=} 1/N_t$, $k = 0, \dots, N_t - 1$ (the same discretization is obtained by applying a backward Euler). This leads in a natural way to the approximation of h and y_0 by their L^2-projection onto Y_Δ and $V_{\Delta x}$, respectively. The discrete control space $U_{\Delta t}$ consists of piecewise constant functions on the same partition of $(0,T]$ and y_d is approximated by its L^2 projection onto the same space. For details we refer to [5], [14], [20].

It was shown in [5] that the resulting implicit scheme admits a unique solution for $\Delta t/\Delta x^2 \le \lambda < (\tau_2/\kappa_1 - \tau_1/\kappa_2)^{-1}/6$ that converges to a solution of (70) as $\Delta t, \Delta x$ tend to zero.

11.2. Numerical tests. For the application of Algorithm 5.16 we use Example 1 of [20] (see also [14]): $T = 0.5$, $\zeta = 1$, and

$$\tau(y) = 4 + y,\ \kappa(y) = 4 - y,\ y_d(t) = 2 - e^{-t},\ y_0(x) = 2 + \cos \pi x,$$
$$h(x,t) = (-6 + 2\pi^2)\, e^{-t} \cos \pi x + \pi^2\, e^{-2t} - (1 + 2\pi^2)\, e^{-2t} \cos^2 \pi x.$$

Then the optimal control for $\alpha = 0$ without bound constraints is $u^*(t) = 2 + e^{-t}$ with associated state $y^*(x,t) = 2 + e^{-t} \cos \pi x$. The regularization parameter was set to $\alpha = 10^{-4}$. The L^2 gradient representation of $f(u) \stackrel{\text{def}}{=} J(y(u), u)$ in U_Δ was computed via the discrete adjoint equation; cf. [5]. Since—at least in the case of constant κ and τ—$\nabla^2 f(u)$ is a compact perturbation of αI, a quasi-Newton approximation of $\nabla^2 f(u)$ like BFGS or PSB is efficient also in the L^2-Hilbert space setting; see [12], [16]. Thus, we may expect that a BFGS- or PSB-approximation of the Hessian in the discrete model performs nearly independent of the discretization; cf. [16] for the mesh-independence of BFGS. For the numerical tests, Algorithm 5.16 was embedded in the trust-region framework of [26] as described in Algorithm 10.1. We took a L^2-trust-region and used an extension of the Steihaug CG-iteration in the scaled variables $\hat{s} \stackrel{\text{def}}{=} d_k^{-1} s$ to compute an approximate solution to (63) satisfying the decrease condition (D): Let u_k^s be the current iterate and B_k the approximation of $\nabla^2 f(u_k^s)$. A CG-iteration in the scaled variable $\hat{s}$ is started. If the process leaves the trust-region or $\mathcal{B}$, or if negative curvature[1] is detected, Steihaug's method yields s_k^{SH} and $s_k^1 = \sigma s_k^{SH}$ is a candidate for step 2.4 in Algorithm 10.1. Here $\sigma \in (0,1]$ is chosen maximal such that $u_k^s + 1.0005\, s_k^1 \in \mathcal{B}$. In contrast to Steihaug's algorithm we continue the CG-iteration as long as no negative curvature is detected even if it leaves the trust-region or $\mathcal{B}$ until an inexact unconstrained minimizer s_k^n of $\psi[u_k^s](u_k^s + .)$ with $\|d_k \nabla \psi[u_k^s](u_k^s + s_k^n)\|_2 \le 10^{-4} \|d_k g_k\|_2$ is found. Then $u_k^s + s_k^n$ is an approximation for u_{k+1}^n in Algorithm 10.1 with $\nabla^2 f(u_k^s)$ replaced by B_k. If the CG-iteration left the trust-region or $\mathcal{B}$, we take $s_k^2 = \min\left(\Delta_k / \|s_k^N\|_2, 1\right) s_k^N$ with the projected step $s_k^N = P[u_k^s](u_k^s + s_k^n) - u_k^s$ according to Algorithm 10.1 and $s_k^3 = \min(\Delta_k / \|s_k^S\|_2, 1) s_k^S$ with s_k^S obtained from s_k^n by the stepsize rule 2.4′ as further candidates. In (39) and 2.4′ we took $\xi = 0.99995$. Now we set $u_{k+1} = u_k^s + s_k$, where $s_k = s_k^i$, $i \in \{1,2,3\}$ is the trial step that provides the best reduction of $\psi[u_k^s]$. As the smoothing step in 2.2 of Algorithm 10.1 we use (cf. section 8)

$$S_k^\circ(u_k) \stackrel{\text{def}}{=} \begin{cases} P[u_k](u_k - \alpha^{-1} g(u_k)) & \text{if } k \ge 1 \text{ and } \|u_k - u_{k-1}^s\|_\infty \ge 3 \|u_k - u_{k-1}^s\|_2, \\ u_k & \text{else.} \end{cases}$$

[1] This does not apply to BFGS-approximations.

For our numerical results we used a BFGS-approximation of the Hessian with $B_0 = \alpha I$. In Algorithm 10.1 we set $\vartheta = 2$, $\Delta_0 = 1$, $\eta_1 = 0.1$, $\eta_2 = 0.75$, $\eta_3 = 0.9$ and $\gamma_1 = 0.5$, $\gamma_2 = \gamma_3 = 2$. The stopping criterion was $\|d(u_k^s)g(u_k^s)\|_2 \le 10^{-10}$. The upper and lower bounds for the control were $a \equiv -1000$, $b \equiv 0.8$ and we used $c \stackrel{\text{def}}{=} 0.075 \min\{b-a, 0.8\}$ in the definition of the discrete scaling function d. The optimization was started with $u_0 \equiv 0.05$.

For $N_t = 100$, $N_x = 20$ Table 4 shows the norm of the step $\|u_{k+1}^s - u_k^s\|_\infty$ and the norm $\|d(u_{k+1}^s)g(u_{k+1}^s)\|_2$ of the scaled gradient for three different algorithms. The first algorithm is as described above. It uses also the projection step s_k^2 as a candidate for the trial step and performs a smoothing step if necessary (see above). The second algorithm is the same but without smoothing. The third algorithm is the same as the second but uses only the trial steps s_k^1, s_k^3 and not the projected step s_k^2 that is suggested by our investigations.

There were no rejected trial steps in all three algorithms. Except for the first two iterations the projected step s_k^2 was chosen by the first two algorithms. Obviously the first algorithm provides the fastest convergence. But if the smoothing steps are omitted also, the usage of the projected step s_k^2 leads to a significant acceleration of the local convergence in comparison to a stepsize-based algorithm. To compare the dependence of the methods on the mesh-size, we list also the number of gradient evaluations and iterations for $(N_t, N_x) = (400, 80)$, $(3000, 200)$, and $(6000, 400)$.

The first algorithm wins on all grids and the second needs at most one additional gradient evaluation. Thus, smoothing is not very important for this example. Both

TABLE 4
Results for $(N_t, N_x) = (100, 20)$, $(400, 80)$, $(3000, 200)$, *and* $(6000, 400)$.

	Proj. and smooth.		Projection		Stepsize rule	
k	$\|s_k^s\|_\infty^\dagger$	$\|d_{k+1}^s g_{k+1}^s\|_2^\dagger$	$\|s_k^s\|_\infty^\dagger$	$\|d_{k+1}^s g_{k+1}^s\|_2^\dagger$	$\|s_k^s\|_\infty^\dagger$	$\|d_{k+1}^s g_{k+1}^s\|_2^\dagger$
	$N_t = 100 \quad N_x = 20$					
	grad-evals: 8		grad-evals: 8		grad-evals: 13	
0	2.041E−01	3.128E−06	2.041E−01	3.128E−06	2.041E−01	3.128E−06
1	4.093E−01	1.752E−06	4.093E−01	1.752E−06	4.093E−01	1.752E−06
2	3.381E−01	1.304E−07	3.381E−01	1.304E−07	1.366E−01	1.146E−06
3	7.382E−02	4.110E−10*	7.381E−02	9.494E−09	6.285E−02	8.221E−07
4	2.821E−04	1.984E−12*	1.590E−02	1.931E−09	7.082E−02	5.951E−07
5			6.501E−03	3.038E−10	7.028E−02	4.029E−07
6			1.634E−03	1.908E−11	6.683E−02	2.379E−07
7					7.704E−02	6.209E−08
8					1.849E−02	1.211E−08
9					9.845E−03	1.640E−09
10					3.251E−03	2.489E−10
11					1.790E−03	4.387E−11
	$N_t = 400 \quad N_x = 80$					
	grad-evals: 8 iterations: 5		grad-evals: 9 iterations: 8		grad-evals: 14 iterations: 13	
	$N_t = 3000 \quad N_x = 200$					
	grad-evals: 9 iterations: 7		grad-evals: 10 iterations: 9		grad-evals: 16 iterations: 15	
	$N_t = 6000 \quad N_x = 400$					
	grad-evals: 10 iterations: 8		grad-evals: 11 iterations: 10		grad-evals: 16 iterations: 15	

† $s_k^s = u_{k+1}^s - u_k^s$, $d_k^s = d(u_k^s)$, $g_k^s = g(u_k^s)$. * Smoothing occurred, i.e., $u_{k+1}^s \ne u_{k+1}$.

algorithms are nearly mesh-independent and need significantly less iterations than the third algorithm, which shows a weak mesh-dependence for the transition from coarse to fine grids. Note, however, that the third algorithm shows strong mesh-dependence for certain other problems; cf. Examples 6.3 and 6.5.

It is also possible to treat the state equation as an equality constraint by combining the SQP-approach with our method. For the finite-dimensional case this was proposed in [8] and [27]. The main advantage of these methods is that they only require the solution of the linearized state equation, whereas in the black-box approach the state equation has to be solved, which in the problem under consideration is about twice as expensive as the solution of the linearized equation. We believe that the extension of our theory to this class of SQP-methods is possible. For the case without box-constraints, an SQP-method for the above problem was analyzed in [20]. Also in this simpler case, an SQP-step requires the computation of the reduced gradient and the solution of the linearized state equation.

Conclusions. We have developed an affine-scaling interior-point Newton algorithm for bound-constrained minimization subject to pointwise bounds in L^p-space. The method is an extension of the algorithms by Coleman and Li [6], [7] for finite-dimensional problems. Our infinite-dimensional framework raised a couple of difficulties which are not present in the finite-dimensional case. A careful analysis led to several modifications of the original algorithm which enabled us to prove superlinear convergence for the resulting method. Under a slightly stronger strict complementarity condition we proved convergence with Q-rate >1. Our main modifications are the introduction of a smoothing step and the implementation of the back-transport by a projection instead of the usual stepsize rule. The smoothing step takes care of the fact that, in general, we can show that only for suitable $q < s$ the affine-scaling Newton step produces a point which is much closer to the solution in L^q (but not necessarily in L^s) than the current iterate was in L^s. The necessity of a smoothing step was also observed by Kelley and Sachs [17] in their study on projected Newton methods. The back-transport is required because the solution of the affine-scaling Newton equation may lie outside of the feasible set $\mathcal{B}$. In the finite-dimensional case one can prove that a stepsize rule to enforce strict feasibility generates stepsizes that converge to one. In our infinite-dimensional setting, however, this is no longer true, as we have demonstrated in Example 6.3. Therefore, we have defined a back-transport on the basis of the pointwise projection onto $\mathcal{B}$. We have discussed how smoothing steps can be obtained for a class of regularized problems. Moreover, we have shown that our theory is applicable under the assumptions used by Kelley and Sachs [17] as well as those by Dunn and Tian [9]. We have demonstrated that our algorithm can be used as an accelerator for the class of globally convergent trust-region interior-point methods introduced in [26]. The good performance of this algorithm is documented by our numerical results for the boundary control of a heating process.

Acknowledgments. The major part of this work was done while the authors were visiting the Department of Computational and Applied Mathematics and the Center for Research on Parallel Computation, Rice University. They would like to thank John Dennis, Rice University, and Klaus Ritter, Technische Universität München, for making this pleasant and fruitful visit possible.

This work was initiated and influenced by many discussions with Matthias Heinkenschloss, Rice University. His support is greatly acknowledged. The authors also are grateful to John Dennis, Rice University, and Luís Vicente, Universidade de Coimbra, for several stimulating discussions.

The authors are grateful to the referees for their insightful reading of this paper and their constructive comments.

REFERENCES

[1] R. Adams, *Sobolev Spaces*, Academic Press, New York, 1975.

[2] W. Alt, *The Lagrange–Newton method for infinite dimensional optimization problems*, Numer. Funct. Anal. Optim., 11 (1990), pp. 201–224.

[3] D. P. Bertsekas, *Projected Newton methods for optimization problems with simple constraints*, SIAM J. Control Optim., 20 (1982), pp. 221–246.

[4] M. A. Branch, T. F. Coleman, and Y. Li, *A Subspace, Interior, and Conjugate Gradient Method for Large-Scale Bound-Constrained Minimization Problems*, Tech. Rep. CTC95TR217, Center for Theory and Simulation in Science and Engineering, Cornell University, Ithaca, NY, 1995.

[5] J. Burger and M. Pogu, *Functional and numerical solution of a control problem originating from heat transfer*, J. Optim. Theory Appl., 68 (1991), pp. 49–73.

[6] T. F. Coleman and Y. Li, *On the convergence of interior-reflective Newton methods for nonlinear minimization subject to bounds*, Math. Programming, 67 (1994), pp. 189–224.

[7] T. F. Coleman and Y. Li, *An interior trust region approach for nonlinear minimization subject to bounds*, SIAM J. Optim., 6 (1996), pp. 418–445.

[8] J. E. Dennis, M. Heinkenschloss, and L. N. Vicente, *Trust-region interior-point SQP algorithms for a class of nonlinear programming problems*, SIAM J. Control Optim., 36 (1998), pp. 1750–1794.

[9] J. C. Dunn and T. Tian, *Variants of the Kuhn-Tucker sufficient conditions in cones of nonnegative functions*, SIAM J. Control Optim., 30 (1992), pp. 1361–1384.

[10] J. C. Dunn, *Second-order optimality conditions in sets of L^∞ functions with range in a polyhedron*, SIAM J. Control Optim., 33 (1995), pp. 1603–1635.

[11] J. C. Dunn, *On L^2-sufficient conditions and the gradient projection method for optimal control problems*, SIAM J. Control Optim., 34 (1996), pp. 1270–1290.

[12] A. Griewank, *The local convergence of Broyden-like methods on Lipschitzian problems in Hilbert spaces*, SIAM J. Numer. Anal., 24 (1987), pp. 684–705.

[13] M. Heinkenschloss, *A trust region method for norm constrained problems*, SIAM J. Numer. Anal., 35 (1998), pp. 1594–1620.

[14] M. Heinkenschloss, *Projected sequential quadratic programming methods*, SIAM J. Optim., 6 (1996), pp. 373–417.

[15] M. Heinkenschloss and L. N. Vicente, *Analysis of Inexact Trust-Region Interior-Point SQP Algorithms*, Tech. Rep. TR95–18, Department of Computational and Applied Mathematics, Rice University, Houston, TX, 1995.

[16] C. T. Kelley and E. W. Sachs, *Quasi-Newton methods and unconstrained optimal control problems*, SIAM J. Control Optim., 25 (1987), pp. 1503–1516.

[17] C. T. Kelley and E. W. Sachs, *Multilevel algorithms for constrained compact fixed point problems*, SIAM J. Sci. Comput., 15 (1994), pp. 645–667.

[18] C. T. Kelley and E. W. Sachs, *Solution of optimal control problems by a pointwise projected Newton method*, SIAM J. Control Optim., 33 (1995), pp. 1731–1757.

[19] C. T. Kelley and E. W. Sachs, *A Trust Region Method for Parabolic Boundary Control Problems*, Tech. Rep. CRSC–TR96–28, Center for Research in Scientific Computing, North Carolina State University, Raleigh, NC, 1996.

[20] F.–S. Kupfer and E. W. Sachs, *Numerical solution of a nonlinear parabolic control problem by a reduced SQP method*, Comput. Optim. Appl., 1 (1992), pp. 113–135.

[21] J. L. Lions, *Quelques Méthodes de Résolution des Problèmes aux Limites Non Linéaires*, Dunod, Paris, 1969.

[22] J. L. Lions, *Optimal Control of Systems Governed by Partial Differential Equations*, Springer, New York, 1971.

[23] H. Maurer, *First and second order sufficient optimality conditions in mathematical programming and optimal control*, Math. Programming Stud., 14 (1981), pp. 163–177.

[24] E. Sachs, *A parabolic control problem with a boundary condition of the Stefan-Boltzmann type*, Z. Angew. Math. Mech., 58 (1978), pp. 443–449.

[25] T. Tian and J. C. Dunn, *On the gradient projection method for optimal control problems with nonnegative L^2 inputs*, SIAM J. Control Optim., 32 (1994), pp. 516–537.

[26] M. Ulbrich, S. Ulbrich, and M. Heinkenschloss, *Global convergence of trust-region interior-point algorithms for infinite-dimensional nonconvex minimization subject to point-*

wise bounds, SIAM J. Control Optim., 37 (1999), pp. 731–764.

[27] L. N. Vicente, *Trust-Region Interior-Point Algorithms for a Class of Nonlinear Programming Problems*, Ph.D. thesis, Department of Computational and Applied Mathematics, Rice University, Houston, TX, 1996.

[28] L. N. Vicente, *On interior-point Newton algorithms for discretized optimal control problems with state constraints*, Optim. Methods Softw., 8 (1998), pp. 249–275.